MEDICAL SCIENCE MONITOR

MOLECULAR BIOLOGY

e-ISSN 1643-3750
© Med Sci Monit, 2019; 25: 6482-6490
DOI: 10.12659/MSM.916275

Received: 2019.03.17
Accepted: 2019.06.03
Published: 2019.08.29

Whole Mitochondrial DNA Sequencing Analysis in 47 Han Populations in Southwest China

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

BCF 1 Lan Yao
BCD 2 Zhen Xu
AE 1 Lihua Wan

1 College of Basic Medicine, Chongqing Medical University, Chongqing, P.R. China
2 Key Laboratory of Forensic Genetics, Institute of Forensic Science, Ministry of Public Security, Beijing, P.R. China

Corresponding Author: Lihua Wan, e-mail: ccfw007@outlook.com
Source of support: Departmental sources

Background: Mitochondrial DNA (mtDNA) sequencing has been used in many areas, including forensic genetics. Due to the rapid development of sequencing technology, whole mtDNA sequencing is now possible and may be used in epidemiological and forensic studies. This study aimed to use whole mtDNA sequencing to investigate 47 Chongqing Han populations in southwest China and the diversity in the mtGenome reference data.

Material/Methods: The mtDNA of 47 Chongqing Han populations was generated using the Ion Torrent Personal Genome Machine (PGM) system. The extent of the effects of the mtDNA on the subpopulations was investigated and compared with six other populations from published studies. Pairwise fixation index (FST), a measure of population differentiation due to genetic structure, were calculated. Analysis of molecular variance (AMOVA) was performed, and 1257 hypervariable region data sets were added to the principal component analysis (PCA).

Results: The whole mtDNA sequencing data of 47 southwest Chinese Han populations were successfully recovered. Expanding the sequencing rage increased the discrimination power of mtDNA from three-times to 25-times based on different populations. The subpopulation effects showed 20 times the differences in match probability when compared with south China regions.

Conclusions: Whole mtDNA sequencing distinguished between individuals from 47 Chongqing Han populations in southwest China and has potential applications that include high-quality forensic identification.

MeSH Keywords: Forensic Genetics • Genetics, Population • Genome, Mitochondrial • High-Throughput Nucleotide Sequencing

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/916275

2486 7 2 40

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Background

Mitochondrial DNA (mtDNA) has an important role in forensic science, especially when nuclear DNA is scarce or degraded [1]. Forensic mtDNA analysis involves sequencing the 600 base pair (bp) hypervariable regions of mtDNA using Sanger sequencing. Considering that the greatest weakness of mtDNA in forensic applications is its relatively low discrimination power, whole mitochondrial genome (mtGenome) sequencing is an improved approach to individual identification in forensics [2,3]. With the advantage of the use of massively parallel sequencing (MPS) technology, the application of whole mtGenome sequencing in forensic practice is increasingly possible in the future.

Whole mtDNA analysis has been based on an established reference population database. In forensic practice, weight is assigned to the results of an mtDNA match comparison by estimating the frequency of the mtDNA haplotype given a relevant reference population, which depends on the size and the genetic structure of the population data [3]. To avoid unreliable estimation of the frequency of mtDNA profiles, forensic mtDNA databases should reflect the different genetic substructures of geographic regions and subpopulations [4,5]. In the past decade, phylogenetic studies have been conducted in China [6–10], but none of these studies have considered the requirements to establish a forensic reference database.

This study aimed to investigate the diversity in the mtGenome reference data within and between subpopulations in the Han Chinese population, which is the largest ethnic group in China that includes more than 1.3 billion individuals [11]. The boundary between the south and north Han populations was identified [12]. This study aimed to understand subpopulation diversity and level at which full mtGenome reference data should be maintained for the Han population.

In this study, for 47 Chongqing Han subpopulations, sequencing data were analyzed in combination with six published studies that contained MPS population data. The population data involving the mtGenome sequence were limited, and 1257 hypervariable region 1 (HV1) and HV2 (HV1/HV2) data from mainland China were also analyzed to investigate the geographic distribution pattern of mtDNA.

Material and Methods

Individuals studied and sequencing

Peripheral blood samples were obtained from 47 unrelated healthy individuals from the Han population from Chongqing city, which located in southwest China, were collected after informed consent was obtained. DNA was extracted using a MagAttract DNA Mini M48 Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. A long amplification sequencing strategy (8 Kb) was used for DNA library construction. Primers and amplification conditions were as previously described [13]. Sequencing libraries were constructed using the Ion Torrent Personal Genome Machine (PGM) system (Thermofisher Scientific, Waltham, MA, USA). The sequencing process on the PGM platform was performed according to the manufacturer's protocol, and two 318 chips (Thermofisher Scientific, Waltham, MA, USA) were loaded for sequencing.

Data acquisition

The keywords, China AND mtGenome OR complete genome, were used to search for population data in PubMed (*https://www.ncbi.nlm.nih.gov/pubmed*). Two publications were identified after filtering using the terms Homo sapiens, sampling method, and sequencing strategy. The two published population studies included 145 Han individuals from southern Shanghai were identified as CHSH [14], and 107 Han individuals from northern Liaoning province were identified as CHLN [15]. The variant call format (VCF) file of all Chinese data from the phase 3 data of the 1000 Genomes Project were downloaded (*ftp://ftp.ncbi.nlm.nih.gov/1000genomes/*) [16,17], which included individuals from Beijing who identified as CHBJ, Hunan province was identified as CHHN and Fujian province was identified as CHFJ, and the Dai population from Yunnan was identified as CDX.

Haplogroup assignment

Each haplogroup was identified and assigned using HaploGrep2 (PhyloTree Build 17) (*http://haplogrep.uibk.ac.at*) and EMPOP (V3/R11) (*https://empop.org*) [18,19]. Any discrepancies between these methods were manually reviewed. During the assignment and subsequent data analysis, point heteroplasmies, length variations, and cytosine insertions after 16193, 309, and 573 were ignored. In total, 16420 base pairs (bp) per sample were included.

Consensus sequences of CHCQ data from our laboratory were generated using UGENE [20] with BAM files. Tab-delimited format (hsd) files for HaploGrep2 were converted using the mtDNA profiler [21]. These data were also submitted to the EMPOP database for quality control. The VCF files of the 1000 Genomes Project data were directly input in HaploGrep2, but due to a previously reported error in input [18], all input data were manually rechecked. All the haplogroups from references were re-assigned following the above workflow. The related data of 1257 hypervariable region 1 (HV1) and HV2 (HV1/HV2) from references covering 13 populations across mainland China were reassigned haplogroups, based on the above methods.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Data analysis

The PGM data were analyzed using the Ion Torrent Software Suite version 5.0 with a variant caller plug-in version 5.0 and coverage analysis version 5.0 (Thermofisher Scientific, Waltham, MA, USA). Phylogenetic analysis was performed for quality control to highlight the potential sample mix or sequencing errors [22,23]. Single mutations and sites of concern from the variant caller file (VCF) report were rechecked using Integrative Genome Viewer (IGV) (Broad Institute, Boston, MA, USA).

From the seven populations, the summary statistics, including shared haplotypes, haplotype frequency, haplotype diversity, and the mean number of pairwise differences were calculated using Arlequin integrated population genetics software version 3.5.1.2 [24]. Random match probability (RMP) and confidence intervals (CIs) were calculated using Excel. Principal component analysis (PCA) was performed with SPSS version 19 software (IBM Corp., Armonk, NY, USA). Pairwise fixation index (FST), a measure of population genetic differentiation, was calculated, and analysis of molecular variance (AMOVA) was performed using Arlequin version 3.5.1.2 [24]. The statistical significance of FST values was estimated by permutation analysis using 10,000 permutations. FST evaluated the subpopulation effect on pairwise subpopulations according to the method described by Salas et al. [5]. The ratio of the corrected match probability to its value was calculated as F/f(h)+1, where F=FST and f(h) was the frequency of the shared haplotype. The size of the database was estimated as 2,000 individuals.

Results

Sequencing using the Ion Torrent Personal Genome Machine (PGM) system

There were 1.23G bases that were generated by two 318 chips. The average sequencing coverage of 47 samples reached 1281×, and 98% of sequencing reads were mapped to the reference. After the phylogeny study and IGV quality check, final CHCQ variants are listed in the Supplementary Table, and these passed EMPOP quality control (EMPOP00706) [Supplemantary/raw data available from the corresponding author on request].

Quality control

Most of the haplogroups were assigned consistently by two methods and discordant samples or samples that had HaploGrep scores less than 80 were manually examined. These discrepancies were mainly due to different phylogenic trees, as EMPOP version 3.0 used Phylotree Build 16, and HaploGrep2 used Build 17, and when they occurred, haplogroups were manually assigned to Build 17, as in sample HG00583. Secondly, in the VCF file from the 1000 Genomes Project, a specific type of input data, was transformed incorrectly as multiple nucleotide polymorphism (MNP), so the second single nucleotide polymorphism (SNP) was ignored. When the ignored variant was a diagnosis mutation, haplogroup assignment may have been affected. Those haplotypes were manually re-assigned, including the haplogroups in samples HG02396, HG00419, and HG00476.

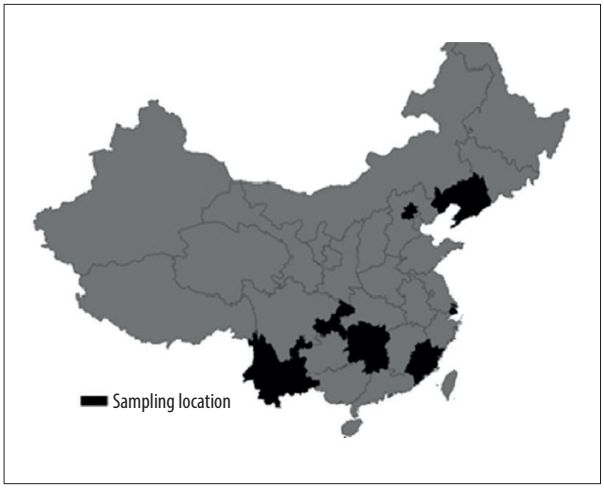

Figure 1. Geographical sampling locations. This map shows the sampling locations for mitochondrial DNA (mtDNA) genome analysis. The study included Han individuals from southern Shanghai (CHSH), northern Liaoning province (CHLN), Beijing (CHBJ), Hunan province (CHHN), and Fujian province (CHFJ), southern region (CHCQ), and the Dai population from Yunnan province (CDX). South Han includes CHHN, CHCQ, CHSH and CHFJ, and the latter two regions also belong to the east coast. North CHBJ is close to CHLN, which also belongs to eastern China.

Haplogroup distribution

The geographical relationships of all sampling locations are shown in Figure 1. In total, there were 600 haplotypes in the seven populations. The summary of the statistical analysis is shown in Table 1, and the basal haplogroup frequencies are listed in Table 2. All haplogroups were covered by the East and Southeast Asia-specific haplogroups [25,26]. Haplogroups D4, F1, B4, and M7 were the major haplogroups in the seven populations, covering 49.17% of all samples. Haplogroup frequencies were compared with previous studies [7,12,27]. Firstly, the super-haplogroup N was mainly found in the southern populations. To better analyze the data, we divided the haplogroups into three components, including northern dominated haplogroups (NDH), southern-dominated haplogroups (SDH) and other haplogroups, according to published data [12]. Generally, the southern populations were mainly represented by SDH. In northern populations, the NDH of CHLN and CHBJ accounted

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 1. Summary of the statistical analysis of the 600 haplotypes in the seven populations.

Sample	N	Haplotypes	MPD	RMP	HD	References
CHHN	70	69	37.0360±16.3016	0.0143	0.9716	1KGP
CHCQ	47	47	38.6966±17.1200	0.0213	0.9583	This study
CHBJ	103	103	36.2578±15.9043	0.0097	0.9807	1KGP
CHLN	107	107	37.8156±16.5690	0.0093	0.9814	Zhou et al. 2016
CHFJ	35	34	35.7714±15.9505	0.0286	0.9436	1KGP
CHSH	145	144	34.9153±15.2912	0.0069	0.9863	Ma et al. 2016
CDX	93	86	36.1351±15.8654	0.0108	0.9786	1KGP

MPD – mean pairwise differences; RMP – random match probability, calculated as 1- (sum of squared frequency); HD – haplotype diversity; 1KGP – the 1000 Genomes Project. The study included Han individuals from southern Shanghai (CHSH), northern Liaoning province (CHLN), Beijing (CHBJ), Hunan province (CHHN), and Fujian province (CHFJ), southern region (CHCQ), and the Dai population from Yunnan province (CDX).

Table 2. The basal haplogroup frequencies (percentages) in the seven populations.

	CDX	CHHN	CHCQ	CHSH	CHFJ	CHBJ	CHLN
A		2.86	17.02	4.14	11.43	6.80	7.48
B4	10.75	11.43	14.90	9.66	14.29	14.56	9.35
B5	10.75	8.57	4.26	2.07	2.86	0.97	2.80
B6	2.15						
C	5.38	1.43	6.38	3.45	2.86	4.85	9.35
D4	4.3	4.29	4.26	16.55	5.71	3.88	10.28
D4a		2.86	2.13	4.83	8.57	3.88	2.80
D4b	6.45	4.29	2.13	3.45	2.86	2.91	2.80
D5	1.08	2.86	2.13	6.21	5.71	9.71	4.67
D5b	1.08	1.43		0.69	2.86	1.94	1.87
D6		1.43					
F*	4.3	4.29	2.13	3.45		8.74	3.74
F1a	15.05	5.71	2.13	6.21	8.57	1.94	3.74
F1	5.38	8.57	12.77	4.14		4.85	5.61
G			4.26	4.83	5.71	4.85	6.54
M*	4.30	4.29	8.51	4.83	5.71	3.88	2.80
M7b	16.13	10	6.38	4.83	2.86	4.85	2.80
M7c	3.23	2.86	4.26	5.52	2.86	3.88	1.87
M8		1.43		4.83	2.86	4.85	10.28
M9a	1.08			1.38		0.97	2.80
M9b			2.13	1.38			
N*		1.43				0.97	
N9a		7.14	2.13	2.07	11.43	6.80	4.67
R*	4.30	4.29		1.38		1.94	0.93
R9	4.30	5.71	2.13	1.38			0.93
Z		2.86		1.38	2.86	0.97	1.87
Y				1.38			
K3						0.97	

The main haplogroups of seven populations that were reassigned according to PhyloTree Build 17. The study included Han individuals from southern Shanghai (CHSH), northern Liaoning province (CHLN), Beijing (CHBJ), Hunan province (CHHN), and Fujian province (CHFJ), southern region (CHCQ) and the Dai population from Yunnan province (CDX).

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 3. Pairwise fixation index (FST) values of population differentiation due to genetic structure, and p-values.

	CHHN	CHCQ	CHBJ	CHLN	CHFJ	CHSH	CDX
CHHN		–	–	+	–	+	–
CHCQ	0.00267		–	+	–	+	+
CHBJ	0.00563	0.0036		–	–	+	+
CHLN	0.01739	0.00862	0.00295		–	–	+
CHFJ	0.00613	0.00195	0.00587	0.00056		–	+
CHSH	0.01922	0.01709	0.00574	0.00209	0.00154		+
CDX	0.00279	0.01702	0.02324	0.03148	0.02542	0.03057	

+ p<0.05; – p>0.05. Permutation test, 10,000. The study included Han individuals from southern Shanghai (CHSH), northern Liaoning province (CHLN), Beijing (CHBJ), Hunan province (CHHN), and Fujian province (CHFJ), southern region (CHCQ), and the Dai population from Yunnan province (CDX).

Table 4. Analysis of molecular variance (AMOVA) percentage of variation and p-values.

Grouping	Among groups	Among populations within groups	Within populations
R.S. *vs.* R.N. (6 populations)	–0.28 (0.93±0.01)	0.86 (0.00±0.00)	99.42 (0.00±0.00)
G.E. *vs.* G.C. *vs.* G.W. (6 populations)	0.95 (0.07±0.01)	0.23 (0.09±0.02)	98.82 (0.00±0.00)
S.W. *vs.* S.E. *vs.* N. (6 populations)	0.53 (0.13±0.01)	0.26 (0.14±0.01)	99.21 (0.00±0.00)
R.S. *vs.* R.N.	0.27 (0.07±0.01)	0.83 (0.00±0.00)	98.9 (0.00±0.00)
G.E. *vs.* G.C. *vs.* G.W.	–0.1 (0.76±0.01)	1.02 (0.00±0.00)	99.08 (0.00±0.00)
C. *vs.* E. *vs.* S. *vs.* N. *vs.* N.W. *vs.* S.W. *vs.* N.E.	0.46 (0.02±0.00)	0.57 (0.00±0.00)	98.97 (0.00±0.00)

R.S. – rough south Han, including CHHN, CHCQ, CHFJ, CHSH, Yunnan, Guangdong, Shanghai, Hong Kong, and Hubei; R.N. – CHBJ, CHLN, Shandong, Henan, Jilin and Xinjiang; G.E. – general east, including CHSH, CHLN, CHFJ, Hong Kong, Shanghai, Guangdong, Liaoning, and Shandong; G.C. – central China, including CHHN, Henan, Hubei, and Jilin; G.W. – general west, including CHCQ, Yunnan, and Xinjiang; C. – CHHN, Hubei and Henan; E. – east, including CHSH, Shandong, Shanghai and CHFJ; S. – south Han, including Hong Kong, and Guangdong; N.W. – northwest Han, including Xinjiang; N.E. – northeast Han, including CHLN, Liaoning, and Jilin; S.W. – southwest Han, including CHCQ, and Yunnan; N. – north Han, including CHBJ.

for 62.62% and 50.49%, respectively. In the Dai population, SDH reached 78.49%, which was greater than its frequency in other the southern Han populations.

In south Han population, CHCQ had the highest frequency of haplogroup A out of all populations, where haplogroup A, D4, and F1 were predominated, which was supported by the finding from a previous study [28]. Haplogroup D6a was specific to CHHN, and the major haplogroups were similar to those previously reported [7,29]. CHSH had the highest frequency of haplogroup D4, which was previously reported to be high in northern China [30]. In CHFJ, haplogroups D and N9a comprised 37.14%. Compared with other southern regions, CHSH had a higher NDH component. As previously reported, haplogroups M7b and M71 were limited to southern China [10].

In the north, CHLN was dominated by haplogroups D4 and M8a, and these haplogroup frequencies have been previously reported [7,27]. For CHBJ, haplogroup D was the major haplogroup, as previously reported [31,32]. However, haplogroup K and N10 were observed specifically in CHBJ, with N10 having only been previously reported in southern China [11].

In the Dai population, CDX was 41.94% involved by haplogroups F1a, M7b, B5a, and this frequency was similar to the frequency of 34.15% previously reported [33]. However, haplogroups A and G were absent in the Dai population, and haplogroups B6a and M61 were specific to the Dai population when compared with other populations. The geographic haplogroup distributions of the mtGenome in this study were consistent with those that have previously been reported.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 5. Summary of the included Han subpopulation data.

References	Subpopulation	Abbreviation	Sample size
Xu et al. [34]	Henan	HEN	208
Yao et al. [35]	Shandong	SD	76
Nie et al. [36]	Shanghai	SH	200
Yao et al. [27]	Hubei Shandong	HUB	42
	Shandong	SD2	50
	Yunnan	YN	43
	Guangdong	GD2	30
	Liaoning	LN	51
Kivisild et al. [37]	Guangdong	GD	69
Chen et al. [38]	Guangdong	GD3	106
Irwin [39]	Hongkong	HK	377
Zhang et al. [40]	Jilin	JL	51

The references populations included in the analysis.

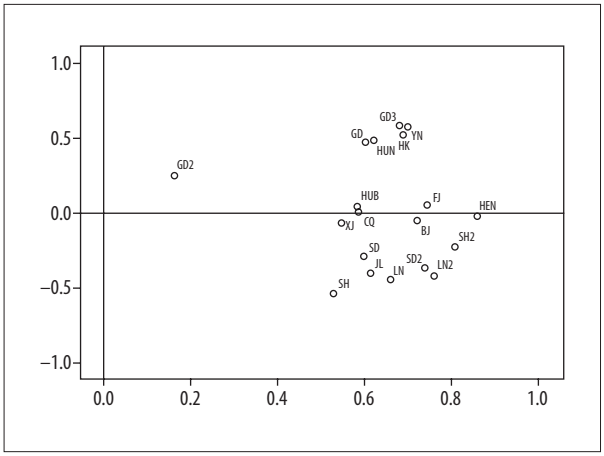

Figure 2. Principal component analysis (PCA) plot in the Han population. SH – Shanghai; HUB – Hubei; GD – Guangdong; FJ – Fujian; HEN – Henan; YN – Yunnan; CQ – Chongqing; SD – Shandong; LN – Liaoning; JL – Jilin; HK – Hong Kong; XJ – Xinjiang; BJ – Beijing.

Genetic relationships

Haplotypes were shared between CHSH and CDX (MK129=N004) and between CHSH and CHLN (HG02356=MK087). The pairwise fixation index (FST) values of population differentiation due to genetic structure and p-values are shown in Table 3. The FST values demonstrated that the Dai population was significantly different from most Han populations, except for CHHN, which was geographically closer to the Dai population. The difference between southern (CHHN and CHCQ), northern (CHLN), and southeastern (CHSH) subpopulations was significant in the Han population. Also, no significant differences were observed between geographically close regions, such as CHHN and CHCQ, CHBJ and CHLN, and CHFJ and CHSH. However, CHSH was different from other southern populations, such as CHHN and CHCQ. However, unlike CHLN, northern CHBJ did not show any significant differences with the southern Han population.

The analysis of molecular variance (AMOVA) results showed that the majority of variance came from within populations considering all regional groupings (Table 4). First, AMOVA was performed on the six Han populations. The variation within populations was much larger than that between groups, and the differentiation between north and south groups were not significant. The east and west groups were different, but these differences were not significant. To better understand the geographic patterns of mtDNA lineage distribution, 1,257 Han individuals, representing 13 populations, were added (Table 5). The variations between groups increased, and the east and west patterns were not as significant as before, which might be due to the larger scale of sampling or relatively low resolution of the partial data set.

Using principal components analysis (PCA), the Han population was generally divided into two groups, which were related to their sampling locations (Figure 2). The Han populations are distributed from southern, central, and northern China, with each area representing a different population. The southern Guangdong, Yunnan, Hong Kong, and Hunan populations were gathered together separately from the others. For Beijing and Shanghai, the northern Beijing area was located near the central regions, and southeastern Shanghai was close to the other northern regions. These populations represented the mixed characteristics of distinct areas. Also, inland regions, including Chongqing, Xinjiang, and Hubei, were located close to each other. In the populations included in this study, the differences were not significant between geographically close populations. In addition to the significant difference between south and north, an east and west difference was also observed, but some populations might include more geographically complex components.

Forensic parameters

Likelihood ratios (LRs) for unobserved haplotypes using two different methods based on the diverse sequencing range are listed in Table 6, including the recommended Clopper–Pearson method and the mtGenome Kappa method [3]. The LR_{KA} showed that the complete sequencing of mtDNA could increase the detection of rare haplotypes by more than 20-times. The Subpopulation effect of each local database was calculated

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 6. Hypervariable region 1 (HV1) and HV2 (HV1/HV2) data from 1257 cases from mainland China with the likelihood ratios calculated by the Clopper-Pearson method and the Kappa method.

	n	HV1		HV1/HV2		Mitochondrial (Mt) genome	
		LR_{CP}	LR_{KA}	LR_{CP}	LR_{KA}	LR_{CP}	LR_{KA}
CHHN	70	24	490	24	544	24	2450
CHCQ	47	16	522	16	1175	16	4700
CHBJ	103	35	530	35	758	35	10300
CHLN	107	36	520	36	818	36	10700
CHSH	145	49	568	49	1168	49	14500
CHFJ	35	12	205	12	613	12	613
CDX	93	32	192	32	333	32	721

LR_{CP} – likelihood ratio calculated by the Clopper-Pearson method. LR_{KA} – likelihood ratio calculated by the Kappa method. The study included Han individuals from southern Shanghai (CHSH), northern Liaoning province (CHLN), Beijing (CHBJ), Hunan province (CHHN), and Fujian province (CHFJ), southern region (CHCQ), and the Dai population from Yunnan province (CDX).

Table 7. Ratios of match probability in the seven populations.

	CHHN	CHCQ	CHBJ	CHLN	CHFJ	CHSH	CDX
CHHN	1						
CHCQ	3.65265	1					
CHBJ	6.63563	4.6036	1				
CHLN	18.40739	9.62862	3.95295	1			
CHFJ	7.13613	2.95195	4.87587	0.43944	1		
CHSH	20.23922	18.10709	6.74574	3.09209	1.08393	1	
CDX	3.79279	18.03702	24.26324	32.51148	26.44542	31.60057	1

Match probability of unobserved haplotype in different subpopulations. The study included Han individuals from southern Shanghai (CHSH), northern Liaoning province (CHLN), Beijing (CHBJ), Hunan province (CHHN), and Fujian province (CHFJ), southern region (CHCQ), and the Dai population from Yunnan province (CDX).

based on the previous pairwise FST values, as shown in Table 3. The ratio ranged from 0.44 to 20.24 in the Han population, which demonstrated that the match probability was 20-times different due to the database (Table 7). Also, the 20-fold difference came from the CHSH and CHHN, which are both located in southern China. The Dai population could reach a 30-times difference against other Han populations, although some geographically close regions might be more similar to each other.

Discussion

In the present study, using samples of DNA obtained from blood, two overlapped amplicons recovered complete 47 mitochondrial DNAs (mtDNAs) by sequencing using the Ion Torrent Personal Genome Machine (PGM) system and the accuracy of the sequencing data was validated. For the forensic reference population study, the complete sequencing of mtDNA could increase the discrimination power by between three-times to 25-times when compared with partial sequencing, and the subpopulation effects within both southern China regions could reach 20-times the difference of matched probability.

The basal haplogroup distribution of the CHCQ population was generally consistent with the previous studies. The inconsistent haplogroups might cause by the whole mtDNA sequencing, as in haplogroup N10 (750A-5581G-16172C-16362C), where half of the diagnosis variations were in the coding region, so that the complete mtDNA sequencing could benefit the haplogroup. Together with the 1257 hypervariable region 1 (HV1) and HV2 (HV1/HV2) data from mainland China, the significant differences ocurred among the northern, central, and southern regions, and the genetic substructures between the east coast and inland regions were also presented. The analysis

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

of molecular variance (AMOVA) results showed that the percentage variations were lower after the southern population was divided into southeast and southwest, and such grouping could better reflect the subpopulation genetic structure. Therefore, the reference population database only that included the southern and northern Han populations was not appropriate. However, the mtGenome diversity of different ethnic populations within provincial China remain to be evaluated.

Conclusions

This study aimed to use whole mitochondrial DNA (mtDNA) sequencing to investigate 47 Chongqing Han populations in southwest China and the diversity in the mtGenome reference data. The findings were based on the complete mtDNA sequencing data generated from sequencing using the Ion Torrent Personal Genome Machine (PGM) system. The diversity of mitochondrial DNA (mtDNA) of the southwest Chongqing Han population was identified and shown to have a potential application for forensic identification. However, when compared with the population of mainland China, the subpopulation effects of the Han population should be noted in establishing a reference population database with a high-quality dataset. This study has shown that the establishment of a whole mtDNA database should include local databases for each province and each ethnic group and that data from nearby regions may be used for quality control.

Conflict of interest

None.

References:

1. Parson W, Bandelt HJ: Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet, 2007; 1(1): 13–19
2. Coble MD, Just RS, O'Callaghan JE et al: Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med, 2004; 118(3): 137–46
3. Just RS, Scheible MK, Fast SA et al: Full mtGenome reference data: Development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. Forensic Sci Int Genet, 2015; 14: 141–55
4. Irwin JA, Saunier JL, Strouss KM et al: Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation. Forensic Sci Int Genet, 2007; 1(2): 154–57
5. Salas A, Bandelt HJ, Macaulay V, Richards MB: Phylogeographic investigations: The role of trees in forensic genetics. Forensic Sci Int, 2007; 168(1): 1–13
6. Yao YG, Nie L, Harpending H et al: Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. Am J Phys Anthropol, 2002; 118(1): 63–76
7. Wen B, Li H, Lu D et al: Genetic evidence supports demic diffusion of Han culture. Nature, 2004; 431(7006): 302–5
8. Zhao M, Kong QP, Wang HW et al: Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. Proc Natl Acad Sci USA, 2009; 6(50): 21230–35
9. Suo C, Xu H, Khor CC et al: Natural positive selection and north-south genetic diversity in East Asia. Eur J Hum Genet, 2012; 20(1): 102–10
10. Kong QP, Sun C, Wang HW et al: Large-scale mtDNA screening reveals a surprising matrilineal complexity in east Asia and its implications to the peopling of the region. Mol Biol Evol, 2011; 28(1): 513–22
11. Zhao YB, Zhang Y, Zhang QC et al: Ancient DNA reveals that the genetic structure of the northern Han Chinese was shaped prior to 3,000 years ago. PLoS One, 2015; 10(5): e0125676
12. Xue F, Wang Y, Xu S et al: A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. Eur J Hum Genet, 2008; 16(6): 705–17
13. Gunnarsdottir ED, Li M, Bauchet M et al: High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res, 2011; 21(1): 1–11
14. Ma K, Li H, Cao Y et al: Haplotype diversity in mitochondrial genome in a Chinese Han population. J Hum Genet, 2016; 61(10): 903–6
15. Zhou Y, Guo F, Yu J et al: Strategies for complete mitochondrial genome sequencing on Ion Torrent PGM platform in forensic sciences. Forensic Sci Int Genet, 2016; 22: 11–21
16. Abecasis GR, Altshuler D, Auton A et al: A map of human genome variation from population-scale sequencing. Nature, 2010; 467(7319): 1061–73
17. Auton A, Brooks LD, Durbin RM et al: A global reference for human genetic variation. Nature, 2015; 526(7571): 68–74
18. Weissensteiner H, Pacher D, Kloss-Brandstatter A et al: HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res, 2016; 44(W1): W58–63
19. Parson W, Dur A: EMPOP – a forensic mtDNA database. Forensic Sci Int Genet, 2007; 1(2): 88–92
20. Okonechnikov K, Golosova O, Fursov M: Unipro UGENE: A unified bioinformatics toolkit. Bioinformatics, 2012; 28(8): 1166–67
21. Yang IS, Lee HY, Yang WI, Shin KJ: mtDNAprofiler: A Web application for the nomenclature and comparison of human mitochondrial DNA sequences. J Forensic Sci, 2013; 58(4): 972–80
22. Bandelt HJ, Salas A, Lutz-Bonengel S: Artificial recombination in forensic mtDNA population databases. Int J Legal Med, 2004; 118(5): 267–73
23. Kong QP, Salas A, Sun C et al: Distilling artificial recombinants from large sets of complete mtDNA genomes. PLoS One, 2008; 3(8): e3016
24. Excoffier L, Lischer HE: Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour, 2010; 10(3): 564–67
25. Allard MW, Wilson MR, Monson KL, Budowle B: Control region sequences for East Asian individuals in the Scientific Working Group on DNA Analysis Methods forensic mtDNA data set. Leg Med (Tokyo), 2004; 6(1): 11–24
26. Kong QP, Yao YG, Sun C et al: Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet, 2003; 73(3): 671–76
27. Yao YG, Kong QP, Bandelt HJ et al: Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2002; 70(3): 635–51
28. Ji F, Sharpley MS, Derbeneva O et al: Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. Proc Natl Acad Sci USA, 2012; 109(19): 7391–96
29. Zhou HY, Wang HW, Tan SN et al: Genetic affinities of central China populations. Genet Mol Res, 2014; 13(1): 616–25
30. Zhao YB, Sun WY, Zhan Y et al: Mitochondrial DNA evidence of southward migration of Manchus in China. Mol Biol (Mosk), 2011; 45(5): 825–30
31. Gu M, Dong X, Shi L et al: Differences in mtDNA whole sequence between Tibetan and Han populations suggesting adaptive selection to high altitude. Gene, 2012; 496(1): 37–44
32. Jin HJ, Tyler-Smith C, Kim W: The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. PLoS One, 2009; 4(1): e4210
33. Yao YG, Zhang YP: Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: New data and a reappraisal. J Hum Genet, 2002; 47(6): 311–18
34. Xu K, Hu S: Population data of mitochondrial DNA HVS-I and HVS-II sequences for 208 Henan Han Chinese. Leg Med (Tokyo), 2015; 17(4): 287–94

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

35. Yao YG, Kong QP, Man XY et al: Reconstructing the evolutionary history of China: A caveat about inferences drawn from ancient DNA. Mol Biol Evol, 2003; 20(2): 214–19
36. Nie Y, Zhang C, Jiao H et al: Development of a multiplex PCR system of 59 mitochondrial SNPs and genetic analysis in Chinese population. Electrophoresis, 2014; 35(12–13): 1903–11
37. Kivisild T, Tolk HV, Parik J et al: The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol, 2002; 19(10): 1737–51
38. Chen F, Wang SY, Zhang RZ et al: Analysis of mitochondrial DNA polymorphisms in Guangdong Han Chinese. Forensic Sci Int Genet, 2008; 2(2): 150–53
39. Irwin JA, Saunier JL, Beh P et al: Mitochondrial DNA control region variation in a population sample from Hong Kong, China. Forensic Sci Int Genet, 2009; 3(4): e119–25
40. Zhang YJ, Xu QS, Zheng ZJ et al: Haplotype diversity in mitochondrial DNA hypervariable region I, II and III in northeast China Han. Forensic Sci Int, 2005; 149(2–3): 267–69

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]