ACADEMIA JOURNAL OF BIOLOGY 2022, 44(2): 21–28
DOI: 10.15625/2615-9023/17115

COMPLETE HUMAN mtDNA GENOME SEQUENCES REVEALED HAPLOTYPE FEATURES OF THE HMONG-MIEN LANGUAGE FAMILY IN VIETNAM

Dinh Huong Thao[1], Nong Van Hai[1,2], Nguyen Thuy Duong[1,2,*]

[1]Institute of Genome Research, VAST, Vietnam
[2]Graduate University of Science and Technology, VAST, Vietnam

Received 9 May 2022; accepted 17 June 2022

ABSTRACT

Vietnam is the homeland of 54 different ethnicities that belong to 5 major language families of the world, including Austroasiatic, Tai-Kadai, Hmong-Mien, Sino-Tibetan and Austronesian. Hmong-Mien, an ethnolinguistic family, presumably stemmed from Southern China and later spread to the Southeast Asia region. In this study, we analyzed the mitochondrial DNA sequences taken from 120 males belonging Hmong-Mien (HM) language family in Vietnam: Dao, Hmong, and Pathen, revealing 352 unique variants. Dao has the most number of polymorphisms (230 unique SNPs occurring 1469 times), followed by Hmong (181 unique SNPs occurring 1367 times) and Pathen (159 SNPs occurring 1243 times). Genetic variations within each population and among Hmong-Mien speakers were further measured by computations of haplotype diversity (H), nucleotide diversity (π) and fixation index (F_{ST}). There are nine major haplogroups (A, B, C, D, F, G, M, N9, and R) detected, with F and B making up over half of each population (Hmong: 56.09% (23/41), Pathen: 58.33% (21/36), Dao: 62.79% (27/43)). Haplotype classification was further divided into 30 haplogroups, of which 80% of them were specific to a single minority. Dao remained the most genetically diverse (H=0.957), while Pathen was the most homogeneous population (H=0.900). In terms of genetic distance, Dao and Hmong were more distinguished from each other, while Hmong and Pathen were more related. Complete mtDNA sequences of Viet HM speakers increased the mtDNA depository, improving the understanding of the genetic structure underlying this language family.

Keywords: Dao, Hmong, Pathen, mtDNA, Vietnam.

Citation: Dinh Huong Thao, Nong Van Hai, Nguyen Thuy Duong, 2022. Complete human mtDNA genome sequences revealed haplotype features of the Hmong-Mien language family in Vietnam. *Academia Journal of Biology*, 44(2): 21–28. https://doi.org/10.15625/2615-9023/17115

**Corresponding author email*: tdnguyen@igr.ac.vn

INTRODUCTION

Covering an approximate area of 331,690 km^2, Vietnam is divided into three regions: Bac Bo (north), Trung Bo (central) and Nam Bo (south). The country shares 4,616 km of borders with China to the North, Laos and Cambodia to the West (CIA, 2022). Vietnam is a long-established domicile to 54 different groups of people with ~98.8 million people (https://danso.org/viet-nam/; accessed May 2022), harboring much genetic information owing to its ethnical diversity. In terms of linguistics, the Vietnamese can be categorized into five major families: Austroasiatic (AA), Tai-Kadai (TK), Hmong-Mien (HM), Sino-Tibetan (ST) and Austronesian (AN) (Eberhard, 2021). Out of the five mentioned above, Hmong-Mien accounts for about 2.1% Vietnamese population. In earlier days, HM had been classified as a member of the ST group but now is recognized as a separate entity of its own (Taguchi, 2021). The homeland of HM can be traced back to the valley between the Yangtze and Mekong rivers, where historical and archaeological evidence of the proto Hmong-Mien language are identified (Ratliff, 2021). Modern HM speakers are widely dispersed over Southern China and mainland Southeast Asia (MSEA) regions (Mortensen, 2017). In Vietnam, the HM language family is spoken by 2.92 million people constituted of 3 minorities: Hmong, Dao and Pathen. They mostly live in the Northern mountainous regions, along the borders with Laos and China, with some Hmong residing in the central highland. As these ethnicities have deep roots in Vietnam, they are well-defined by their cultural elements such as language, social etiquette, architecture, and traditional practices. However, the biological aspects, which are no less informative and applicable, have not received sufficient attention, urging additional systematic research.

So far, the population genetics landscape in Vietnam is grounded on the variation data extracted from autosomal, Y-chromosome or mitochondrial DNA (mtDNA). To name a few, Li et al. (2007) underscored the high intra-population diversity of Vietnamese when juxtaposing its mtDNA dataset with that of the Chinese. Liu et al. (2020) explored single-nucleotide polymorphisms (SNPs) profiles of 22 ethnic groups, highlighting the dissimilarity between cultural and genetic diversities. In 2018, Duong et al. (2018) sequenced the whole mtDNA of 609 individuals, encompassing all five language families, revealing novel discoveries about the genetic compositions of the Vietnamese. While these studies unraveled the significance of the Vietnamese populations, their scopes were either on the major Kinh people or multiple language families. Understanding the various structure of Vietnamese people requires multifaceted approaches, including investigating sub-entities within the language family. In this study, we aimed to focus on the Hmong-Mien family by examining the molecular variants and haplotypic features of Dao, Hmong and Pathen from their complete mtDNA sequences. The findings here will provide a detailed stratification of these specific ethnolinguistic groups and expand the genetic knowledge of the Vietnamese population as a whole.

MATERIALS AND METHODS

Study subjects

Whole blood samples of 43 Dao, 41 Hmong, and 36 Pathen males were collected in Ha Giang and Dien Bien provinces of Vietnam. As declared in the written informed consent, all participants were biological unrelated and had at least three generations of ancestral of the same ethnicity. This study followed the protocol reviewed and received approval from the Institutional Review Board of the Institute of Genome Research, Vietnam Academy of Science and Technology (No: 2-2019/NCHG-HĐĐĐ).

Mitochondrial DNA sequencing and multiple alignments

Genomic DNA was isolated from peripheral blood samples using GeneJET Whole Blood Genomic DNA Purification Mini Kit (ThermoFisher Scientific, USA) following the manufacturer's procedure. The quality of DNA samples was assessed by

NanoDrop One/One machine. All DNA samples were stored in EDTA-containing tubes at -20 ^{o}C degrees. Construction of genomic libraries and capture-enrichment for mtDNA were performed using the method described by Maricic et al. (2010). Long-range PCR products were sonicated into 150–800 base-pair fragments and captured by bait-coated beads. After 48 hours, the enriched library pool was purified with carboxyl-coated magnetic beads (Agencourt AMPure XP, Agencourt, Beverly, MA, USA). Libraries were sequenced on Illumina platform. Reads were aligned to the Reconstructed Sapiens References Sequence (RSRS).

Haplotype analysis

The haplotype of each individual was assigned by Haplogrep2 and subjected to PhyloTree mtDNA tree Build 17 as the reference. A network of major haplogroups was built using the median-joining method on Network5 (https://www.fluxus-engineering.com/sharepub.htm#a1) and visualized by Network Publisher (Bandelt et al., 1999). The intra-population variation was assessed by haplotype diversity (H), and nucleotide diversity (π) (Nei, 1987). Using Arlequin 3.5.2.2, an analysis of molecular variance (AMOVA) was performed to establish the pairwise genetic distances (F_{ST}) (Excoffier & Lischer, 2010).

RESULTS

Full-length mtDNA sequences in the Vietnamese population at an average read depth of 840X discovered 352 unique SNPs (Fig. 1a), occurring 4,079 times on 120 participants. On the molecular scale, about a quarter of them (85/352) are repeatedly detected 3,015 times in all three ethnic groups (Fig. 1b). The numbers of unique SNPs overlapped within Vietnamese Hmong-Mien are 23, 6 and 19 in the following couples: Dao-Hmong, Hmong-Pathen and Dao-Pathen. The number of unique SNPs in the Dao, Hmong and Pathen are 230, 181 and 159, in which the percentages of these SNPs exclusive to the respective populations are 44.78% (103/230), 37.02% (67/181) and 30.82% (49/159). The numbers of times the unique SNPs appear in Dao, Hmong and Pathen are 1,469; 1,367 and 1,243; respectively.

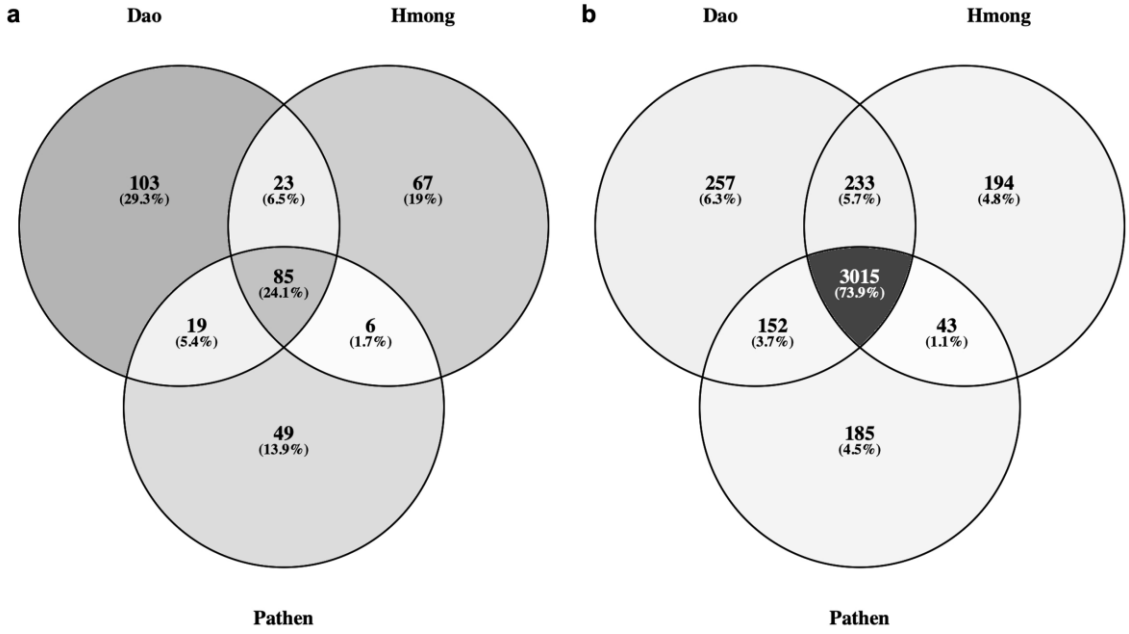

Figure 1. Distribution of variants among Vietnamese Hmong-Mien speakers, (a) Unique variants (352); (b) the total detected times of variants (4,079)

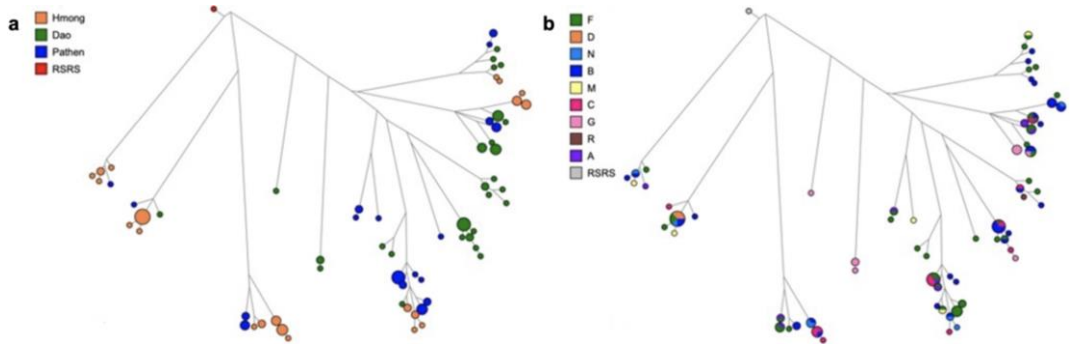

Figure 2. Median-joining networks of HM speakers with a color scheme based on a) ethnicities and b) major haplogroups (RSRS: Reconstructed Sapiens References Sequence)

Table 1. Distribution of haplogroups and their subclades within Hmong-Mien speaker populations

Haplogroup	Subhaplogroup	Dao (n = 43)	Hmong (n = 41)	Pathen (n = 36)	Total
F	F1a1a	7 (16.28%)	2 (4.88%)	1 (2.78%)	10
	F1a2	0	1 (2.44%)	0	1
	F1a4a1	1 (2.33%)	0	0	1
	F1c1a1	0	0	1 (2.78%)	1
	F1d	0	0	5 (13.89%)	5
	F1f	0	1 (2.44%)	0	1
	F1g1	1 (2.33%)	1 (2.44%)	0	2
	F2a	0	0	7 (19.44%)	7
	F2e	1 (2.33%)	0	0	1
	F3a1	1 (2.33%)	2 (4.88%)	5 (13.89%)	8
B	B4a1c4	3(6.98%)	0	0	3
	B4a5	3 (6.98%)	6 (14.63%)	0	9
	B4c2c	1 (2.33%)	0	0	1
	B4g2	1 (2.33%)	0	0	1
	B4h	3 (6.98%)	0	0	3
	B5a1c1a	1 (2.33%)	2 (4.88%)	0	3
	B5a1c1a1	0	8 (19.51%)	2 (5.56%)	10
	B6a	4 (9.3%)	0	0	4
C	C5d	0	5 (12.2%)	0	5
	C7a	4 (9.3%)	0	4 (11.11%)	8
N	N9a10+16311	0	7 (17.07%)	0	7
A	A14	0	0	9 (25%)	9
D	D4	0	3 (7.32%)	0	3
G	G*	6 (13.95%)	0	0	6
	G1a1	3 (6.98%)	0	0	3
M	M12a1a1	0	1 (2.44%)	0	1
	M74a	0	2 (4.88%)	0	2
	M7 b1a1a3	0	0	2 (5.56%)	2
R9	R9b1a3	2 (4.65%)	0	0	2
	R9c1b1	1 (2.33%)	0	0	1

Note: n: Number of individuals; the percentage is calculated based on the sample size (n) of each minority.

Our dataset can be classified into 30 haplogroups, in which 33% (9/30) are singleton, and 23.33% (7/30) are shared in at least two populations. All defined groups are encompassed by two macro-haplogroups, M and N, and both are considered indigenous to Eurasia (Palanichamy et al., 2004). While representatives of group M here comprise six haplogroups (C, G, M12, M74, M7 and D) and are responsible for 25% (30/120) of the samples, those of group N include five haplogroups (F, R9, A, B and N9) and takes the account for 75% (90/120) of the samples. Overall, F and B are the most prevalent among Vietnamese HM speakers, accounting for a total of 59.17% of mtDNA (71/120). A summary of haplogroup composition with respect to ethnicity is shown in Table 1, and the visualization of the haplotype network is demonstrated in Figure 2.

In terms of genetic variation, within each population, Dao people have the most diversity (H = 0.957, π = 0.0020), followed by Hmong (H = 0.925, π = 0.0019) and Pathen (H = 0.900, π = 0.0020). This analysis is in concordance with the mean number of variants per individual in each entity, as Dao has 34.07, Hmong has 33.2, and Pathen has 32.94. In general, the HM population in Vietnam is more homogeneous than that in Thailand (H = 0.97) and East Asia (H ranges: 0.928–1). Pairwise comparisons show that the most and least closely related in terms of genetic distance are Dao-Pathen (F_{ST} = 0.12205) and Dao-Hmong (F_{ST} = 0.25124), respectively.

DISCUSSION

A major haplogroup in East Asian, group F comes up frequently in Dao, Hmong and Pathen, corresponding to the percentages of 25.58% (11/43), 17.07% (7/41) and 52.78% (19/36**)** in each group. Two of F haplogroup, F1a1a and F3a1, are displayed in all three ethnicities. Previously reported in the Chinese Miao (Hmong) (Le et al., 2019) and French Guianese Hmong (Brucato et al., 2012), F3a1 also occurs at a high frequency in our Pathen (13.89%) (Brucato et al., 2012; Le et al., 2019). On the other hand, F1a1a, first discovered in HM East Asians (Wen et al., 2005) and later described in Hmong French Guianese (Brucato et al., 2012), is most prevalent in Dao (16.28%). Compared to the other HM population in East Asia, our samples share similar major haplogroups, yet differentiate based on their subhaplogroup composition and frequencies. In particular, B5a, being considered a highly homogeneous group that exists in 15/17 studied East Asian HM populations with a distribution ranging from 0 to 29.2% (Wen et al., 2005), was further classified into sub-haplogroups B5a1c1a (2.33% in Dao; 4.88% in Hmong) and B5a1c1a1 (19.51% in Hmong; 5.56% in Pathen) in our study. Given the fact that the previous research focused on D-loop regions of mtDNA, our approach of sequencing the complete mtDNA may provide more insights regarding haplotype classification.

Besides seven overlapping branches (B4a5, B5a1c1a, B5a1c1a1, C7a, F1a1a, F1g1 and F3a1) among HM speakers, it is notable that certain haplogroups are distinctive to each ethnicity (Table 1). In Dao, there are five major haplogroups (F, B, C, G and R), with B responsible for a substantial portion of 37.21% (16/43). Presumably, a native founder arising 50000 years before present (ybp), haplogroup B is characterized by a 9-bp deletion in mtDNA region V (Schurr & Wallace, 2002). Out of 16 Dao displaying haplogroup B, 25% of them have subhaplogroup B6a, which is absent in Thai and other Viet HM speakers but present in Austroasiatic Southern Mon-Khmer in Thailand (Kutanan et al., 2017). Group G*, a subhaplogroup belonging to macro-haplogroup M, is detected in 20.93% (9/43) of the population, in which 4 out of 9 retain sub-haplogroup G1a1, which is regarded to emerge from East Asian and has the highest frequency in Japanese (Tanaka et al., 2004). Similar to G, R9 only shows up in Dao (6.98%). Commonly detected in Southeast Asia, its descendants R9b1a3 (Hill et al., 2007; Hill et al., 2006) and R9c1b1 can be found in Taiwan and the Malay archipelago (Ko et al., 2014).

Hmong comprises six main haplogroups: B, C, D, F, M and N. About 12.2% of studied Hmong could be classified as C5d, a sub-haplogroup absent in Dao and Pathen. There was also another 14.63% (6/41) from M12a1a1, M74a, and D4, which are also descendants from macro-haplogroup M. In the N branch, N9a10+16311 appears exclusively in Hmong (17.07%). Interestingly, considered specific only to Hmonic people in Thailand, B4a5 is simultaneously observed in our Dao (6.98% (3/43)) and Hmong (14.63% (6/41)) participants (Kutanan et al., 2020). As HM speakers in Thailand consist of Hmong and Iumien branches, it is possible that the Viet HM speakers are more related to the former than the latter.

Pathen has a total of nine sub-haplogroups derived from five major haplogroups (F, B, C, A, M). The majority are predominated by group F (52.78% (19/36)), in which F1d (13.89% (5/36)) and F2a (19.44% (7/36)) are assigned solely to Pathen. Two haplogroups also exclusive to this minority are A14 (25% (9/36)) and M7b1a1a3 (5.56% (2/36)). B5a1c1a has a high frequency in HM people in Hunan, China, especially in Pathen (48%), thus suggesting being a signature haplogroup to the Hunan Pathen (Xia et al., 2019). In our dataset, B5a1c1a and its sub-haplogroup B5a1c1a1 show up in 2.33% (1/43) of Dao, 5.56% (2/36) of Pathen and take up a considerable portion of Hmong (24.39% (10/41)) in Vietnam.

CONCLUSION

Here, we report a total of 30 haplogroups from the complete mtDNA sequences of 120 HM individuals divided into three minorities Dao, Hmong, and Pathen. About 23.33% (7/30) of the defined sub-haplogroups appear in two or more populations, while the rest are unique to a single ethnicity. These haplogroups are present among HM speakers in other regions with differential distribution, emphasizing the similarity and divergence between HM speakers in Vietnam and other geographical locations.

Acknowledgements: We thank Prof. Mark Stoneking for making this work possible. We thank all sample donors for contributing to this research. This research was funded by the Ministry of Science and Technology, Vietnam (DTDL.CN-05/15) and by the Max Planck Society. Dinh Huong Thao was funded by Vingroup JSC and supported by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Institute of Big Data, code VINIF.2021.TS.153.

REFERENCES

Bandelt H. J., Forster P. and Rohl A., 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1): 37−48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Brucato N., Mazières S., Guitard E., Giscard P. H., Bois É., Larrouy G. and Dugoujon J. M., 2012. The Hmong Diaspora: Preserved South-East Asian genetic ancestry in French Guianese Asians. *Comptes Rendus Biologies*, 335(10−11): 698−707. https://doi.org/10.1016/j.crvi.2012.10.003

CIA, Vietnam, in The World Factbook, 2022. Central Intelligence Agency.

Duong N. T., Macholdt E., Ton N. D., Arias L., Schröder R., Van Phong N., Thi Bich Thuy V., Ha N. H., Thi Thu Hue H., Thi Xuan N., Thi Phuong Oanh K., Hien L. T. T., Hoang N. H., Pakendorf B., Stoneking M. and Van Hai N., 2018. Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia. *Scientific Reports*, 8(1): 11651. https://doi.org/10.1038/s41598-018-29989-0

Eberhard D. M. F., Simons C. D., 2021 G. F., S.l.: Sil International, global (Ethnologue: languages of asia).

Excoffier L. and Lischer H. E. L., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3):

564−567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Hill C., Soares P., Mormina M., Macaulay V., Clarke D., Blumbach P. B., Vizuete-Forster M., Forster P., Bulbeck D., Oppenheimer S. and Richards M., 2007. A Mitochondrial Stratigraphy for Island Southeast Asia. *The American Journal of Human Genetics*, 80(1): 29−43. https://doi.org/10.1086/510412

Hill C., Soares P., Mormina M., Macaulay V., Meehan W., Blackburn J., Clarke D., Raja J. M., Ismail P., Bulbeck D., Oppenheimer S. and Richards M., 2006. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Molecular Biology and Evolution*, 23(12): 2480−2491. https://doi.org/10.1093/molbev/msl124

Ko A. M. S., Chen C. Y., Fu Q., Delfin F., Li M., Chiu H. L., Stoneking M. and Ko Y. C., 2014. Early Austronesians: Into and Out Of Taiwan. *The American Journal of Human Genetics*, 94(3): 426−436. https://doi.org/10.1016/j.ajhg.2014.02.003

Kutanan W., Kampuansai J., Srikummool M., Kangwanpong D., Ghirotto S., Brunelli A. and Stoneking M., 2017. Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. *Human Genetics*, 136(1): 85−98. https://doi.org/10.1007/s00439-016-1742-y

Kutanan W., Shoocongdej R., Srikummool M., Hübner A., Suttipai T., Srithawong S., Kampuansai J. and Stoneking M., 2020. Cultural variation impacts paternal and maternal genetic lineages of the Hmong-Mien and Sino-Tibetan groups from Thailand. *European Journal of Human Genetics*, 28(11): 1563−1579. https://doi.org/10.1038/s41431-020-0693-x

Le C., Ren Z., Zhang H., Wang Q., Yang M., Liu Y., Huang J. and Wang J., 2019. The mitochondrial DNA control region sequences from the Chinese Miao population of southeastern China. *Annals of Human Biology*, 46(7-8): 606-609. https://doi.org/10.1080/03014460.2019.1694701

Li H., Cai X., Winograd-Cort E. R., Wen B., Cheng X., Qin Z., Liu W., Liu Y., Pan S., Qian J., Tan C. C. and Jin L., 2007. Mitochondrial DNA diversity and population differentiation in southern East Asia. *American Journal of Physical Anthropology*, 134(4): 481−488. https://doi.org/10.1002/ajpa.20690

Liu D., Duong N. T., Ton N. D., Van Phong N., Pakendorf B., Van Hai N. and Stoneking M., 2020. Extensive Ethnolinguistic Diversity in Vietnam Reflects Multiple Sources of Genetic Diversity. *Molecular Biology and Evolution*, 37(9): 2503-2519. https://doi.org/10.1093/molbev/msaa099.

Maricic T., Whitten M. and Pääbo S., 2010. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE*, 5(11): e14004. https://doi.org/10.1371/journal.pone.0014004

Mortensen D. R., 2017. Hmong-Mien Languages, in Oxford Research Encyclopedia of Linguistics., Oxford University Press.

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

Palanichamy M. G., Sun C., Agrawal S., Bandelt H. J., Kong Q. P., Khan F., Wang C. Y., Chaudhuri T. K., Palla V. and Zhang Y. P., 2004. Phylogeny of Mitochondrial DNA Macrohaplogroup N in India, Based on Complete Sequencing: Implications for the Peopling of South Asia. *The American Journal of Human Genetics*, 75(6): 966−978. https://doi.org/10.1086/425871

Ratliff M., 2021. 14 Classification and historical overview of Hmong-Mien languages. The Languages and Linguistics of Mainland Southeast Asia: A

comprehensive guide, 8, pp. 247. https://doi.org/10.1515/9783110558142

Schurr T. G. and Wallace D. C., 2002. Mitochondrial DNA Diversity in Southeast Asian Populations. *Human Biology*, 74(3): 431−452. https://doi.org/10.1353/hub.2002.0034

Taguchi Y., 2021 Historiography of Hmong-Mien linguistics, in The Languages and Linguistics of Mainland Southeast Asia, P. Sidwell and M. Jenny, Editors, De Gruyter. pp. 139−148.

Tanaka M., Cabrera V. M., González A. M., Larruga J. M., Takeyasu T., Fuku N., Guo L. J., Hirose R., Fujita Y., Kurata M., Shinoda K. I., Umetsu K., Yamada Y., Oshida Y., Sato Y., Hattori N., Mizuno Y., Arai Y., Hirose N., Ohta S., Ogawa O., Tanaka Y., Kawamori R., Shamoto-Nagai M., Maruyama W., Shimokata H., Suzuki R. and Shimodaira H., 2004. Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Research*, 14(10A): 1832−1850. https://doi.org/10.1101/gr.2286304

Wen B., Li H., Gao S., Mao X., Gao Y., Li F., Zhang F., He Y., Dong Y., Zhang Y., Huang W., Jin J., Xiao C., Lu D., Chakraborty R., Su B., Deka R. and Jin L., 2005. Genetic Structure of Hmong-Mien Speaking Populations in East Asia as Revealed by mtDNA Lineages. *Molecular Biology and Evolution*, 22(3): 725−734. https://doi.org/10.1093/molbev/msi055

Xia Z. Y., Yan S., Wang C. C., Zheng H. X., Zhang F., Liu Y. C., Yu G., Yu B. X., Shu L. L. and Jin L., 2019. Inland-coastal bifurcation of southern East Asians revealed by Hmong-Mien genomic history. *bioRxiv*, 730903. https://doi.org/10.1101/730903