Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

# Sequence-length variation of mtDNA HVS-I C-stretch in Chinese ethnic groups*

Feng CHEN[§1,3,4], Yong-hui DANG[§1], Chun-xia YAN[1,2], Yan-ling LIU[1], Ya-jun DENG[3], David J. R. FULTON[4], Teng CHEN[†‡1,2]

([1]*Department of Forensic Medicine, School of Medicine, Xi'an Jiaotong University, Xi'an 710061, China*)

([2]*Ministry of Education Key Laboratory of Environment and Genes Related to Diseases, Xi'an Jiaotong University, Xi'an 710061, China*)

([3]*Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 100029, China*)

([4]*Medical College of Georgia, 1459 Laney Walker Blvd, Augusta, Georgia 30912, USA*)

[†]E-mail: chenteng@mail.xjtu.edu.cn

Received May 11, 2009; Revision accepted July 8, 2009; Crosschecked Aug. 4, 2009

**Abstract:** The purpose of this study was to investigate mitochondrial DNA (mtDNA) hypervariable segment-I (HVS-I) C-stretch variations and explore the significance of these variations in forensic and population genetics studies. The C-stretch sequence variation was studied in 919 unrelated individuals from 8 Chinese ethnic groups using both direct and clone sequencing approaches. Thirty eight C-stretch haplotypes were identified, and some novel and population specific haplotypes were also detected. The C-stretch genetic diversity (*GD*) values were relatively high, and probability (*P*) values were low. Additionally, C-stretch length heteroplasmy was observed in approximately 9% of individuals studied. There was a significant correlation ($r$=−0.961, $P$<0.01) between the expansion of the cytosine sequence length in the C-stretch of HVS-I and a reduction in the number of upstream adenines. These results indicate that the C-stretch could be a useful genetic maker in forensic identification of Chinese populations. The results from the Fst and dA genetic distance matrix, neighbor-joining tree, and principal component map also suggest that C-stretch could be used as a reliable genetic marker in population genetics.

**Key words:** Mitochondrial DNA (mtDNA), Clone sequencing, Length heteroplasmy, Population genetics
**doi:**10.1631/jzus.B0920140 **Document code:** A **CLC number:** R89

## INTRODUCTION

Mitochondrial DNA (mtDNA) is well suited for forensic genetics, molecular anthropology, and medicine studies because of its stability, lack of recombination, rapid evolutionary changes, and high population specific polymorphisms (Anderson *et al*., 1981; Brown *et al*., 1982; Bowling *et al*., 1993; Allen *et al*., 1998; Macaulay *et al*., 1999; Ingman *et al*., 2000). The mtDNA sequence evolves much faster than nuclear DNA (Brown *et al*., 1982). A great number of mtDNA variations have been defined, including transition, transversion, deletion, inversion, insertion, and complex rearrangement (Brandon *et al*., 2005). MtDNA hypervariable segment I (HVS-I) contains a C-continuous tract termed the C-stretch, which is associated with sequence-length variations (Lee HY *et al*., 2004; Lutz-Bonengel *et al*., 2004) located in 16180~16193 nt. Due to slipping of the DNA polymerase during replication, the C-stretch evolves much faster than other regions of mtDNA, and variations in this region have been demonstrated widely among unrelated individuals (Malik *et al*., 2002a; Lee HY *et al*., 2004; Lutz-Bonengel *et al*., 2004). Furthermore, the mtDNA control region, especially the C-stretch, might be involved in the

‡ Corresponding author
§ The two authors contributed to this work equally
* Project supported by the Sciences and Technological Fundamental Resources Data of the Ministry of Education, China (No. 505015) and the Key Project for Science and Technology of Shaanxi Province, China (No. 2004K09-G12)

development of human diseases such as cancer (Tan *et al*., 2002; Lee HC *et al*., 2004; Meierhofer *et al*., 2004; Montanini *et al*., 2005; Sangkhathat *et al*., 2005).

C-stretch length heteroplasmy within both HVS-I and HVS-II has been observed, and the rates of heteroplasmic length variation revealed significant differences among distinct populations (Salas *et al*., 2001; Chen *et al*., 2002; Imaizumi *et al*., 2002; Mabuchi *et al*., 2007). An individual can exhibit two or more different C-stretch lengths in different tissues (Kirches *et al*., 2001; Lee *et al*., 2006; Lutz-Bonengel *et al*., 2008). Furthermore, even in the same tissue, especially in hairs, C-stretch length heteroplasmy was demonstrated in different hair shafts (Pfeiffer *et al*., 2004; Lee *et al*., 2006) and even in different parts of the same hair (Salas *et al*., 2001). As the 'out-of-phase' nucleotide pattern, C-stretch is not easily detected when compared to other regions of mtDNA, and C-stretch length heteroplasmy increases the difficulties of DNA sequencing (Lee HY *et al*., 2004; Lutz-Bonengel *et al*., 2004; Bini and Pappalardo, 2005). C-stretch is located in the middle of mtDNA HVS-I, and failure to interpret these sequence variations may hinder the application of the mtDNA control region to forensic and population genetics (Malik *et al*., 2002a; Lee HY *et al*., 2004). Therefore, the C-stretch might be highly significant in forensic identification and population genetic studies.

In the present study, we investigated the mtDNA HVS-I C-stretch sequence-length variation in 919 unrelated individuals from 8 Chinese ethnic groups by using direct and clone sequencing approaches and explored the significance of these findings to forensic, population genetics, and medicine studies.

## MATERIALS AND METHODS

### Sample collection

A total of 919 unrelated healthy individuals from 8 different Chinese ethnic groups were analyzed for C-stretch haplotypes in the mtDNA HVS-I region. Northern Han (*N*=188) samples were obtained from Xining, Xi'an and Beijing. Southern Han (*N*=117) samples were from Kunming, Wuhan and Zhanjiang. Tu (*N*=146), Tibetan (*N*=120) and Salar (*N*=100) were all collected from Qinghai Province. Oroqen (*N*=50) samples were collected from Inner Mongolia. Yugur (*N*=100) and Bonan (*N*=98) were collected from Gansu Province. The locations and the language families of the studied ethnic groups are shown in Table 1 and Fig.1 (Du and Yip, 1993; Ma, 1994). The geographic origin, nationality, and maternal pedigree (unrelated at least through three generations) of each individual were ascertained at first. Informed consent was obtained before sampling. The whole blood samples were obtained by venipuncture and put into ethylenediaminetetraacetic acid (EDTA) tubes.

**Table 1 Sample information of Chinese ethnic groups in the present study**

| Population | Location | Longitude | Latitude | Language[a] | Census size[b] | Sample size |
|---|---|---|---|---|---|---|
| 1. Northern Han | Xining, Qinghai | 101.75° E | 36.57° N | Chinese, Sino-Tibetan family | 1979200 | 37 |
| | Xi'an, Shaanxi | 108.95° E | 34.27° N | | 7410000 | 109 |
| | Beijing | 116.42° E | 39.92° N | | 14640000 | 42 |
| 2. Southern Han | Kunming, Yunnan | 102.73° E | 25.05° N | Chinese, Sino-Tibetan family | 5781000 | 43 |
| | Wuhan, Hubei | 114.32° E | 30.52° N | | 8312600 | 42 |
| | Zhanjiang, Guangdong | 110.30° E | 21.20° N | | 6072900 | 32 |
| 3. Tu | Huzhu Tu autonomous county, Qinghai | 101.90° E | 36.80° N | Mongolian group, Altaic family | 199470 | 146 |
| 4. Tibetan | Yushu Tibetan autonomous prefectures, Qinghai | 96.97° E | 33.03° N | Tibetan branch, Tibetan-Burman group, Sino-Tibetan family | 1134236 | 120 |
| 5. Oroqen | Hulun Buir League, inner Mongolia | 123.70° E | 50.58° N | Tungus branch, Manchu-Tungusic group, Altaic family | 8196 | 50 |
| 6. Salar | Xunhua Salar autonomous county, Qinghai | 102.40° E | 35.80° N | Tujue branch, Altaic family | 95815 | 100 |
| 7. Yugur | Yugur autonomous county, Gansu | 99.60° E | 38.80° N | Turkic branch, Altaic family | 13719 | 100 |
| 8. Bonan | Jishishan autonomous county, Gansu | 102.80° E | 35.70° N | Mongolian branch, Altaic family | 16505 | 98 |
| Total | | | | | | 919 |

[a] The language family is from Du and Yip (1993) and Ma (1994); [b] The total population size of each population is based on the fifth national Census (2000)

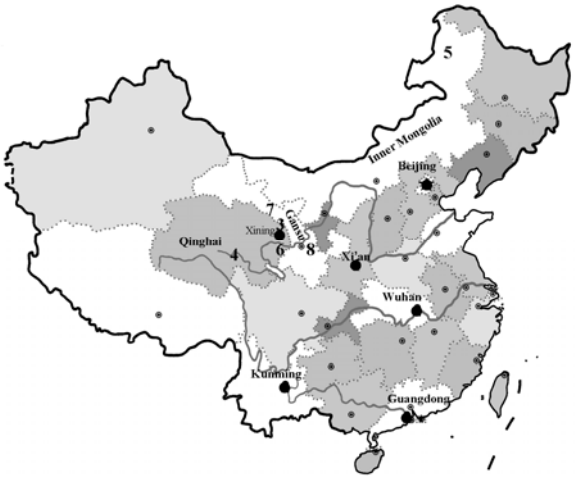


**Fig.1 Locations of Chinese ethnic groups in the current study. Numbers correspond to the population names in Table 1**

## DNA extraction, amplification, and direct sequencing

Genomic DNA was extracted using the Chelex-100 method as previously described (Walsh *et al*., 1991). Polymerase chain reaction (PCR) was performed in a 25-μl reaction volume using GeneAmp PCR System 9700 (PE Applied Biosystems, Foster City, CA, USA) with primers L15990 (5′-ACTCCACCATTAGCACC-3′) and H16503 (5′-CAGATGTCGGATACAGTTC-3′). Each reaction mixture contained 50 mmol/L KCl, 10 mmol/L Tris-HCl, 1.5 mmol/L $MgCl_2$, 50 mmol/L of dNTP each, 0.25 mmol/L of each primer, and 1.25 U AmpliTaqGold DNA polymerase. The thermal cycling conditions were 10 min at 95 °C followed by 36 cycles of 30 s at 95 °C, 30 s at 60 °C, and 45 s at 72 °C, and a final extension of 10 min at 72 °C. PCR products were purified by Centrisep (PE Applied Biosystems) and sequenced using BigDye® Terminator Cycle Sequencing Ready Reaction Kit (PE Applied Biosystems) according to the manufacturer's manual. Then, the products were purified with DyeEx-columns (Qiagen, Hilden, Germany) and dried on a heat block. The dried specimens were dissolved in HiDi formamide and analyzed using ABI 3730 capillary electrophoresis (PE Applied Biosystems). Sequencing of both strands was performed to reduce ambiguities in sequence determination as recommended by the DNA Commission of the International Society for Forensic Genetics (Carracedo *et al*., 2000).

## Cloning and sequencing

For samples that provided ambiguous results by direct sequencing, a clone sequencing approach was used to define the exact sequence of each sample. Firstly, the ambiguous samples were correctly selected and then amplicons were purified by 6% (w/v) denaturing polyacrylamide gel electrophoresis (Acr: Bis=19:1, urea as denaturant). PCR products (530 bp) were cloned into pGem-T Easy vector system (Promega, Madison, WI, USA) and more than 20 clones were selected for each sample. Plasmid DNA was extracted from bacterial cells using QIAprep Spin Miniprep Kit (Qiagen). Purified plasmids were sequenced by the above mentioned sequencing method. To decrease the artificial generation of sequencing errors by Taq DNA polymerase during PCR, Pfu DNA polymerase (Promega, Madison, WI, USA) was used for amplification of samples with length heteroplasmy, and two independently generated PCR products from each sample were ligated into pGem-T Easy vector for cloning and randomly selected clones were sequenced in two directions by ABI 3730 capillary electrophoresis. For samples with greater C number ($n$>13), junction primers, which bind to a part of the C-stretch and the first 2~4 bases downstream of the C-stretch region, were used in the experiment to facilitate sequencing.

## Data analysis

The edited data were aligned with the revised Cambridge Reference Sequence (rCRS) (Andrews *et al*., 1999) using DNAStar software ver. 5.0. Haplotypes and gene diversities were estimated according to Nei (1987). The population diversity indices were estimated by Arlequin software ver. 2000 (Schneider *et al*., 2000) and Dispan computer program (Ota, 1993). The probability ($P$) of two randomly selected individuals from a population having identical haplotype types was $P=\sum f^2$ and gene diversity ($GD$) was calculated based on the following formula: $GD=n(1-\sum f^2)/(n-1)$ (Nei, 1996), in which $n$ stands for the sample number and $f$ is the frequency of each C-stretch haplotype (Stoneking *et al*., 1991). A comparative analysis between our haplotype data and the previously published data (Watson *et al*., 1996) was carried out by using Arlequin software ver. 2000 (Schneider *et al*., 2000) and Dispan computer program (Ota, 1993). The Africa data were obtained from

241 individuals from 9 ethnic groups at hospitals and rural medical clinics in Kenya, Nigeria, and Niger in Africa (Watson *et al*., 1996). According to the Fst genetic distances, the phylogenetic tree was constructed based on Neighbor Joining (NJ) method using the molecular evolutionary genetics analysis (MEGA) software ver. 4.0 (Tamura *et al*., 2007) and Dispan computer program (Ota, 1993). The principal component analysis (PCA) of the studied populations and the correlation study were performed by using SPSS 11.0 software.

## RESULTS

### C-stretch variation of HVS-I

MtDNA sequences from all the subjects were successfully obtained using both direct and clone sequencing approaches. In the HVS-I region, we observed 38 C-stretch haplotypes, and the most common one, which was identical with the Cambridge Reference Sequence, was AAAACCCCC TCCCC and was found in all Chinese ethnic groups (Table A1). Novel haplotypes and population specific haplotypes were also detected. The C-stretch haplotypes (AAAACCCCCACCCCC and AAAAGCCCC TCCCC) in HVS-I region, which contained a mutation of T to A transversion at position 16189 and a mutation of C to G transversion at position 16184, respectively, were observed for the first time. The haplotype (AAAGCCCCCTCCCC) was identified for the first time in Chinese ethnic groups (Fig.A1); and AAAACCCCTCCCCC and AACCCCCCT CCCCCC were population specific haplotypes detected in Chinese Tibetan and Salar ethnic groups. Approximately 25.14% of all the samples were found to show T to C transition at position 16189. The gene diversity values of HVS-I C-stretch haplotype in the studied ethnic groups were 0.3989, 0.7290, 0.6614, 0.4908, 0.5331, 0.6814, 0.5756 and 0.5803 across the various ethnic groups (Table A1). This finding has been deposited into the GenBank database (accession numbers: EU920130 to EU920667). The novel mutations have been submitted to MITOMAP (http://www.mitomap.org/), and the sequence numbers are from 20080730001 to 20080730003.

### Length heteroplasmy and correlation study of C-stretch

In the present study, C-stretch length heteroplasmy was detected in HVS-I region of approximately 9% of all the samples, which occurred exclusively in the samples with T to C transition at position 16189. For the samples with C-stretch length heteroplasmy, we chose the average C-length of all clones as the representative of C-length for these samples (Table 2 and Fig.2). For the samples with T to C transition at position 16189 in HVS-I region, we

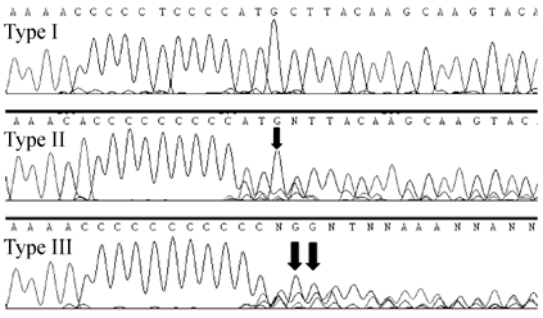


**Fig.2 Different kinds of sequence types. Arrows indicated dominant G**

**Table 2 Direct and cloning sequencing analyses of samples with length heteroplasmy**

| Sample | Predominant sequence type by direct sequencing | C7 | C8 | C9 | C10 | C11 | C12 | Total clone | Average C-length[a] | Sequence type[b] |
|---|---|---|---|---|---|---|---|---|---|---|
| A | C7 | 19 | 1 | 0 | 0 | 0 | 0 | 20 | 7.1 | I |
| B | C8 | 0 | 17 | 2 | 1 | 0 | 0 | 20 | 8.2 | I |
| C | C9 | 0 | 1 | 16 | 3 | 0 | 0 | 20 | 9.1 | I |
| D | C10 | 0 | 0 | 2 | 16 | 1 | 1 | 20 | 10.1 | I |
| E | C7/C8 | 13 | 6 | 1 | 0 | 0 | 0 | 20 | 7.4[d] | II |
| F | C9/C10 | 0 | 0 | 10 | 9 | 1 | 0 | 20 | 9.6[d] | III |
| G[c] | C8 | 0 | 20 | 0 | 0 | 0 | 0 | 20 | 8.0 | I |
| H[c] | C9 | 0 | 0 | 20 | 0 | 0 | 0 | 20 | 9.0 | I |

[a] Average C-length=(7×C7+8×C8+9×C9+10×C10+11×C11+12×C12)/20; [b] Categorized into three characteristic types (Types I, II, and III, as shown in Fig.2). Type I contains a predominant sequence type and no unreadable sequence. Type II shows a predominant sequence type and the unreadable sequence downstream. Type III contains two or more predominant sequence types and shows severely blurred sequences; [c] Cloning samples that were amplified and cloned again; [d] The values closed to 0.5 may be inconclusive results because two or more predominant types of C-length seem to be equally represent in direct sequencing

analyzed the correlation between the expansion of cytosine length in the C-stretch and the reduction of adenines upstream to it, and the result shows a significant correlation ($r$=−0.961, $P$<0.01).

**Phylogenetic analysis and principal component analysis of populations**

The Fst and dA genetic distance matrix of the eight Chinese populations and native African population is shown in Table 3. The two different genetic distances showed a very close correlation ($r$=0.853, $P$<0.01). All pairwise Fst comparisons between African and eight Chinese populations were significant with values ranging between 0.1599 (African vs Southern Han) and 0.3105 (African vs Northern Han). There were no significant differences of Tu vs Oroqen (*Fst*=0.0129, $P$>0.05), Tu vs Bonan (*Fst*=0.0089, $P$>0.05), Oroqen vs Bonan (*Fst*=0.0044, $P$>0.05), or Salar vs Southern Han (*Fst*=−0.0025, $P$>0.05), which indicates that they are closely related populations.

Based on the Fst value (Table 3), an unrooted NJ tree (Fig.3) of the eight Chinese populations and native African population was constructed. The eight Chinese populations showed a close affinity. The two Gansu populations (Yugur and Bonan) clustered together in the tree, the same as Tu and Salar (Qinghai populations). The genetic distances between Tibetan and the other Qinghai populations (Tu and Salar) were relatively far, which could be expected as the Tibetan population belong to a different language family (Sino-Tibetan family) (Table 1).

Fig.4 displays the principal component (PC) analysis map of the nine populations. The PC map represents the first two principal components. The first PC incorporated about 85.93% of the total original variation, while the second accounted for 11.23%. The cluster pattern in the PC map, with high diversity in northern origin Chinese ethnic groups (except Tibetan), was in good agreement with the

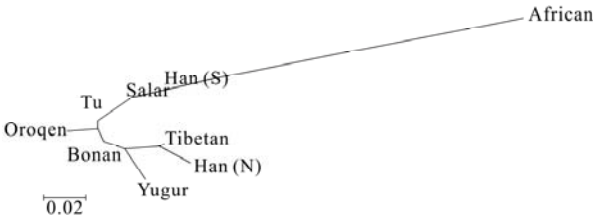


**Fig.3 NJ tree of nine populations based on Fst distance given in Table 3**
Han (N) represents Northern Chinese Han ethnic group; Han (S) represents Southern Chinese Han ethnic group

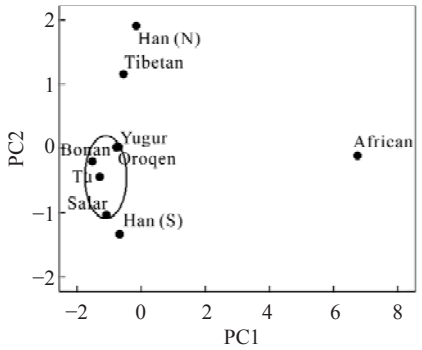


**Fig.4 Principal component (PC) map of nine populations based on the distance matrix**
The first PC incorporated about 85.93% of the total original variation, while the second accounted for 11.23%

**Table 3 Pairwise Fst and dA values between nine populations based on HVS-I C-stretch haplotype**

| | Han (N) | Han (S) | Tu | Tibetan | Oroqen | Salar | Yugur | Bonan | African |
|---|---|---|---|---|---|---|---|---|---|
| Han (N) | | 0.1266 | 0.1115 | 0.0929 | 0.1198 | 0.1128 | 0.1176 | 0.0678 | 0.3816 |
| Han (S) | 0.1015 | | 0.1690 | 0.2587 | 0.1803 | 0.0739 | 0.1847 | 0.0869 | 0.4479 |
| Tu | 0.0515 | 0.0255 | | 0.1236 | 0.1062 | 0.1444 | 0.2559 | 0.1128 | 0.4385 |
| Tibetan | 0.0192 | 0.0798 | 0.0191 | | 0.1682 | 0.2216 | 0.2363 | 0.1524 | 0.4220 |
| Oroqen | 0.0613 | 0.0427 | 0.0129* | 0.0510 | | 0.1687 | 0.3334 | 0.0859 | 0.5478 |
| Salar | 0.0738 | −0.0025* | 0.0178 | 0.0569 | 0.0375 | | 0.1821 | 0.0547 | 0.4566 |
| Yugur | 0.0373 | 0.0413 | 0.0412 | 0.0494 | 0.0685 | 0.0300 | | 0.1740 | 0.3360 |
| Bonan | 0.0267 | 0.0268 | 0.0089* | 0.0257 | 0.0044* | 0.0138 | 0.0222 | | 0.3766 |
| African | 0.3105 | 0.1599 | 0.1909 | 0.2609 | 0.2326 | 0.1810 | 0.2079 | 0.2190 | |

Fst values are shown below diagonal, and dA values above diagonal; *$P$>0.05

results of the NJ tree. The results are coincident with the geographic location and distribution of the language family.

## DISCUSSION

Due to slipping of the DNA polymerase during replication (Malik *et al*., 2002a; Lutz-Bonengel *et al*., 2004) or nuclear factor dysregulation (Malik *et al*., 2002b), the C-stretch exhibits significant variation among unrelated individuals. A recent study indicated that selective and neutral mechanisms might also be involved in the C-stretch length heteroplasmy in the mtDNA control region (Irwin *et al*., 2009). In the current study, we determined the C-stretch haplotype profile of HVS-I in eight Chinese ethnic groups. The results indicate that the most common haplotype of HVS-I was AAAACCCCCTCCCC, and some novel haplotypes, population specific haplotypes, and novel mutations were also detected (Fig.A1). The polymorphisms in HVS-I C-stretch resulted in relatively high values of gene diversity (*GD*), and low probability (*P*) values of two randomly selected individuals from a population having identical haplotype types in the eight Chinese ethnic groups. There was a significant correlation ($r=-0.961$, $P<0.01$) between the expansion of cytosine length and the reduction of upstream adenines in the C-stretch of HVS-I, supporting that sequence of cytosines vs adenosines was under some selective constraint (Howell and Smejkal, 2000). These data provide further evidence that the C-stretch could be a useful genetic marker in forensic identification where genomic DNA is degraded or absent in a variety of forensic samples such as loose hairs, teeth or bone. However, in this study, mtDNA C-stretch length heteroplasmy was observed in the blood sample from selected Chinese populations, but its ratio was much lower than those in other populations (Chen *et al*., 2002; Imaizumi *et al*., 2002; Mabuchi *et al*., 2007; Barbosa *et al*., 2008). To ensure reproducible results, we used the average C-length of all clones as the representative of C-length for the samples with C-stretch length heteroplasmy. This method was certified to be useful in defining the actual C-length variant of C-stretch, as the average C-length gave us comprehensive information of all the different sequence-length variations from the same sample, rather than the limited information from the predominant sequence type (Table 2).

The mtDNA nucleotide sequence evolves 6 to 17 times faster than nuclear DNA (Brown *et al*., 1982), and the C-stretch of the HVS-I region evolves even faster than the other regions of mtDNA (Malik *et al*., 2002a; Lutz-Bonengel *et al*., 2004). We hypothesize that the mtDNA HVS-I C-stretch might therefore be useful as a marker in population genetics and molecular evolution studies of closely related populations, as its rapid rate of change can more accurately reveal more acute evolutionary events in recent human history. In this study, we chose eight closely related Chinese ethnic groups, including Northern Han, Southern Han, Tu, Tibetan, Salar, Yugur, Bonan, and Oroqen ethnic groups, to investigate the utility of the C-stretch genetic marker. These ethnic groups live very close to each other (Fig.1), Northern Han (Xining and Xi'an), Tu, Tibetan, Salar, Yugur, Bonan ethnic groups live in Gansu and Qinghai Provinces, which are located in the northwest of China, and Oroqen and Han (Beijing) populations dwell in the northeast of China. All these ethnic groups belong to Sino-Tibetan or Altaic Language family (Table 1). The results from genetic matrix, phylogenetic tree, and PC map show highly coincident findings and indicate that the genetic relationships between the eight Chinese ethnic groups are very close, but relatively far from the Africans. A caveat here is that African populations are very diverse and the Africa sample data cited in this study were limited to certain regions of Africa (Kenya, Nigeria, and Niger). The Chinese ethnic minority groups, including Yugur, Bonan, Tu, and Sala, with the same language family, cluster together, while Tibetan ethnic group, which belongs to the Sino-Tibeta language family, was in another cluster. Ethnic groups within a similar geographic location (Yugur vs Bonan, Tu vs Salar) revealed a closer relationship. Therefore, the C-stretch variation is in agreement with the geographic location and distribution of the language family. These findings indicate that the C-stretch of mtDNA control region could be a reliable genetic marker when applied to population genetics and human evolution studies.

A broad spectrum of degenerative diseases (Yen *et al*., 1989; Cooper *et al*., 1992; Bowling *et al*., 1993; Brandon *et al*., 2005) involving the central nervous

system, heart, muscle, endocrine system, kidney and liver has been associated with mtDNA mutations, either base substitutions or insertion-deletions. The sequence-length variation of C-stretch may therefore affect mtDNA replication or transcription, and multiple diseases are associated with C-stretch variations (Tan *et al*., 2002; Lee HC *et al*., 2004; Meierhofer *et al*., 2004; Montanini *et al*., 2005; Sangkhathat *et al*., 2005). Thereby, the mtDNA C-stretch might also be considered a useful biological marker for human disease. In this study, we identified three different novel mutations, which were T to A transversion at position 16189, C to G transversion at position 16184, and A to G transition at position 16183. These novel mutations could be applied to a number of different research fields to further establish the relationships between mtDNA mutations and human disease.

When using the direct sequencing approach, the 'out-of-phase' nucleotide pattern (Fig.2), C-stretch could not easily be detected when compared to the other regions of mtDNA. Moreover, C-stretch length heteroplasmy further increases the difficulties of sequencing (Lee HY *et al*., 2004; Lutz-Bonengel *et al*., 2004; Bini and Pappalardo, 2005). These obstacles might affect the application of mtDNA C-stretch in forensic identification, population genetics, and other studies. Several sensitive detection methods (Lee HY *et al*., 2004; Lutz-Bonengel *et al*., 2004), such as direct sequencing, fluorescently labeled restriction fragment analysis, cloning, and fragment analysis, have been reported. Of these methods, cloning is less likely to be affected by artifacts and shows the highest accuracy and sensitivity in defining the exact sequence variants. In this study, we directly sequenced all the samples at first. Unfortunately, some samples were no amenable to this procedure, in particularly, those with the influence of C-continuous tract or length heteroplasmy. For the samples that were ambiguously typed by direct sequencing, a clone sequencing approach was used to define the exact sequence.

In summary, we have obtained genetic information regarding the variation of the C-stretch in mtDNA HVS-I region in eight Chinese ethnic groups. Both direct and clone sequencing approaches were employed in our study and three novel mutations, haplotypes, and population specific haplotypes were identified. The results indicate that the C-stretch could be a useful genetic maker in forensic identification and population genetics of Chinese populations.

## ACKNOWLEDGEMENT

We thank Zhan-hai Wang of Forensic Medical Identification Center of Qinghai Public Security Bureau of China for his assistance of collecting samples.

## References

Allen, M., Engstrom, A.S., Meyers, S., Handt, O., Saldeen, T., von Haeseler, A., Paabo, S., Gyllensten, U., 1998. Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. *J Forensic Sci.*, **43**(3):453-464.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., *et al*., 1981. Sequence and organization of the human mitochondrial genome. *Nature*, **290**(5806):457-465. [doi:10.1038/290457a0]

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N., 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**(2):147. [doi:10.1038/13779]

Barbosa, A.B., da Silva, L.A., Azevedo, D.A., Balbino, V.Q., Mauricio-da-Silva, L., 2008. Mitochondrial DNA control region polymorphism in the population of Alagoas state, north-eastern Brazil. *J. Forensic. Sci.*, **53**(1):142-146. [doi:10.1111/j.1556-4029.2007.00619.x]

Bini, C., Pappalardo, G., 2005. mtDNA HVI length heteroplasmic profile in different tissues of maternally related members. *Forensic. Sci. Int.*, **152**(1):35-38. [doi:10.1016/j.forsciint.2005.03.006]

Bowling, A.C., Mutisya, E.M., Walker, L.C., Price, D.L., Cork, L.C., Beal, M.F., 1993. Age-dependent impairment of mitochondrial function in primate brain. *J. Neurochem.*, **60**(5):1964-1967. [doi:10.1111/j.1471-4159.1993.tb13430.x]

Brandon, M.C., Lott, M.T., Nguyen, K.C., Spolim, S., Navathe, S.B., Baldi, P., Wallace, D.C., 2005. MITOMAP: a human mitochondrial genome database—2004 update. *Nucleic. Acids. Res.*, **33**(Database Issue):D611-D613. [doi:10.1093/nar/gki079]

Brown, W.M., Prager, E.M., Wang, A., Wilson, A.C., 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, **18**(4):225-239. [doi:10.1007/BF01734101]

Carracedo, A., Bar, W., Lincoln, P., Mayr, W., Morling, N., Olaisen, B., Schneider, P., Budowle, B., Brinkmann, B., Gill, P., *et al*., 2000. DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. *Forensic. Sci. Int.*, **110**(2):79-85. [doi:10.1016/S0379-0738(00)00161-4]

Chen, M.H., Lee, H.M., Tzen, C.Y., 2002. Polymorphism and heteroplasmy of mitochondrial DNA in the D-loop region in Taiwanese. *J. Formos. Med. Assoc.*, **101**(4):268-276.

Cooper, J.M., Mann, V.M., Schapira, A.H., 1992. Analyses of mitochondrial respiratory chain function and mitochondrial DNA deletion in human skeletal muscle: effect of ageing. *J. Neurol. Sci.*, **113**(1):91-98. [doi:10.1016/0022-510X(92)90270-U]

Du, R., Yip, V.F., 1993. Ethnic Groups in China. Science Press, Beijing, China.

Howell, N., Smejkal, C.B., 2000. Persistent heteroplasmy of a mutation in the human mtDNA control region: hypermutation as an apparent consequence of simple-repeat expansion/contraction. *Am. J. Hum. Genet.*, **66**(5):1589-1598. [doi:10.1086/302910]

Imaizumi, K., Parsons, T.J., Yoshino, M., Holland, M.M., 2002. A new database of mitochondrial DNA hypervariable regions I and II sequences from 162 Japanese individuals. *Int. J. Legal. Med.*, **116**(2):68-73. [doi:10.1007/s004140100211]

Ingman, M., Kaessmann, H., Pääbo, S., Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, **408**(6813):708-713. [doi:10.1038/35047064]

Irwin, J.A., Saunier, J.L., Niederstatter, H., Strouss, K.M., Sturk, K.A., Diegoli, T.M., Brandstatter, A., Parson, W., Parsons, T.J., 2009. Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. *J. Mol. Evol.*, **68**(5):516-527. [doi:10.1007/s00239-009-9227-4]

Kirches, E., Michael, M., Warich-Kirches, M., Schneider, T., Weis, S., Krause, G., Mawrin, C., Dietzmann, K., 2001. Heterogeneous tissue distribution of a mitochondrial DNA polymorphism in heteroplasmic subjects without mitochondrial disorders. *J. Med. Genet.*, **38**(5):312-317. [doi:10.1136/jmg.38.5.312]

Lee, H.C., Li, S.H., Lin, J.C., Wu, C.C., Yeh, D.C., Wei, Y.H., 2004. Somatic mutations in the D-loop and decrease in the copy number of mitochondrial DNA in human hepatocellular carcinoma. *Mutat. Res.*, **547**(1-2):71-78. [doi:10.1016/j.mrfmmm.2003.12.011]

Lee, H.Y., Chung, U., Yoo, J.E., Park, M.J., Shin, K.J., 2004. Quantitative and qualitative profiling of mitochondrial DNA length heteroplasmy. *Electrophoresis*, **25**(1):28-34. [doi:10.1002/elps.200305681]

Lee, H.Y., Chung, U., Park, M.J., Yoo, J.E., Han, G.R., Shin, K.J., 2006. Differential distribution of human mitochondrial DNA in somatic tissues and hairs. *Ann. Hum. Genet.*, **70**(Pt 1):59-65. [doi:10.1111/j.1529-8817.2005.00217.x]

Lutz-Bonengel, S., Sanger, T., Pollak, S., Szibor, R., 2004. Different methods to determine length heteroplasmy within the mitochondrial control region. *Int. J. Legal. Med.*, **118**(5):274-281. [doi:10.1007/s00414-004-0457-0]

Lutz-Bonengel, S., Schmidt, U., Sanger, T., Heinrich, M., Schneider, P.M., Pollak, S., 2008. Analysis of mitochondrial length heteroplasmy in monozygous and non-monozygous siblings. *Int. J. Legal. Med.*, **122**(4):315-321. [doi:10.1007/s00414-008-0240-8]

Ma, Y., 1994. China's Minority Nationalities. Foreign Languages Press, Beijing, China.

Mabuchi, T., Susukida, R., Kido, A., Oya, M., 2007. Typing the 1.1 kb control region of human mitochondrial DNA in Japanese individuals. *J. Forensic. Sci.*, **52**(2):355-363. [doi:10.1111/j.1556-4029.2006.00366.x]

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-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**(1):232-249. [doi:10.1086/302204]

Malik, S., Sudoyo, H., Pramoonjago, P., Sukarna, T., Darwis, D., Marzuki, S., 2002a. Evidence for the de novo regeneration of the pattern of the length heteroplasmy associated with the T16189C variant in the control (D-loop) region of mitochondrial DNA. *J. Hum. Genet.*, **47**(3):122-130. [doi:10.1007/s100380200013]

Malik, S., Sudoyo, H., Pramoonjago, P., Suryadi, H., Sukarna, T., Njunting, M., Sahiratmadja, E., Marzuki, S., 2002b. Nuclear mitochondrial interplay in the modulation of the homopolymeric tract length heteroplasmy in the control (D-loop) region of the mitochondrial DNA. *Hum. Genet.*, **110**(5):402-411. [doi:10.1007/s00439-002-0717-3]

Meierhofer, D., Mayr, J.A., Foetschl, U., Berger, A., Fink, K., Schmeller, N., Hacker, G.W., Hauser-Kronberger, C., Kofler, B., Sperl, W., 2004. Decrease of mitochondrial DNA content and energy metabolism in renal cell carcinoma. *Carcinogenesis*, **25**(6):1005-1010. [doi:10.1093/carcin/bgh104]

Montanini, L., Regna-Gladin, C., Eoli, M., Albarosa, R., Carrara, F., Zeviani, M., Bruzzone, M.G., Broggi, G., Boiardi, A., Finocchiaro, G., 2005. Instability of mitochondrial DNA and MRI and clinical correlations in malignant gliomas. *J. Neuro-Oncol.*, **74**(1):87-89. [doi:10.1007/s11060-004-4036-5]

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M., 1996. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.*, **30**(1):371-403. [doi:10.1146/annurev.genet.30.1.371]

Ota, T., 1993. DISPAN: Genetic Distance and Phylogenetic Analysis. Institute of Molecular Evolutionary Genetics, Pennsylvania State University, PA, USA.

Pfeiffer, H., Lutz-Bonengel, S., Pollak, S., Fimmers, R., Baur, M.P., Brinkmann, B., 2004. Mitochondrial DNA control region diversity in hairs and body fluids of monozygotic triplets. *Int. J. Legal. Med.*, **118**(2):71-74. [doi:10.1007/s00414-003-0409-0]

Salas, A., Lareu, M.V., Carracedo, A., 2001. Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report. *Int. J. Legal. Med.*, **114**(3):186-190. [doi:10.1007/s004140000164]

Sangkhathat, S., Kusafuka, T., Yoneda, A., Kuroda, S., Tanaka, Y., Sakai, N., Fukuzawa, M., 2005. Renal cell carcinoma

in a pediatric patient with an inherited mitochondrial mutation. *Pediatr. Surg. Int.*, **21**(9):745-748. [doi:10.1007/s00383-005-1471-0]

Schneider, S., Roessli, D., Excoffier, L., 2000. Arlequin: A Software for Population Genetic Data Analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Stoneking, M., Hedgecock, D., Higuchi, R.G., Vigilant, L., Erlich, H.A., 1991. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am. J. Hum. Genet.*, **48**(2):370-382.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: molecular evolutionary genetics analysis. *Mol. Biol. Evol.*, **24**(8):1596-1599. [doi:10.1093/molbev/msm092]

Tan, D.J., Bai, R.K., Wong, L.J., 2002. Comprehensive scanning of somatic mitochondrial DNA mutations in breast cancer. *Cancer Res.*, **62**(4):972-976.

Walsh, P.S., Metzger, D.A., Higuchi, R., 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques*, **10**(4):506-513.

Watson, E., Bauer, K., Aman, R., Weiss, G., von Haeseler, A., Paabo, S., 1996. mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.*, **59**(2):437-444.

Yen, T.C., Chen, Y.S., King, K.L., Yeh, S.H., Wei, Y.H., 1989. Liver mitochondrial respiratory functions decline with age. *Biochem. Biophys. Res. Commun.*, **165**(3):994-1003. [doi:10.1016/0006-291X(89)92701-0]

## APPENDIXES

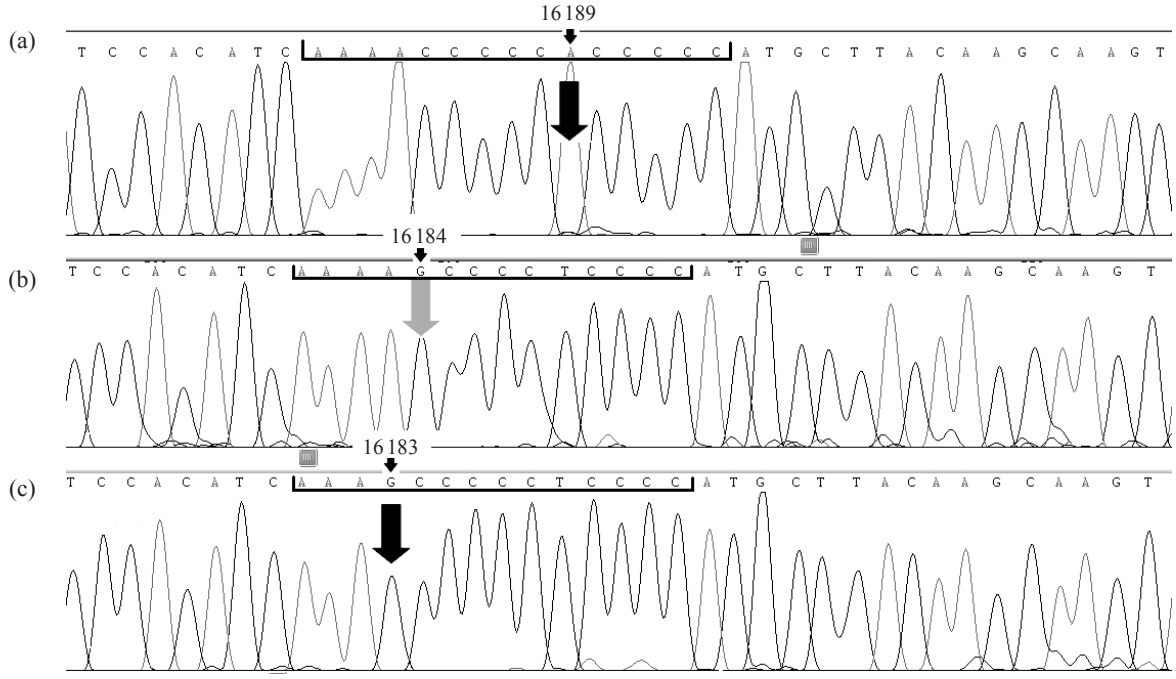


**Fig.A1 Novel C-stretch haplotypes detected in HVS-I region**

(a) AAAACCCCCACCCCC haplotype, which contained a mutation of T to A transversion at position 16189 and a C insertion at position 16193, was reported for the first time; (b) AAAAGCCCCTCCCC haplotype, which contained a mutation of C to G transversion at position 16184, was reported for the first time; (c) AAAGCCCCCTCCCC haplotype, which contained a mutation of A to G transition at position 16183, was reported for the first time in Chinese ethnic groups. The arrows in the figure indicated base position 16189, 16184, and 16183, respectively. The novel mutations have been submitted to MITOMAP (http://www.mitomap.org/), and the sequence numbers are 20080730001~20080730003. We also deposited our variation data to GenBank database (accession Numbers: EU920130~EU920667)

**Table A1 C-stretch haplotypes in HVS-I of eight Chinese ethnic groups**

| Haplotype | Population | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Han (N) (*n*=188) | Han (S) (*n*=117) | Tu (*n*=146) | Tibetan (*n*=120) | Oroqen (*n*=50) | Salar (*n*=100) | Yugur (*n*=100) | Bonan (*n*=98) |
| AAAACCCCCTCCCC | 0.7713 | 0.4701 | 0.5548 | 0.6917 | 0.6400 | 0.5300 | 0.6300 | 0.6327 |
| AAAACTCCCTCCCC | 0.0106 | 0.0171 | 0.0274 | 0.0583 | 0.0200 | 0.0300 | | 0.0612 |
| AAAATCCCCTCCCC | 0.0266 | 0.0171 | 0.0068 | 0.0083 | | 0.0200 | 0.1200 | 0.0306 |
| AAAACCCCCTCCTC | 0.0372 | 0.0769 | 0.0068 | 0.0083 | | 0.0300 | 0.1100 | 0.0306 |
| AAAAGCCCCTCCCC | | | 0.0068 | | | | | |
| AAAGCCCCCTCCCC | | | 0.0068 | | | | | |
| AAAACCCCCACCCCC | 0.0053 | | | | | | | |
| AAAACCCCCTCCAC | | 0.0085 | | | | | | |
| AAAACCCTCTCCCC | | | 0.0137 | | | | 0.0300 | |
| AAAACCCCTCCCTC | | | 0.0068 | | | | | |
| AAAACCCCTTCCCC | | | | | | | 0.0300 | |
| AAAACCCCCTCTCC | | | | 0.0083 | | | | |
| AAAACCCCCTCCCT | 0.0053 | | 0.0068 | | | | | |
| AAAACCCCTCCCCC | | | | 0.0167 | | | | |
| AAAACCCCCCCCTC | 0.0053 | | 0.0137 | 0.0167 | | | | |
| AAAACCCCCCCCCC | 0.0691 | 0.0598 | 0.1507 | 0.1750 | 0.1000 | 0.0600 | 0.0100 | 0.0510 |
| AAAATCCCCCCCCC | | 0.0085 | | | | | | |
| AAAACTCCCCCCC | | 0.0085 | | | | | | |
| AAAACTCCCCCCCC | | 0.0085 | | | | | | |
| AAAACCTCCCCCCC | 0.0053 | | | | | | | |
| AAAACCTCCTCCTCC | | | | | | | | 0.0102 |
| AAAACCCCCCTCCCC | | | 0.0068 | 0.0083 | | | | |
| AAAACCCCCTCCCCC | | | 0.0068 | | | | | |
| AAAACCCCCCCCCCC | | | 0.0342 | 0.0083 | | | | |
| AAAACCCCCCCCCCCC | | | 0.0137 | | | | | |
| AAAACCCCCCCCCCCCC | 0.0053 | | | | | | | |
| AAACCCCCCTCCCC | 0.0053 | | | | | | | |
| AAACCCCCCCCCCC | 0.0213 | 0.1709 | 0.0205 | | | 0.1700 | 0.0700 | 0.0408 |
| AAACCCCCCCCCCCC | | | 0.0068 | | | 0.0200 | | |
| AACCCCCCCCCCCC | 0.0266 | 0.1282 | 0.0959 | | 0.2400 | 0.0900 | | 0.1224 |
| AACCCCCCCTCCCC | | | 0.0068 | | | | | |
| AACCCCCCTCCCCC | | | | | | 0.0300 | | 0.0204 |
| AACCCCCCTCCCCCC | | | | | | 0.0200 | | |
| AACCCCCCCCCCC | | 0.0085 | | | | | | |
| AACCCCCCCCCCCCC | | | 0.0068 | | | | | |
| ACCCCCCCCCCCC | | 0.0085 | | | | | | |
| ACCCCCCCCCCCCC | | 0.0085 | | | | | | |
| ACCCCCCCCCCCCCC | 0.0053 | | | | | | | |
| Total | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| *GD* | 0.3989 | 0.7290 | 0.6614 | 0.4908 | 0.5331 | 0.6814 | 0.5756 | 0.5803 |
| *P* | 0.6032 | 0.2772 | 0.3431 | 0.5133 | 0.4776 | 0.3254 | 0.4302 | 0.4257 |