SHORT COMMUNICATION

Positive selection on mitochondrial M7 lineages among the Gelong people in Hainan

Kun Yang[1,2], Hongxiang Zheng[2], Zhendong Qin[2], Yan Lu[2], Sara E Farina[2], Shilin Li[2], Li Jin[2], Dongna Li[1,3], Hui Li[2] and the Genographic Consortium[4]

Selections on human mitochondrial variations are difficult to examine. In this study, we found possible signs of selection on mitochondrial M7 lineages among the Gelong people who migrated from Guizhou to Hainan (the hottest province in China) throughout the last 1000 years. The genetic structure of the Gelong people shows an obvious sex-biased population admixture pattern with only 4.9% paternal contribution but 30.7% maternal contribution from indigenous Hlai people. According to frequency spectrum tests for deviation from neutrality and mismatch tests of demographic expansion, part of the maternal mitochondrial M7 lineages among the Gelong came from the Hlai had spread quickly and therefore might have undergone positive selection. In the future, whole mitochondrial genome sequencing might reveal the functional advantage of the M7 lineages.

Population admixture may cause various genetic effects, including different disease incidences. In some instances, a population may show different admixture patterns in paternal and maternal lineages, a condition that is called 'sex bias population admixture'.[1] Sex bias admixture may result from specific population histories, such as military migration or frequent intermarriage. A military migration might involve a mobile male soldier population that marries indigenous females, which results in two different lineage origins. Intermarriage between neighboring populations might also eventually alter maternal lineages. In a patriarchy, women usually move more frequently than men because of virilocal residence practices. Although military migration and intermarriage account for the majority of sex bias admixture, the Gelong population of Hainan is an exception.

The Gelong are a small population derived from the Gelao people in Guizhou, China. They arrived in Hainan, China, around 1000 years ago and kept a low frequency of genetic contact with the neighboring Han and Hlai people.[2] In this study, we investigated the mitochondrial DNA (mtDNA) diversity of 110 Gelong individuals. Our study was approved by the Fudan School of Life Sciences Ethics Committee. All subjects signed informed consent. We employed the same methods used in our previous study.[1,3,4] The hypervariable segments of mtDNA were sequenced and 23 coding regions of single-nucleotide polymorphisms (SNPs) were typed using restriction fragment length polymorphism (Supplementary Table 1). All sequences have been submitted to the GenBank (accession numbers: HM145965–6074). The haplogroups were determined according to the most updated nomenclature for mtDNA.[5]

The Gelong people show moderate genetic diversity (haplotype diversity[6] 0.968 ± 0.008, close to the average of southern China[3]) in mtDNA lineages (Figure 1a) and have high M7 lineage frequency (30%). Previous studies show that the Gelong originated in Gelao and were also influenced by the neighboring Han and Hlai in Y chromosome lineages.[2] Therefore, all the three populations may serve as genetic contributors for Gelong mtDNA lineages as well. To assess the possible sex bias admixture of the Gelong, we performed 5000 simulations for both Y chromosome and mtDNA data using Admix 2.0[7] (reference data are from our previous study,[2–4,8] including Hlai,[3] Gelao[3] and southern Han Chinese[4]). The admixture patterns show an obvious sex bias with only 4.9% paternal contribution and 30.7% maternal contribution from the Hlai. Also, there is less genetic material from the Gelao in maternal lineages (37.6%) than in paternal lineages (67.9%) (Figure 1b).

To further investigate which lineages were obtained from the Hlai, we compared the haplotype diversity of mtDNA among East Asian populations[1,3,4,9] using Network4.201.[10] The Gelong share several haplotypes with the Hlai in lineages M7e, R9c and R9b (Figure 2). There is a unique sublineage (locus 1d in Figure 2) of populations from Hainan: within M7e, only one haplotype was found among 14
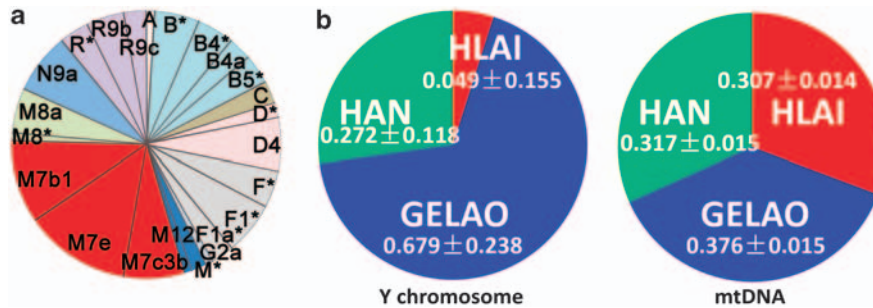
Figure 1 Mitochondrial DNA haplogroup frequencies in Gelong (**a**) and sex bias admixture pattern of Gelong (**b**). Detailed frequency data of each haplogroup are shown in Supplementary Table 1. The admixture analyses were done using Admix 2.0 (see Excoffier et al.[6]) running at the default settings of the program.
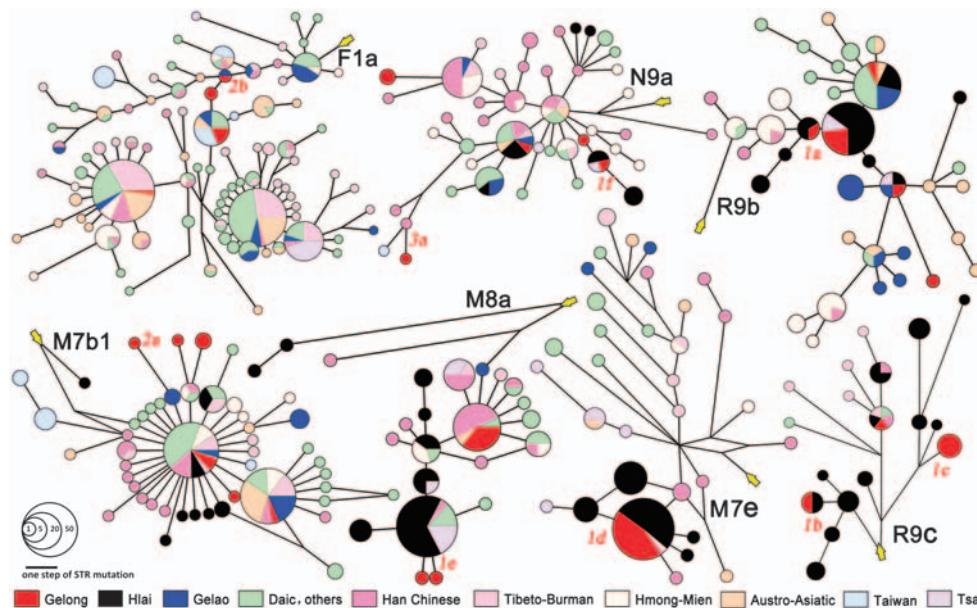

Figure 2 Networks of mtDNA exhibit the gene flow from Hlai to Gelong. The Network program was run at the default settings with the same weight for each SNP.[9] Node sizes are proportional to the individual numbers of haplotypes (see the lower left corner for scale). The lengths of the lines between nodes are proportional to the step numbers of the STR mutations. Locus 1a and 1b contain haplotypes shared only by the Gelong and Hlai. On locus 1c, the clade is shared only by the Gelong and Hlai. The clade on Locus 1d is a Hainan-specific clade; individuals originate from Hainan, including the Hlai, Gelong, Tsat, Hainan Han and Hainan Kimmun. The haplotypes on loci 1e and 1f are more widely shared among populations. Loci 2a and 2b show a close genetic relationship between the Gelong and Gelao.

Gelong individuals, whereas five were found among the Hlai. It is most likely that this haplotype was passed from the Hlai to the Gelong and spread quickly because of a genetic advantage. The high frequencies of the M7 lineages might also result from positive selection.

Selection signals on mtDNA were examined by frequency spectrum tests for deviation from neutrality. We employed a series of classical frequency spectrum tests, including Tajima's D,[11] Fu and Li's D, $D^\star$, F and $F^\star$[12,13] using DnaSP5.0[14] to detect deviation from neutrality (Table 1). Significant negative values of these tests indicate positive selection, purifying selection or demographic expansion effect; significant positive values imply balance selection or population subdivision effect. From neutral constant size simulation, we detected significant negative values for M7b1 and M7c3b lineages in the Gelong but not for M7e as was expected. Interestingly, some values for all lineages were also significantly negative, indicating that selection might have affected various lineages.

These significant values might be signals for either selection and/or demographic expansion. To distinguish the effect of demographic

expansion, we performed demographic coalescence simulations 10 000 times using a MS program.[15] Reference East Asian and African population models[16,17] were set as best-fit demographic models. The mutation rate was set as 3.24×10^{-6} per site per generation[18] (20 years per generation). According to the demographic history of Gelong, we assumed three scenarios for simulation (Supplementary Table 2). The $P=0.05$ significant cut-off values for each statistic are shown in Table 1. The previously observed significant values are still significant when compared with the simulation results of demographic expansion. Therefore, these values are most likely signals for selection.

To validate the authenticity of these signals for selection, we employed one more test, the mismatch test of demographic expansion,[19] using Arlequin3.11[6] to show the bias of the expansion effects among lineages. The less significant the result, the stronger the expansion effect on the lineage. The result of M7b1 and M7c3b is not significant ($P=0.128$), whereas those of total lineages ($P=0.011$) and all the non-M7 lineages ($P=0.006$) are significant, indicating that the M7 lineages expanded much faster than other lineages in the

Table 1 Frequency spectrum tests for deviation from neutrality

Lineages	Tajima's D	Fu and Li's D	Fu and Li's D*	Fu and Li's F	Fu and Li's F*
Observed					
Total	−1.463[a]	−1.751	−1.571	−1.958[a]	−1.831[a]
M7	−0.034	−1.862	−1.587	−1.485	−1.277
Non-M7	−1.296	−0.929	−0.799	−1.314	−1.196
M7e	−1.155	−1.477	−1.397	−1.618	−1.514
M7C3b	−1.448[a]	−1.942[a]	−1.565	−2.149[a]	−1.686
M7b1	−1.417	−2.010[a]	−1.522	−2.220[a]	−1.690
Simulated for demographic expansion					
Scenario A	−1.031	−1.267	−1.249	−1.254	−1.227
Scenario B	−1.276	−1.582	−1.485	−1.470	−1.407
Scenario C	−1.443	−1.465	−1.443	−1.577	−1.527

[a]Means $P<0.05$ when estimated by 10 000 coalescence simulations supposing neutral constant demographic sizes. A haplotype L3e is used as an out group. Some values are critical, such as −1.448, demonstrating that only weak signals were found for relevant lineages. We also applied these tests to 20 other Gelong lineages. No significant negative signals were found among other lineages, indicating that M7 lineages have undergone selective effects different from the other lineages. The significant signals found for the total lineages, M7b1 and M7c3b, were yielded by two or more tests, and therefore were unlikely caused by chance. Furthermore, if compared with the simulated results for demographic expansion, more values will be significant, including Fu and Li's D and D* values of the M7 lineage, further supporting the authenticity of the positive selection signals for the M7 lineages.

population. This might be another signal for the selective advantage for M7 lineages among Gelong people.

As a source of cellular energy, mtDNA may affect human's adaption to different climates. Therefore, some variants of mtDNA may have undergone selection. Hainan is the hottest province of China, a stark contrast to the mountain areas of Guizhou from where the Gelong originated. Tropical climates might serve as a selective force on mtDNA M7 lineages. However, in the phylogeny of mtDNA (http://www.phylotree.org/),[5] we found only three synonymous variations (C6455 T, T9824A and C4071 T) and one control region variation (T199C) in M7b1 and M7c3b. A future genome re-sequencing of mtDNA for these M7 samples is expected to reveal the functional relevance of M7 lineages.

1 Wen, B., Xie, X., Gao, S., Li, H., Shi, H., Song, X. *et al.* Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am. J. Hum. Genet.* **74,** 856–865 (2004).
2 Li, D., Sun, Y., Lu, Y., Mustavich, L. F., Ou, C., Zhou, Z. *et al.* Genetic origin of Kadai-speaking Gelong people on Hainan Island viewed from Y chromosomes. *J. Hum. Genet.* **55,** 462–468 (2010).
3 Li, H., Cai, X. Y., Winograd-Cort, E. R., Wen, B., Cheng, X., Qin, Z. D. *et al.* Mitochondrial DNA diversity and population differentiation in southern East Asia. *Am. J. Phys. Anthropol.* **134,** 481–488 (2007).
4 Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y. *et al.* Genetic evidence supports demic diffusion of Han culture. *Nature* **431,** 302–305 (2004).
5 van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mut.* **30,** E386–394 (2009).
6 Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **1,** 47–50 (2005).
7 Dupanloup, I. & Bertorelle, G. Inferring admixture proportions from molecular data: extension to any number or parental populations. *Mol. Biol. Evol.* **18,** 672–675 (2001).
8 Li, H., Wen, B., Chen, S. J., Su, B., Pramoonjago, P., Liu, Y. *et al.* Paternal relationships between Austronesian and Daic populations. *BMC Evol. Biol.* **8,** 146 (2008).
9 Wen, B., Li, H., Gao, S., Mao, X. Y., Gao, Y., Li, F. *et al.* Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol. Biol. Evol.* **22,** 725–734 (2005).
10 Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16,** 37–48 (1999).
11 Tajima, F. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
12 Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709 (1993).
13 Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147,** 915–925 (1997).
14 Librado, P., Rozas, J. & Dna, S. P. v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25,** 1451–1452 (2009).
15 Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
16 Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J. & Altshuler, D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15,** 1576–1583 (2005).
17 Helgason, A., Pálsson, S., Thorleifsson, G., Grant, S. F. A., Emilsson, V., Gunnarsdottir, S. *et al.* Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39,** 218–225 (2007).
18 Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A. *et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84,** 740–759 (2009).
19 Schneider, S. & Excoffier, L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152,** 1079–1089 (1999).

APPENDIX

The Genographic Consortium: Syama Adhikarla[1], Christina J Adler[2], Elena Balanovska[3], Oleg Balanovsky[3], Doron M Behar[4], Jaume Bertranpetit[5], Andrew C Clarke[6], David Comas[5], Alan Cooper[2], Clio SI Der Sarkissian[2], Matthew C Dulik[7], Christoff J Erasmus[8], Jill B Gaieski[7], ArunKumar GaneshPrasad[1], Wolfgang Haak[2], Angela Hobbs[8], Asif Javed[9], Matthew E Kaplan[10], Begoña Martínez-Cruz[5], Elizabeth A Matisoo-Smith[6], Marta Melé[5], Nirav C Merchant[10], R John Mitchell[11], Amanda C Owings[7], Laxmi Parida[9], Ramasamy Pitchappan[1], Daniel E Platt[9], Lluis Quintana-Murci[12], Colin Renfrew[13], Daniela R Lacerda[14], Ajay K Royyuru[9], Fabrício R Santos[14], Theodore G Schurr[7], Himla Soodyall[8], David F Soria Hernanz[15], Pandikumar Swamikrishnan[16], Chris Tyler-Smith[17], Kavitha Valampuri John[1], Arun Varatharajan Santhakumari[1], Pedro Paulo Vieira[18], Janet S Ziegle[19] and R Spencer Wells[15]

Affiliations for participants: [1]Madurai Kamaraj University, Madurai, Tamil Nadu, India. [2]University of Adelaide, South Australia, Australia.

[3]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia. [4]Rambam Medical Center, Haifa, Israel. [5]Universitat Pompeu Fabra, Barcelona, Spain. [6]University of Otago, Dunedin, New Zealand. [7]University of Pennsylvania, Philadelphia, PA, USA. [8]National Health Laboratory Service, Johannesburg, South Africa. [9]IBM, Yorktown Heights, NY, USA. [10]University of Arizona, Tucson, AZ, USA. [11]La Trobe University, Melbourne, Victoria, Australia. [12]Institut Pasteur, Paris, France. [13]University of Cambridge, Cambridge, UK. [14]Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [15]National Geographic Society, Washington, DC, USA. [16]IBM, Somers, NY, USA. [17]The Wellcome Trust Sanger Institute, Hinxton, UK. [18]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. [19]Applied Biosystems, Foster City, CA, USA.