International Journal of Legal Medicine
https://doi.org/10.1007/s00414-020-02290-4

POPULATION DATA

Close genetic relationship between central Thai and Mon people in Thailand revealed by autosomal microsatellites

Suparat Srithawong[1] · Kanha Muisuk[2] · Metawee Srikummool[3] · Jatupol Kampuansai[4,5] · Pittayawat Pittayaporn[6] · Sukhum Ruangchai[7] · Dang Liu[8] · Wibhu Kutanan[1]

Received: 7 February 2020 / Accepted: 30 March 2020

Abstract

Central Thailand is home to diverse populations with the central Thai constituting the major group, while the Mon, who migrated from southern Myanmar, are sparsely distributed within the region. A total of 338 individuals of eight central Thai (246 samples) and three Mon populations (92 samples) were newly genotyped. When combined with our previously published Mon data, this provides a total of 139 Mon samples. We found genetic similarity between the central Thai and Mon and weak sub-structuring among Thais from central, northern, and northeastern Thailand. The forensic parameter results show high discrimination values which are appropriate for forensic personal identification and paternity testing in both the central Thai and Mon; the probabilities of excluding paternity are 0.999999112 and 0.999999031, respectively, and the combined discrimination power is 0.99999999999999999999 in both groups. This regional allelic frequency on forensic microsatellites may serve as a useful reference for further forensic investigations in both Thailand and Myanmar.

Keywords Autosomal microsatellites · Allelic frequency · Central Thai · Mon

With the population as tallied in the most recent census falling at 65.9 million, Thailand is home to around 68 ethnic groups who speak five major languages: Kra–Dai (KD), Austroasiatic (AA), Sino–Tibetan, Hmong–Mein, and Austronesian [1]. KD is the most widely spoken in all regions of Thailand while the AA is the second most common language family. Spoken by ~ 29.72% of the population [1], Central Thai, the country's official language, also belongs to this family. In addition to central Thais, the central region of Thailand is also home to numerous ethnic groups. Among the most prominent are the AA-speaking Mon who settled in scattered enclaves throughout the region. However, the Thai Mon are not directed descendants of the ancient Monic populations in Thailand but are refugees that fled Myanmar during the sixteenth to nineteenth centuries A.D. [2]. The population of the Mon is ~ 100,000 in Thailand [1] and ~ 743,000 in Myanmar [3]. Previous mtDNA and Y chromosomal studies show contrasting maternal and paternal genetic histories of the major

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00414-020-02290-4) contains supplementary material, which is available to authorized users.

✉ Wibhu Kutanan
wibhu@kku.ac.th

[1] Department of Biology, Faculty of Science, Khon Kaen University, Mittapap Road, Khon Kaen, Thailand

[2] Department of Forensic Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[3] Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand

[4] Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

[5] Research Center in Bioresources for Agriculture, Industry and Medicine, Chiang Mai University, Chiang Mai, Thailand

[6] Department of Linguistics and Southeast Asian Linguistics Research Unit, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand

[7] Department of Physics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand

[8] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Published online: 13 April 2020

Thai groups in each region and also genetic difference within these major Thai groups. For the central Thai groups, either maternal genetic mixing with the AA-speaking Mon groups or cultural diffusion from the Mon on the paternal side is possible [4–6].

Microsatellites or short tandem repeats (STRs) on autosome show many advantages for both population genetic and forensic studies, i.e., distribution throughout the human genome, high mutation rate, and high polymorphism [7]. In Thailand, studies on forensic microsatellites have been focused on the north and northeast, leaving the central region understudied [8–10]. In this way, we newly genotype a total of 338 samples belonging to 11 populations, eight central Thai and three Mon (Fig. S1). The genomic DNA of CT1 to CT7 and all Mon populations were obtained from our previous studies [4, 5], while the CT8 samples were newly collected using buccal swab with informed consent. Ethical approval for this study was provided by Khon Kaen University and Naresuan University. Fifteen autosomal STR loci were amplified using a commercial AmpFlSTR Identifiler kit (Applied Biosystem, Foster City, CA, USA) according to manufacturer protocol. The amplicons were genotyped by multi-capillary electrophoresis on an ABI 3130 DNA sequencer (Applied Biosystem), and allele calling was performed by Gene Mapper software v.3.2.1 (Applied Biosystem). For genetic comparison analyses, we also retrieved 951 genotypic data of one Mon group and other relevant Thai and Indian populations from previous studies [8–11] (Fig. S1).

We used Arlequin v.3.5.2.2 [12] to calculate allele frequency, Hardy–Weinberg (HWE) *P* values, observed heterozygosity (H_o), expected heterozygosity (H_e), total alleles, and gene diversity (GD). We adjusted significance levels for the HWE according to the sequential Bonferroni correction ($\alpha = 0.05/15$) [13]. We used the Excel PowerStats spread sheet [14] to calculate several forensic parameters, including power of discrimination (PD), matching probability (MP), polymorphic information content (PIC), power of exclusion (PE), and typical paternity index (TPI) as well as the combined PD (CPD), combined MP (CMP), and combined PE (CPE). To characterize population affinity and genetic structure, the Arlequin was used to generate a genetic distance matrix based on a number of different alleles (F_{st}), and the matrix was then plotted in two dimensions by means of multidimensional scaling (MDS) using Statistica v.10 demo (StatSoft, Inc., USA). The Arlequin was also used to perform an analysis of molecular variance (AMOVA). The Bayesian clustering method implemented in STRUCTURE version 2.3.4 was performed under the following prior parameters: admixture, correlated allele frequencies, and assistance of sampling locations (LOCPRIOR model) [15–17]. We performed 10 replications for each number of clusters (*K*) from 1 to 12 and used a burn-in length of 100,000 iterations, followed by 200,000 iteration running length. The STRUCTURE Harvester [18] was employed to compute a second-order rate of change logarithmic probability between subsequent *K* values (ΔK) in order to identify the optimal *K* value in the data [19]. Then, CLUMPAK [20] was used to produce a single set resulting from 10 runs, and outputs from the CLUMPAK were graphically modified by DISTRUCT [21]. We used TreeMix v1.12 software [22] to construct maximum-likelihood trees with the parameters "-micro" for STR data and "-noss" for avoiding overcorrection of sample size differences. With zero to five migration events and 10 independent runs, we found the topology with five migration events with the highest likelihood and hence selected it for further investigation. To evaluate genetic relationship with other Asian populations, a neighbor-joining (NJ) tree based on F_{st} computation by allele frequency of 13 CODIS STR loci was carried out using POPTREE v.2 [23]. To estimate admixture coefficient (m_Y), we employed ADMIX 2.0 [24].

A total of 338 individual raw genotypes are provided in Table S1. Loci departure from the HWE, average H_e, total alleles, GD, and forensic parameters (CMP, CPE, and CPD) of 26 individual populations are shown in Table S2. To present allelic frequency figures for the 15 STR loci, we combined data from eight central Thai populations as a central Thai allelic frequency table (Table S3) and that of the four Mon populations as a Mon allelic frequency table (Table S4) because the Mon and central Thai showed low genetic variation (Table S5) and close genetic relatedness (Table S6) within their respective populations. The 15 loci were in agreement with the HWE ($P > 0.05$). A total of 142 alleles, varying from 5 alleles at *TPOX* to 17 alleles at *FGA*, were found in the Mon (Table S4), while there were 148 alleles, ranging from 6 alleles at *TH01* and *TPOX* to 18 alleles at *FGA*, in the central Thai (Table S3). Their allele frequencies varied from 0.0020 to 0.5408 for the central Thai and 0.0036 to 0.4779 for the Mon. The lowest H_e was observed at *TPOX* (0.6204) for the central Thai and 0.6669 for the Mon, while the highest H_e was the *FGA* for the central Thai (0.8837) and Mon (0.8738) (Tables S3 and S4). The total gene diversity of the central Thai group (0.7838 ± 0.3931) was higher than that of the Mon (0.7681 ± 0.3862) (Table S3 and S4). For the central Thai group, the PIC and MP ranged from 0.5649 (*TPOX*) to 0.8686 (*FGA*) and from 0.0283 (*FGA*) to 0.1894 (*TPOX*), respectively. The PD ranged from 0.8106 (*TPOX*) to 0.9717 (*FGA*), with a value of 0.999999999999999999999 for the combined PD. The PE ranged from 0.5649 (*TPOX*) to 0.7671 (*FGA*), with a combined PE value of 0.999999112. For the Mon, the PIC and MP ranged from 0.6098 (*TPOX*) to 0.8566 (*D2S1339*) and from 0.0383 (*D18S51*) to 0.1729 (*TPOX*), respectively. The PD ranged from 0.8271 (*TPOX*) to 0.9617 (*D2S1338*), with a value of 0.999999999999999999999 for the combined PD. The PE ranged from 0.4106 (*D13S317*) to 0.8586 (*FGA*), with a combined PE value of 0.999999031.

To further visualize the relationships, we constructed a MDS plot with two dimensions based on the F_{st} matrix which shows the genetic differentiation of four outlier populations positioned farther away from the central cloud of KD-speaking populations and four Mon groups (Fig. S2). However, there was some clustering of populations at the central cloud, albeit with some overlapping between them. There seems to be a separation between northeastern populations on the right side and northern populations on the other side in the first dimension. The central Thai groups overlapped the Mon and other KD-speaking groups along the second dimension (Fig. S2). To elucidate a cryptic population structure by the STRUCTURE, at $K = 3$, the suitable cluster (Fig. S3) [19], the first cluster detected was in the Indian population and is represented by the orange color, while the second and third clusters (purple and blue) stood out in the northeastern Thai populations and northern Thai populations, respectively, reflecting sub-structuring between these two major Thai groups (Figs. S1 and S4). The central Thai and Mon populations exhibited similar genetic components composed roughly equally of the northern Thai blue (30 to 56%) and the northeastern Thai purple (30 to 55%) and having the least percentages of the Indian orange (10 to 30%) (Table S7). When K was increased, the distinct components emerged in differentiated populations (Fig. S4), which are the outliers shown in the MDS results (Fig. S2). In AMOVA results, although the genetic variations among the two language families (AA and KD) (0.34%, $P < 0.01$) and three geographical areas (north/northeast/central) (0.19%, $P < 0.01$) were smaller than the variation among populations assigned to each group (1.00%, $P < 0.01$ for language and 0.48%, $P < 0.01$ for geography), significant linguistic and geographic differences ($P < 0.01$) could indicate that both language and geography do correspond to the genetic structure of populations (Table S5). Both the AMOVA and STRUCTURE results support genetic sub-structuring among the KD in each region of Thailand. To investigate population relationship and migration events, we generated a maximum likelihood tree with TreeMix which showed potential migrations, especially among northeastern Thai populations (both AA and KD groups), supporting assertions of a different genetic structure of the KD speaking northeastern Thai groups from the other regions (Fig. S5). The admixture results show that most of the central groups had received more contributions from the parental Mon (greater than 62.48%) than the parental Khmer and Tai (Table S8). The pooled central Thai and Mon were compared with other Asian populations by the NJ tree (Fig. S6). Genetic clustering between Asian populations (South Asian, East Asian, Mainland and Island Southeast Asian) is in concordance with their macro-geography. The central Thais are placed near the Laotian, Yuan, Vietnamese, and Mon. Interestingly, the Mon are located between the Burmese from Yangon and Mandalay, indicating genetic relationship of the Mon in Thailand and Burmese populations (Fig. S6).

In sum, the present study emphasizes that this set of markers is useful and reliable for the study of bio-anthropological processes at local scale. In general, the central Thai groups show more genetic similarity to the Mon than other populations, in accordance with previous mtDNA and Y chromosomal studies [5, 6]. That is, the central Thai populations could have originated from the Mon. The forensic microsatellite data newly generated here have strengthened the regional forensic database which is useful for further forensic investigation in both Thailand and Myanmar.

Acknowledgments We would like to thank all volunteers for donating their buccal cells and village chiefs for their participation. J.K. acknowledges the support provided by Chiang Mai University in Thailand.

Authors' contributions S.S., M.S., J.K., S.R., P.P., and W.K. collected the samples. S.S., M.S., and K.M. extracted the DNA and performed genotyping. S.S. and D.L. analyzed the data. S.S. drafted the first manuscript with input from all authors. W.K. designed the project and drafted and edited the manuscript.

Funding information S.S. was supported by Khon Kaen University under a research fund for supporting lecturers to admit high-potential students to study and research (592T224). W.K. was supported by the Thailand Research Fund (RSA6180058).

Data availability The raw genotyped data are reported in the Supplementary Materials.

Compliance with ethical standards

Competing interests The authors declare that they have no conflict of interest.

Ethics approval Ethical approval for this study was provided by Khon Kaen University and Naresuan University.

References

1. Simons GF, Fennig CD (2018) Ethnologue: languages of the world. Dallas (TX), SIL International
2. Ocharoen S (1998) Mon in Thailand. Bangkok (in Thai). the Thailand Research Fund, Bangkok, Thailand
3. Eberhard DM, Gary FS, Charles DF (2019) Ethnologue: languages of the world. Dallas (TX), SIL International
4. Kutanan W, Kampuansai J, Srikummool M et al (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. Hum Genet 1:85–98
5. Kutanan W, Kampuansai J, Brunelli A, Ghirotto S, Pittayaporn P, Ruangchai S, Schröder R, Macholdt E, Srikummool M, Kangwanpong D, Hübner A, Arias L, Stoneking M (2018) New insights from Thailand into the maternal genetic history of mainland Southeast Asia. Eur J Hum Genet 26(6):898–911
6. Kutanan W, Kampuansai J, Srikummool M, Brunelli A, Ghirotto S, Arias L, Macholdt E, Hübner A, Schröder R, Stoneking M (2019) Contrasting paternal and maternal genetic histories of Thai and Lao populations. Mol Biol Evol 36(7):1490–1506

7. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
8. Kutanan W, Kampuansai J, Colonna V, Nakbunlung S, Lertvicha P, Seielstad M, Bertorelle G, Kangwanpong D (2011) Genetic affinity and admixture of northern Thai people along their migration route in northern Thailand: evidence from autosomal STR loci. J Hum Genet 56:130–137
9. Srithawong S, Srikummool M, Pittayaporn P, Ghirotto S, Chantawannakul P, Sun J, Eisenberg A, Chakraborty R, Kutanan W (2015) Genetic and linguistic correlation of the Kra–Dai–speaking groups in Thailand. J Hum Genet 60:371–380
10. Srithawong S, Muisuk K, Srikummool M et al (2020) Genetic structure of the ethnic Lao groups from mainland Southeast Asia revealed by forensic microsatellites. Ann Hum Genet 1–13. https://doi.org/10.1111/ahg.12379
11. Seah LH, Jeevan NH, Othman MI, Jaya P, Ooi YS, Wong PC, Kee SS (2003) STR data for the AmpFlSTR Identifiler loci in three ethnic groups (Malay, Chinese, Indian) of the Malaysian population. Forensic Sci Int 138:134–137
12. Excoffier L, Lischer L (2010) Arlequin suitever 3.5.: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567
13. Rice WR (1989) Analyzing tables of statistical tests. Evolution 43: 223–225
14. Promega (1999) Powerstats version 1.2. tools for analysis of population statistics. https://www.promega.com.cn/products/geneticidentity. Accessed 1 July 2018
15. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
16. Falush D, Stepheas M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:156–187
17. Hubisz M, Falush D, Stephens M, Pritchard J (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9:1322–1332
18. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361
19. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620
20. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 15(5):1179–1191
21. Rosenberg NA (2003) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138
22. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome–wide allele frequency data. PLoS Genet 8(11):e1002967
23. Takezaki N, Nei M, Tamura K (2014) POPTREEW: webersion of POPTREE for constructing population trees from allele frequency data and computing some other quantities. Mol Biol Evol 31(6): 1622–1524
24. Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. Mol Biol Evol 15:1298–1311