European Journal of Human Genetics
https://doi.org/10.1038/s41431-020-0693-x

ARTICLE

Cultural variation impacts paternal and maternal genetic lineages of the Hmong-Mien and Sino-Tibetan groups from Thailand

Wibhu Kutanan[1,2] · Rasmi Shoocongdej[3] · Metawee Srikummool[4] · Alexander Hübner[2] · Thanatip Suttipai[1] · Suparat Srithawong[1] · Jatupol Kampuansai[5,6] · Mark Stoneking[2]

Received: 8 February 2020 / Revised: 17 June 2020 / Accepted: 26 June 2020

Abstract

The Hmong-Mien (HM) and Sino-Tibetan (ST) speaking groups are known as hill tribes in Thailand; they were the subject of the first studies to show an impact of patrilocality vs. matrilocality on patterns of mitochondrial (mt) DNA vs. male-specific portion of the Y chromosome (MSY) variation. However, HM and ST groups have not been studied in as much detail as other Thai groups; here we report and analyze 234 partial MSY sequences (~2.3 mB) and 416 complete mtDNA sequences from 14 populations that, when combined with our previous published data, provides the largest dataset yet for the hill tribes. We find a striking difference between Hmong and IuMien (Mien-speaking) groups: the Hmong are genetically different from both the IuMien and all other Thai groups, whereas the IuMien are genetically more similar to other linguistic groups than to the Hmong. In general, we find less of an impact of patrilocality vs. matrilocality on patterns of mtDNA vs. MSY variation than previous studies. However, there is a dramatic difference in the frequency of MSY and mtDNA lineages of Northeast Asian (NEA) origin vs. Southeast Asian (SEA) origin in HM vs. ST groups: HM groups have high frequencies of NEA MSY lineages but lower frequencies of NEA mtDNA lineages, while ST groups show the opposite. A potential explanation is that the ancestors of Thai HM groups were patrilocal, while the ancestors of Thai ST groups were matrilocal. Overall, these results attest to the impact of cultural practices on patterns of mtDNA vs. MSY variation.

Introduction

Thailand occupies the center of Mainland Southeast Asia (MSEA) and shares borders with several countries: Laos and Myanmar in the North and West, Laos in the Northeast, Cambodia in the East, and Malaysia in the South (Fig. 1). With a census size of ~68.61 million in 2018, there are 73 different recognized languages belonging to five different linguistic families: Tai-Kadai (TK, 89.4%), Austroasiatic (AA, 4.0%), Sino-Tibetan (ST, 3.2%), Austronesian (AN, 2.8%), and Hmong-Mien (HM, 0.2%) [1]. Archaeological evidence indicates the presence of modern humans in the area of Thailand since the late Pleistocene [2]. More recently, archaeogenetics studies indicate that modern AA-speaking groups in Southeast Asia (SEA) are descended from a dispersal of Neolithic farmers from southern China that occurred ~4000 years ago (kya) [3, 4]. Archaeological and linguistic evidence supports the presence of AA people by at least 2.5 kya, while a later expansion of the TK-speaking groups from southern China reached present-day Thailand ~2 kya [2, 5]. And, historical evidence indicates that the homeland of many ST and HM speaking hill tribe

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-020-0693-x) contains supplementary material, which is available to authorized users.

✉ Wibhu Kutanan
wibhu@kku.ac.th

✉ Mark Stoneking
stoneking@eva.mpg.de

[1] Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen 40002, Thailand

[2] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

[3] Department of Archaeology, Faculty of Archaeology, Silpakorn University, Bangkok 10200, Thailand

[4] Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok 65000, Thailand

[5] Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai 50202, Thailand

[6] Research Center in Bioresources for Agriculture, Industry and Medicine, Chiang Mai University, Chiang Mai 50202, Thailand

Published online: 20 July 2020

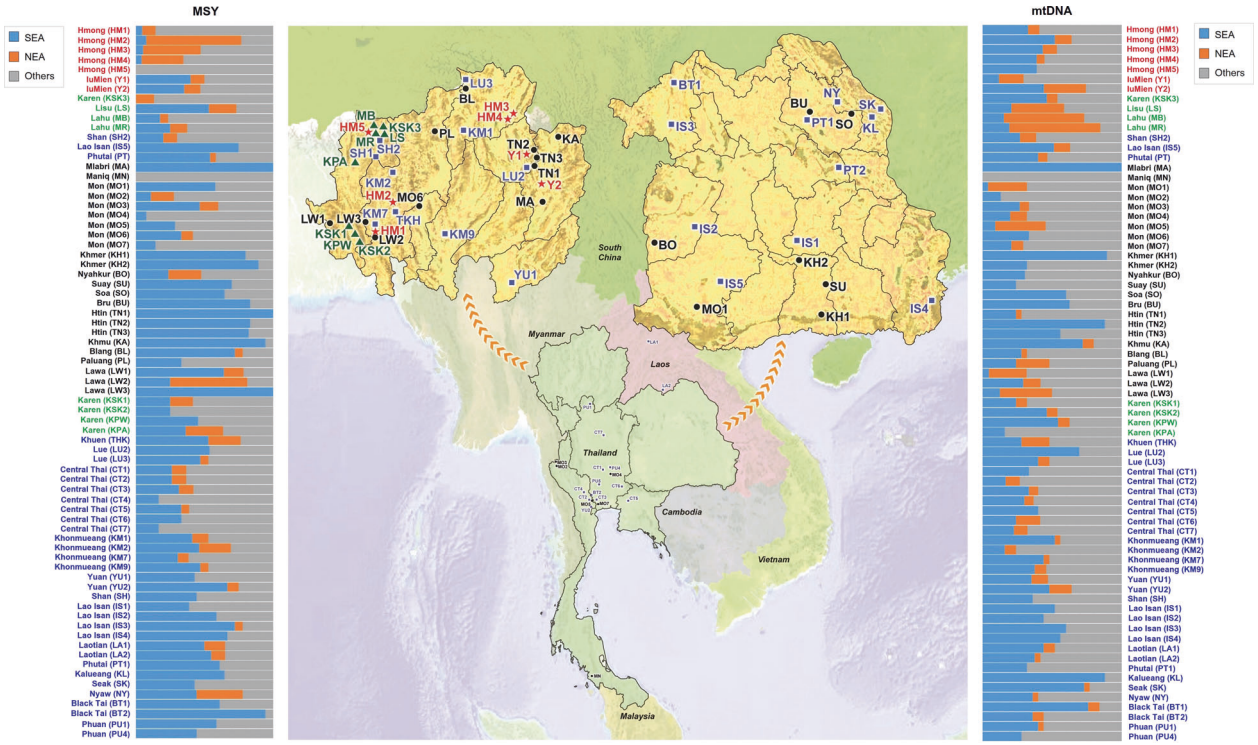

Fig. 1 Map of sampling locations. There are 14 populations sampled in the present study from northern Thailand (HM1–HM5, Y1–Y2, KSK3, MR, MR, LS and SH2) and northeastern Thailand (PT2 and IS5), together with 59 Thai/Lao populations sampled in previous studies [12–15]. Red stars, green triangles, black circles, and blue squares represent Hmong-Mien, Sino-Tibetan, Austroasiatic, and Tai-Kadai speaking populations, respectively. The barplots on the left and right sides of the map depict the proportion of MSY and mtDNA haplogroups specific to Northeast Asia (NEA), Southeast Asia (SEA), or of unknown/other origin.

groups (e.g., Akha, Lisu, Lahu, Karen, Hmong, and IuMien) is in the area further north of Thailand, i.e., northern Myanmar, northern Laos, and southern China, and most of these groups migrated to present-day Thailand ~200 ya [6, 7]. Thus, the present-day cultural and linguistic variation in Thailand has multiple sources, but the HM and ST groups have not been studied in as much detail and the impact of this variation on genetic variation is still poorly understood.

The HM language family is one of the major language families in MSEA, comprising some 39 languages (35 Hmongic and 4 Mienic) distributed across China, northern Vietnam, northern Laos, and northern Thailand [8]. Although the Hmongic and Mienic languages are relatively similar to one another, there are differences due to divergent developments in the phonology [8]. The heartland of the Hmong people is considered to be in the present-day southern Chinese province of Kweichow, where they were established at least 2 kya, probably arriving from the east [6]. Migrations into Thailand through Laos are documented since the second half of the 19th century A.D. The IuMien are, like the Hmong, thought to have an origin in south-eastern China, from which they started to migrate south-wards to Vietnam in the 13th century A.D., entering Thailand about 200 ya [6, 9]. With two main ST sub-families, Chinese and Tibeto-Burman, the ST family is both large (~460 languages spoken by over a billion people) and spread across many countries in South, East and SEA, including China, Nepal, Bhutan, northeastern India, Pakistan, Myanmar, Bangladesh, Thailand, Viet-nam, and Laos [10].

The HM and ST groups in Thailand are regarded as the hill tribes who inhabit the high mountainous northern and western region of Thailand. Consisting of ~700,000 people, there are nine officially recognized hill tribes: the AA-speaking Lawa, Htin and Khmu; the HM-speaking Hmong and IuMien; and the ST-speaking Karen, Lahu (or Mussur), Akha, and Lisu [6, 7]. In addition to living in a remote and isolated region of Thailand, the hill tribes are of interest for their cultural variation. In particular, postmarital residence pattern varies among the hill tribes, with some practicing patrilocality (i.e., following marriage, the woman moves to the residence of the man) while others are matrilocal (i.e., the man moves to the residence of the woman). The first study to document an effect of patrilocality vs. matrilocality on patterns of human mitochondrial (mt) DNA vs. male-specific portion of the Y chromosome (MSY) variation was carried out on the hill tribes [11], and has been further investigated in subsequent studies [9, 12].

Our previous studies on the paternal and maternal genetic lineages and structure of many TK and AA groups indicated

different and complex demographic histories in populations from Thailand and Laos [12–15]. However, the ST and HM speaking groups have not been studied in as much detail. Here we generated and analyzed 416 complete mtDNA genome sequences and 234 partial sequences of the MSY from 14 populations belonging to 11 HM and ST speaking hill tribe populations, and from three non-hill tribe TK populations: the Shan, who migrated recently from Myanmar and live in the mountainous area of northern Thailand; and the Phutai and Lao Isan from the Northeast of Thailand (Fig. 1). This is the first detailed genetic study of Thai HM speaking groups and we also revisit the impact of patrilocality vs. matrilocality on patterns of mtDNA and MSY variation with higher-resolution methods and more populations than studied previously.

Materials and methods

Samples

Samples were collected from 416 individuals belonging to 14 populations classified into three linguistic groups: (1) HM groups, consisting of five Hmong populations (HM1–HM5) and two IuMien populations (Y1 and Y2); (2) ST groups, consisting of two Lahu populations (MR and MB), one Lisu (LS), and one Karen subgroup Skaw (KPW3); and (3) Tai-Kadai groups, consisting of one Shan population (SH2), one Phutai population (PT2), and one Lao Isan population (IS5); (Supplementary Table S1; Fig. 1). Genomic DNA samples of HM1–HM4 and Y1 and Y2 were from a previous study [16], while the remaining groups were newly-collected buccal samples obtained with written informed consent and with ethical approval from Khon Kaen University and the Ethics Commission of the University of Leipzig Medical Faculty. We extracted DNA using the Gentra Puregene Buccal Cell Kit (Qiagen, Germany) according to the manufacturer's directions.

Sequencing

Genomic libraries with double indices were prepared and enriched for mtDNA as described previously [17, 18]. The libraries were sequenced on an Illumina Hiseq 2500 to obtain mtDNA consensus sequences as described in our previous studies [14, 15]. We used Bustard for Illumina standard base calling and the read length was 76 bp. We also enriched for ~2.3 mB of the MSY from the same genomic libraries for male samples via in-solution hybridization-capture using a previously designed probe set [12, 14] and the Agilent Sure Select system (Agilent, CA). The target MSY regions are positioned between positions 9927192 and 13199427 on the human reference genome *hg19* [12, 14]. Sequencing was carried out on the Illumina HiSeq 2500 platform with paired-end reads of 125 bp length and we used Bustard for Illumina standard base calling. The process for manipulating raw sequencing data, alignment and post-processing pipeline of the sequencing data for both mtDNA and MSY were carried out as previously described [12–15]. We then manually checked and manipulated sequences with Bioedit (www.mbio.ncsu.edu/BioEdit/bioedit.html). The complete mtDNA sequence data set can be found at GenBank (accession number MT418943–MT419358). All reads that aligned to the region of the MSY that was targeted by the capture-enrichment array were deposited in the European Nucleotide Archive (study ID PRJEB36639). Final SNP genotypes and their chromosomal positions on *hg19* are provided in Supplementary Table S2.

Statistical analysis

The newly-generated 234 MSY sequences from 14 populations were combined with 928 sequences from 59 populations from our previous studies [12, 14] for a total of 1162 sequences belonging to 73 populations. For mtDNA, combining the 416 new sequences from this study with 1434 sequences from our previous studies [13–15] brings the total to 1850 sequences from 73 populations. Summary statistics of the genetic diversity within populations, the matrix of pairwise genetic distances (Φ_{st}), and analyses of molecular variance (AMOVA) were obtained with Arlequin 3.5.1.3 [19]. To visualize population relatedness, the R package [20] was used to carry out the nonmetric MDS analysis (based on the Φ_{st} distance matrices for the MSY and mtDNA) (R function: isoMDS package: MASS) and to construct heat plots of the Φ_{st} distance matrix and the matrix of shared haplotypes (R function: ape, pegas, adegenet and ggplot2 packages). STATISTICA 13.0 (StatSoft, Inc., USA) was used to carry out a correspondence analysis (CA) based on MSY and mtDNA haplogroup frequencies. For the MSY, haplogroup assignment was performed by yHaplo [21]. Haplogroups were assigned to the maximum depth possible given the phylogeny of ISOGG Y-DNA Haplogroup Tree 2015 (http://www.isogg.org/) and the available genetic markers in our target region. The derived SNPs for each sample are provided in Supplementary Table S2. The mtDNA haplogroups were assigned by Haplogrep2 [22] with PhyloTree mtDNA tree Build 17 (http://www.phylotree.org) [23] and the polymorphisms for each sample are provided in Supplementary Table S3. To obtain a broader picture of population relationships within SEA, we included publicly-available sequences from relevant populations for the MSY and mtDNA (Supplementary Table S4). To compare Northeast Asia (NEA) and SEA prevalent haplogroups with our studied populations, we calculated the frequency of MSY/mtDNA NEA and SEA prevalent

haplogroups in the HM and ST speaking populations from China and Vietnam (Supplementary Table S5). To construct Bayesian skyline plots (BSP) per population and maximum clade credibility (MCC) trees per haplogroup, based on Bayesian Markov Chain Monte Carlo (MCMC) analyses, we used BEAST 1.8.4 [24]. For BSP plots by population, we conducted analyses both pooling all populations within the same ethnicity (e.g., pooling HM1-HM5), and for the individual populations. The Bayesian MCMC estimates (BE) and 95% highest posterior density (HPD) intervals of haplogroup coalescent times were calculated using the CongPy6 sequence (haplogroup A1b1-M14) for rooting the tree for MSY haplogroup C [25] and the Mbuti-3 sequence (haplogroup B-M182) for rooting the tree for haplogroups F and O [26]. BEAST input files were created with BEAUTi v1.8.2 after first running jModel test 2.1.7 in order to choose the most suitable model of sequence evolution [27]. The best substitution models are shown in Supplementary Table S6. We used an MSY mutation rate of 8.71×10^{-10} substitutions/bp/year [28], and the BEAST input files were modified by an in-house script to add in the invariant sites found in our data set. Both strict and log normal relaxed clock models were run, with marginal likelihood estimation [29, 30]. After each BEAST run, the Bayes factor was computed from the log marginal likelihood of both models to choose the best-fitting BSP/MCC tree (Supplementary Table S6). For mtDNA, we executed BSP analyses per population and the BEAST runs by haplogroup, with mutation rates of 1.708×10^{-8} and 9.883×10^{-8} for data partitioned between the coding and noncoding regions, respectively [31]. The RSRS was used for rooting the tree for mtDNA [32]. The strict clock model was used, and the best substitution models are shown in Supplementary Table S6. The Bayesian skyline piecewise linear tree prior for the dating and Bayesian skyline generation were applied, so as to allow for population size changes over time. The number of chains for each MCMC were varied based on sample sizes (Supplementary Table S6), but were always sufficient for successful Bayesian estimation and to reach ESS values above 200. Tracer 1.6 was used to generate the BSP plot from the BEAST results. The Bayesian MCC trees were assembled with TreeAnnotator and drawn with FigTree v 1.4.3.

Results

Genetic lineages

MSY

We combined the 234 newly-generated sequences with 928 sequences from our previous studies [12, 14] for a total of 1161sequences, of which 818 are distinct, from 73 populations; population details are listed in Supplementary Table S1. The mean coverage of the newly-generated 234 MSY sequences of ~2.3 mB ranges from 7× to 60× (overall average coverage 18×) (Supplementary Table S2).

When combined with our previous Thai/Lao data, there are a total of 90 haplogroups identified (Supplementary Table S7); 10 of these were not found in our previous studies. O1b1a1a (O-M95) (26.61%), O2a2b1a1 (O-Page23) (9.75%), and O1b1a1a1a1 (O-M88) (7.15%) are prevalent in almost all groups (overall frequency 43.52%) (Supplementary Fig. S1). Haplogroup O2a2a1a2a1a2 (O-N5) and C-F845 are mostly prevalent in HM groups while haplogroup F is the dominant haplogroup of the Lahu (MR and MB). The coalescent ages of these three haplogroups are ~2.45 kya (HPD: 2.88–1.13 kya) for O2a2a1a2a1a2 (O-N5), ~12.54 kya (HPD: 17.16–4.09 kya) for C-F845 and ~16.00 kya (HPD: 22.59–12.26 kya) for haplogroup F. However, if we focus on HM or Lahu clades of the MCC tree of haplogroup C-F845, the age is ~2.85 kya (HPD: 4.25–1.08 kya) and ~0.58 ya (HPD: 1.55–0.29) for haplogroup F (Supplementary Fig. S2).

When we focus on the MSY lineages that are prevalent in NEA, i.e., C2e*, D-M174 and N* in our samples, the frequency of NEA lineages is >30% among the Hmong (HM2-HM4), Lawa (LW2), and Karen (KPA) from northern Thailand, and the Nyaw (NY) from northeastern Thailand. In contrast O1b*, which is the predominant lineage in SEA and at high frequency in AA speaking people [12], is the major lineage in the Thai/Lao AA speaking group (except for some Mon populations who show evidence of admixture with Central Thai populations). Interestingly, there is heterogeneity in the frequency of NEA/SEA lineages in the Hmong and Lawa groups: among the five Hmong populations, HM5 completely lacks NEA lineages, while within the Lawa groups, LW2 has 56% NEA lineages while LW3 has exclusively SEA lineages (Fig. 1).

mtDNA

We generated 416 complete mtDNA sequences with mean coverage ranging from 35× to 7752× (overall average coverage 1934×) (Supplementary Table S3). When combined with 1434 sequences from our previous studies [13–15] there are in total 1850 sequences belonging to 73 populations, with 1125 haplotypes. When combined with our previous data there are a total of 285 haplogroups (Supplementary Table S8); several were not reported previously from Thai/Lao populations and these are specific to some populations, e.g., B4a5 (specific to the Hmong), and B5a1c1a, B5a1c1a1, D4e1a3, and F1g1 (specific to HM groups).

The coalescent ages (Supplementary Fig. S3) of the prevalent lineages of HM groups are ~10.67 kya (HPD: 11.27–3.31 kya)

for B5a1c1a*, ~1.53 kya (HPD: 3.96–0.94 kya) for B5a1c1a1, ~6.83 kya (HPD: 7.09–1.55 kya) for D4e1a3, ~11.54 kya (HPD: 16.04–5.84 kya) for F1g1. B4a5, specific for the Hmong, has a coalescent age of ~6.63 (HPD: 11.16–3.22 kya). The coalescent ages of D4j1a1 and G1c, abundant in the Lahu, are ~9.23 kya (HPD: 12.07–4.85 kya) and ~3.88 kya (HPD: 4.70–0.21 kya). In addition, the Lahu-specific clade of D4j1a1 sequences is dated to 2.49 kya (HPD: 5.83–1.53 kya). The coalescent age of B6a1a, abundant in the Karen, is ~6.69 kya (HPD: 11.62–4.25 kya).

Haplogroup A*, D* and G* are predominant in NEA populations and we find that the frequency of these NEA lineages is >30% in Lahu (MB and MR), Lawa (LW3), Lisu (LS), and IuMien (Y2) from northern Thailand, and in Mon (MO5) from central Thailand. In contrast, the predominant SEA haplogroups (B5*, F1a*, M7b* and R9b*) are at highest frequency in TK and AA speaking groups, indicating genetic similarity between these two groups. Strikingly, the populations with the highest frequencies of NEA MSY lineages are not the same as the populations with the highest frequencies of NEA mtDNA lineages (Fig. 1), which we suggest below may reflect differences in ancestral postmarital residence patterns.

Genetic diversity

MSY

Overall, the HM, AA and ST groups tend to have more heterogeneous and lower genetic diversity values than the TK groups (Fig. 2; Supplementary Table S9). By contrast, genetic diversities of the HM groups are not statistically different from the AA and ST groups, nor do the ST and AA groups differ significantly in genetic diversity values (Supplementary Table S9). At the individual population level, out of 63 populations, the Hmong (HM1) shows lower haplotype diversity than all other groups except for two hunter-gatherer groups (Mlabri (MA) and Maniq (MN)) and the Htin Mal (TN1). HM1 and HM5 have lower genetic diversity than the other HM populations, although HM2 shows the highest MPD values. Generally, the Hmong (HM1–HM5) groups show lower haplotype and haplogroup diversity than the In Mien (Y1–Y2) (Fig. 2a, b). Of the eight ST speaking populations, the newly-studied Karen group (KSK3) exhibits lower genetic diversities while the Lisu (LS) and KSK1 have higher genetic diversities than the other ST speaking populations (Fig. 2a–d). Although low genetic diversities are observed in HM1, a significantly low Tajima's *D* value (Fig. 2d) suggests recent paternal expansion in this group. Significant negative Tajima's *D* values were observed more frequently in the TK than in the AA and HM groups ($P < 0.05$: 11/34 for TK, 6/24 for AA, and 2/7 for HM) but no significant Tajima's *D* values were observed in any of the ST-speaking groups (Fig. 2d).

mtDNA

Along with the Mlabri (MA), Htin (TN1 and TN2), and Seak (SK), the ST speaking Lahu or Mussur (MR) shows low mtDNA haplotype and haplogroup diversities whereas the Lisu (LS) shows higher genetic diversities than the other ST populations (Fig. 2a–c). In contrast to the MSY, the Hmong groups exhibited generally higher genetic diversities than the ST and AA speaking groups (Supplementary Table S9). Both the ST and AA groups have lower genetic diversity values than the TK groups (Supplementary Table S9). As also seen in the MSY results, both IuMien populations (Y1 and Y2) show higher genetic diversities than the other Hmong and ST speaking groups (Fig. 2a–c). In agreement with the MSY results, a significantly negative Tajima's *D* value was observed more frequently for the TK than for the AA and HM groups ($P < 0.05$: 21/34 for TK, 5/24 for AA, and 2/7 for HM). Interestingly, the ST-speaking groups show no significant Tajima's *D* values while the two IuMien groups both show significant negative Tajima's *D* values (Fig. 2d).

The analysis of molecular variance (AMOVA)

MSY

The AMOVA results indicate that the variation among populations accounts for 13.72% of the total MSY genetic variance (Table 1). The HM group shows the greatest genetic heterogeneity among populations, followed by the AA and ST groups; the TK group shows the lowest among-population variation. The Thai Hmong, with five populations sampled, shows higher variation among the populations than do the other hill tribe groups. When HM and ST populations from Vietnam [33] were included in the analysis, genetic variation among populations of the ST and HM groups increased substantially, suggesting some differentiation between Vietnamese and Thai populations belonging to these two groups. However, a direct comparison of Thai ST vs. Vietnamese ST, and Thai HM vs. Vietnamese HM groups, showed no significant differences between groups. The variation among populations within groups of the IuMien and Lahu were much lower than for the Hmong, indicating genetic heterogeneity of the Hmong and ST populations and more homogeneity for the IuMien and Lahu. The MSY genetic variation showed significant differences among the four language families (HM, ST, AA, and TK), but the variation among groups was lower than the variation among populations within each group, indicating that language families do not correspond to genetic structure. All pairwise comparisons of the four language families

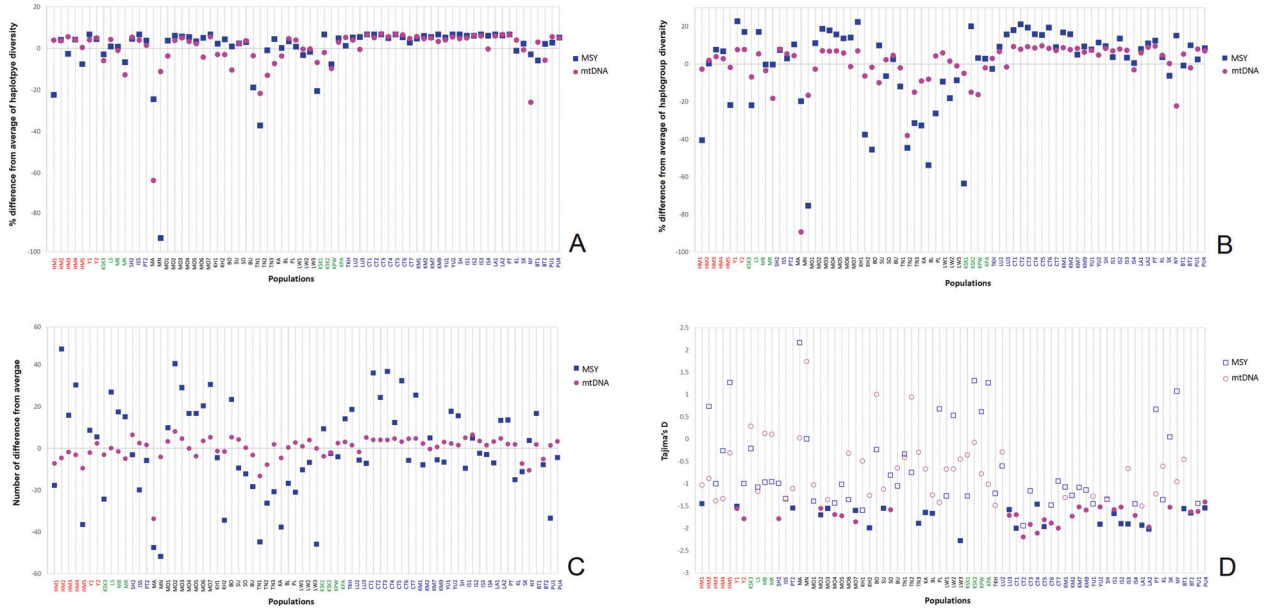

Fig. 2 Genetic diversity values shown as the percent difference from the average. (**a**) haplotype diversity, (**b**) haplogroup diversity and (**c**) MPD. The gray line shows the mean across populations. (**d**) Tajima's *D* values; solid symbols indicate values significantly different from zero ($P < 0.05$). More information and all genetic diversity values are provided in Supplementary Table S1. The new populations are placed at the left of the figure. Population names are color-coded according to language family; red, green, black, and blue represent HM, ST, AA, and TK- speaking populations, respectively.

showed significant differences among groups, but these were on the same order as the differences among populations within the same group. However, when the HM were separated into Hmong and IuMien, the pairwise comparisons of Hmong with other language families remained significant, while for the IuMien there were no significant differences with other language families, suggesting some differences between Hmong and IuMien groups. The lowest variation was between the TK and AA groups, indicating a relatively close genetic relationship between these two.

mtDNA

The total mtDNA variation among populations of 8.46% was lower than for the MSY (Table 1). The mtDNA variation for the AA, ST, and TK groups was about the same as for the MSY, but was substantially less for the HM. The Htin have by far the largest variation among populations, while the Hmong show nonsignificant mtDNA variation among populations (0.53%), reflecting genetic homogeneity in their maternal side. The mtDNA genetic variation among the four language families (HM, ST, AA, and TK) was much smaller (1.09%) than the variation among populations assigned to each group (7.7%), indicating that as with the MSY, language families do not correspond to the genetic structure of these populations. The variation between pairs of linguistic groups shows in all comparisons that the variation between groups is lower than the variation among populations within groups. As with the MSY, the lowest variation (which is not significantly different from zero) is between the TK and AA groups, further supporting a close relationship between these two groups in Thailand.

The pooled mtDNA data of Lahu from Vietnam and Thailand revealed much larger variation for mtDNA (13.47%) than for the MSY (4.88%), in contrast to the larger MSY than mtDNA variation observed when pooling data from other groups from Thailand and Vietnam. In particular, the mtDNA variation among Hmong groups from Thailand and Vietnam was only 1.08%, which is not significantly different from zero. When the Hmong and IuMien were separately compared with other linguistic groups, significant differences were observed for the Hmong but not for IuMien, similar to the MSY results and further supporting the difference between Hmong and Mien groups.

Population affinity

MSY

Shared haplotypes within populations are an indication of smaller population size and increase relatedness among individuals, while shared haplotypes between populations are an indication of recent shared ancestry or contact. There were shared MSY haplotypes within the HM groups, and some sharing between them and a few TK-speaking groups, except for HM5 and Y1, who did not share any haplotypes with any other populations (Fig. 3a). The Lisu shared haplotypes with both Lahu (MB and MR) populations,

Table 1 AMOVA results.

Groups	a	b	Percent variation					
			Within populations		Between populations within groups		Among groups	
			MSY	mtDNA	MSY	mtDNA	MSY	mtDNA
Total	1	73 (71)	86.28 (86.85)	91.54 (92.33)	13.72** (13.15**)	8.46** (7.67**)		
Hmong-Mien (HM)	1	7	81.83	96.6	18.17**	3.4**		
Austroasiatic (AA)	1	24 (22)	82.15 (83.92)	85.96 (88.47)	17.85** (16.08**)	14.04** (11.53**)		
Sino-Tibetan (ST)	1	8	87	89.76	13**	10.24**		
Tai-Kadai (TK)	1	34	95.59**	95.86	4.41**	4.14**		
HM (Thailand + Vietnam)	1	11	82	94.35	18**	5.65**		
ST (Thailand + Vietnam)	1	12	74.96	87.64	25.04**	12.36**		
Patrilocal	1	13	72.57**	93.64**	27.43**	6.36**		
Matrilocal	1	10	75.51**	83.54**	24.49**	16.46**		
Hmong	1	5	81.87	99.47	18.13**	0.53		
Karen	1	5	90.89	94.15	9.11**	5.85**		
Mon	1	7	96.5	93.1	3.5**	6.90*		
H'tin	1	3	90.11	74.29	9.89**	25.71*		
Lawa	1	3	64.35	92.22	35.65**	7.78*		
Hmong (Thailand + Vietnam)	1	6	78.75	98.92	21.25**	1.08		
IuMien (Thailand + Vietnam)	1	3	94.42	95.5	5.58**	4.5**		
Lahu (Thailand + Vietnam)	1	3	95.12	86.53	4.88*	13.47**		
Language (AA, TK, HM, ST)	4	73 (71)	84.69** (85.26**)	91.21** (91.95**)	10.10** (9.58**)	7.7** (6.85**)	5.21** (5.16**)	1.09** (1.20**)
HM vs. AA	2	31 (29)	70.24** (71.52**)	86.92** (88.46**)	15.18** (14.17**)	11.56** (9.53**)	14.58** (14.31**)	1.52* (2.02**)
HM vs. ST	2	15	78.13**	89.23**	14.59**	6.69**	7.28**	4.09**
HM vs. TK	2	41	82.85**	94.3**	6.11**	4.00**	11.04**	1.70**
Hmong vs. AA	2	29 (27)	65.55** (66.86**)	85.44** (87.02**)	14.06** (13.04**)	11.79** (9.59**)	20.39** (20.10**)	2.77** (3.40**)
Hmong vs. ST	2	13	75.18**	87.78**	13.52**	6.38**	11.30**	5.84**
Hmong vs. TK	2	39	78.04**	93.20**	5.19**	3.72**	16.77**	3.08**
IuMien vs. AA	2	26 (24)	83.01** (84.54**)	89.00** (90.68**)	16.92** (15.21**)	13.97** (11.41**)	0.07 (0.26)	−2.97 (−2.09)
IuMien vs. ST	2	10	88.45**	90.69**	11.46**	9.35**	0.09	−0.04
IuMien vs. TK	2	36	95.92**	95.93**	4.28**	4.14**	−0.19	−0.08
Thai HM vs. Vietnam HM	2	11	80.82**	93.86**	15.98**	5.01**	3.2	1.13
AA vs. ST	2	30	78.04** (79.40**)	86.02** (87.98**)	15.54** (14.33**)	13.03** (11.13**)	6.41** (6.27**)	0.95* (0.9)
AA vs. TK	2	58 (56)	89.71** (90.51**)	91.87** (92.83**)	9.27** (8.57**)	7.92** (6.87**)	1.02* (0.92*)	0.21 (0.3*)
ST vs. TK	2	42	90.22**	92.94**	5.56**	5.02**	4.22**	2.05**
Thai ST vs. Vietnam ST	2	12	74.65**	86.85**	24.44**	11.23**	0.91	1.92

*indicates $P < 0.05$; **indicate $P < 0.01$.

[a]Number of groups.

[b]Number of populations; numbers/values in parentheses were calculated by excluding the two hunter-gather groups, Mlabri (MA) and Maniq (MN).

while both Lahu populations shared haplotypes among themselves and also with one group of central Thai (CT5). The newly studied Karen (KSK3) shared haplotypes with the other Karen populations (KSK1, KSK2, and KPW), and also with their neighbors, i.e., Shan (SH2) and Lawa (LW1) (Fig. 3a).

Genetic distance values are a further indication of genetic relationships among populations; the genetic distances (Φ_{st} values) indicate, in general, genetic heterogeneity among AA populations and homogeneity among TK populations, as well as genetic differences between the AA (except the Mon) and TK populations. For the newly-studied HM groups, significant genetic differences between the Hmong and almost all other populations were observed, whereas the Φ_{st} values for comparisons of the IuMien (Y1 and Y2) and Lisu with many populations were not significant (Fig. 3b). The new Karen group is significantly different from almost all populations except SH2 and KSK1. The two Lahu populations are genetically distinct from all other populations (except each other).

To further visualize the relationships based on the Φ_{st} distance matrix, we carried out MDS analysis. The MDS plot for three dimensions indicates genetic distinction of the Maniq (MN), the hunter-gatherer group from southern Thailand, the Hmong groups (HM1-HM5) and the Karen (KSK3) (Supplementary Fig. S4), as further indicated in the MDS heat plot (Supplementary Fig. S5). Based on the MDS results for both the MSY and mtDNA, we removed five highly-diverged populations (MA, MN, TN1, TN2, and SK); a three-dimension MDS for the remaining 68 populations has an acceptable stress value (Fig. 4a–c). There was overall some clustering of populations according to

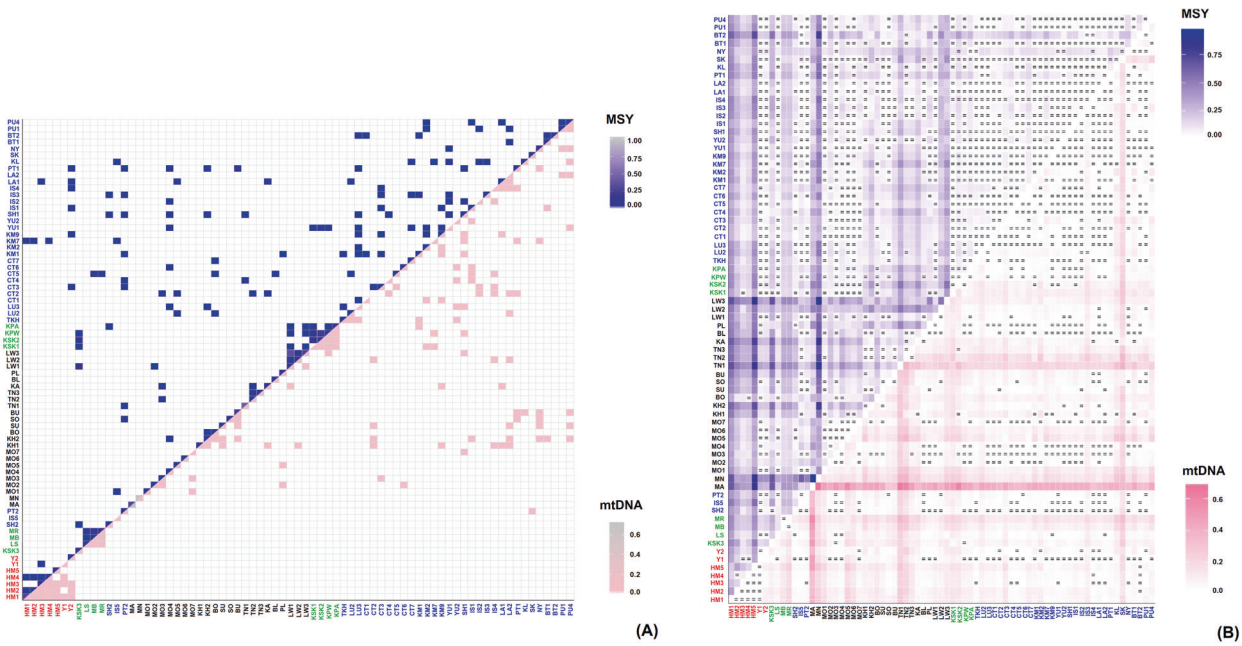

Fig. 3 Frequency of shared haplotypes and heat plot of Φ_{st} values. (**a**) Frequency of shared MSY (above diagonal) and mtDNA (below diagonal) haplotypes within and between populations. (**b**) Heat plot of Φ_{st} values based on MSY (above diagonal) and mtDNA (below diagonal) haplotypes. The "=" symbol indicates Φ_{st} values that are not significantly different from zero ($P > 0.05$). The new populations are placed at the left of the figure. Population names are color-coded according to language family; red, green, black and blue represent HM, ST, AA, and TK-speaking populations, respectively.

language family, albeit with some overlapping between them. The Hmong populations are quite distinct from all other groups, whereas the IuMien populations are more similar to other groups than to the Hmong groups. The TK overlap with AA groups, but the AA are more spread out, indicating more genetic divergence of AA groups. KSK3 and, to a lesser extent, KSK2 and both Lahu populations are distinct from the other ST groups.

To investigate the relationships of Thai/Lao populations with other SEA populations, we included available comparable sequencing data from populations from Vietnam, southern China, and Myanmar. The MDS plot based on the Φ_{st} distance matrix of 88 populations (Fig. 5a–c; the same outliers are excluded as in Fig. 4a–c) shows clustering of the Vietnamese HM-speaking Pathen and Hmong with the Thai Hmong populations, while the Vietnamese HM-speaking Yao are more similar to the Thai IuMien groups. The Karen (KSK3) remains distinct, but shows a close genetic relatedness to Burmese. Some of the TK-speaking groups from Vietnam, i.e., Nung, Tay and Thai, are close to the TK populations from northeastern Thailand, e.g., Phutai (PT1 and PT2), Kalueang (KL), and Black Tai (BT1).

To investigate which MSY haplogroups might be driving population relationships, we carried out a correspondence analysis (CA), which is based on haplogroup frequencies. The results indicate that the genetic distinctiveness of the Hmong reflects high frequencies of haplogroups O2a2a1a2a1a2 (O-N5) and C2e2 (C-F845) (Supplementary Fig. S6). One IuMein population (Y1) is positioned between the Hmong and other Thai/Lao populations, reflecting haplogroups D (D-M174) and C2e1b1 (C-F1319). The second dimension distinguishes the two Lahu groups (MB, MR) and Seak (SK), based on haplogroup F (F-M89). The Soa (SO) occupy an intermediate position, based on O1b1a1a1a2a1a (O-Z24091). The third dimension distinguishes three of the Karen populations (KSK1, KSK2, and KPW), based on haplogroups O1b1a1a1b1a1 (O-FGC29907) and G1 (G-M342). Further dimensions distinguish an AA group (LW2) based on haplogroups O2a2b2a2 (O-F706) and N (N-M231), while the AA group MO2 and TK group CT7 are distinguished based on haplogroups R1a1a1b2a1b (R-Y6) and J2a1 (J-L26).

mtDNA

With respect to mtDNA haplotype sharing (Fig. 3a), the HM populations (HM1–HM5) share mtDNA haplotypes extensively with each other, including with the IuMien populations (Y1 and Y2), but do not share haplotypes with any other population, reflecting their unique genetic structure. As with the MSY results, the Lisu (LS) only shares haplotypes with both Lahu populations (MB and MR), indicating contact between them. The newly studied Karen population (KSK3) also exhibits large differences from the other Karen groups: there is no mtDNA haplotype sharing between the KSK3 and the other Karen populations, while

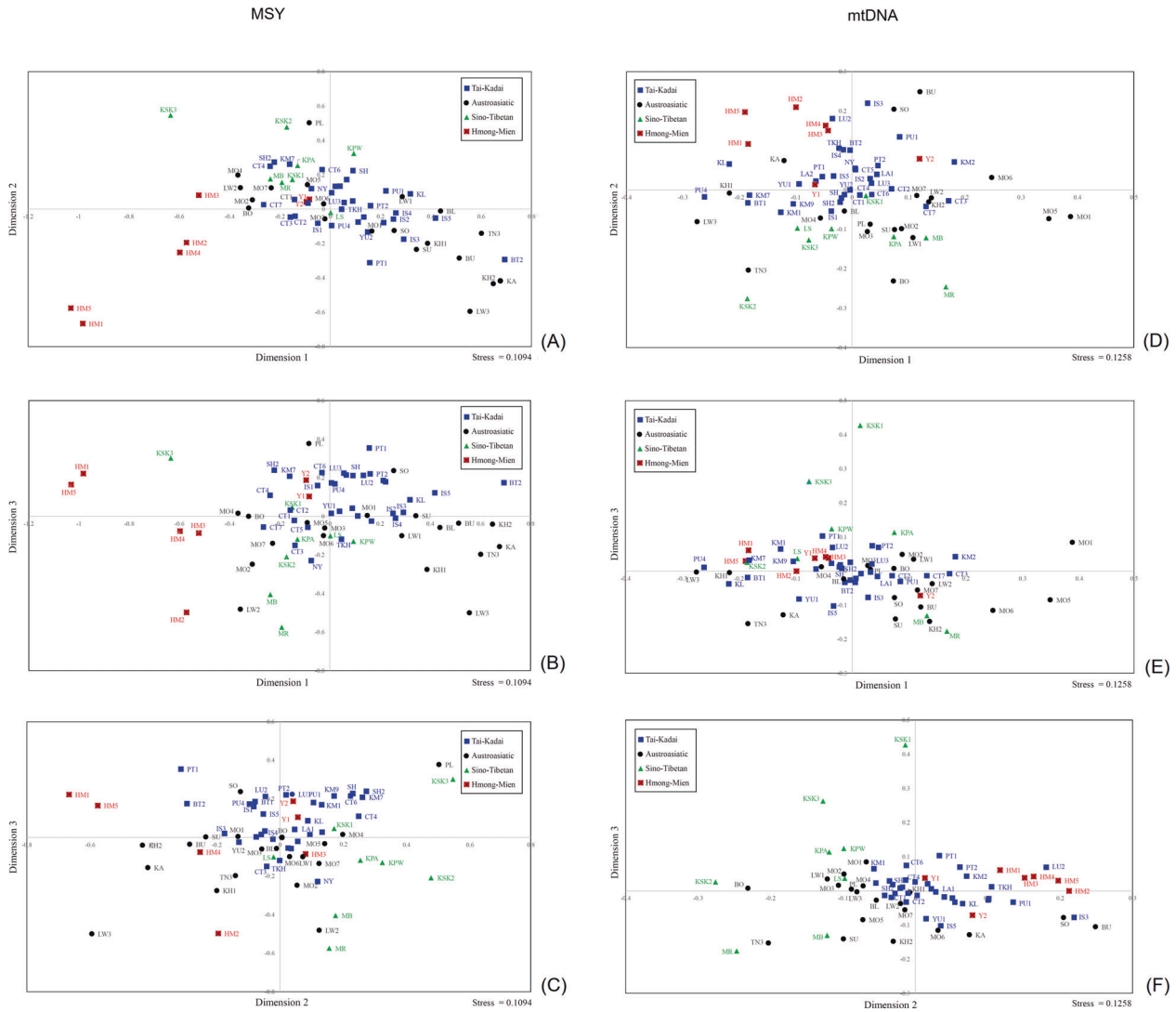

Fig. 4 MDS plots based on the Φ_{st} distance for Thai/Lao populations. The three-dimensional MDS plots for 68 Thai/Lao populations (after removal of Maniq, Mlabri, Htin (TN1, TN2) and Seak (SK)) for (**a**–**c**) MSY and (**d**–**f**) mtDNA. The stress values are 0.1094 for MSY and 0.1258 for mtDNA.

some of the other Karen populations do share haplotypes with one another.

The heat plot of Φ_{st} genetic distances (Fig. 3b) also supports genetic distinction of the HM from other Thai/Lao populations, and mostly nonsignificant Φ_{st} values among them, with the exception of the IuMien populations and HM3, who show more similarity to other Thai/Lao populations. However, consistent with the MSY results, Lisu are not significantly different from several TK, AA, and ST speaking populations, while the two Lahu populations do not differ significantly from each other, but do show significant differentiation from the other Thai/Lao groups.

The MDS plots based on Φ_{st} values for the Thai/Lao populations show greatest genetic divergence for the MA, the hunter-gatherer group from northern Thailand, followed by their linguistic relatives, the Htin (TN1, TN2) and Seak (SK) (Supplementary Fig. S4). After removal of the same five outliers as for the MSY analysis (MA, MN, TN1, TN2, and SK), the MDS analysis based on dimensions 1 and 2 shows separation between the Hmong populations and the ST populations, with the IuMien populations rather closer to the cloud of TK populations around the center of the plot. The Karen groups are further differentiated by dimension 3 (Fig. 4d–f).

The MDS plot based on the Φ_{st} distance matrix that includes comparative data from other SEA populations (Fig. 5d–f) shows that the HM speaking populations from Thailand and Vietnam tend to cluster together, except the Thai/Vietnamese IuMien are closer to other populations, consistent with the MSY results. The Vietnamese Lahu are quite distinct from the Thai Lahu, and in fact are closer to the Thai HM groups. Interestingly, the ST speaking populations are about as heterogeneous as the AA speaking groups.

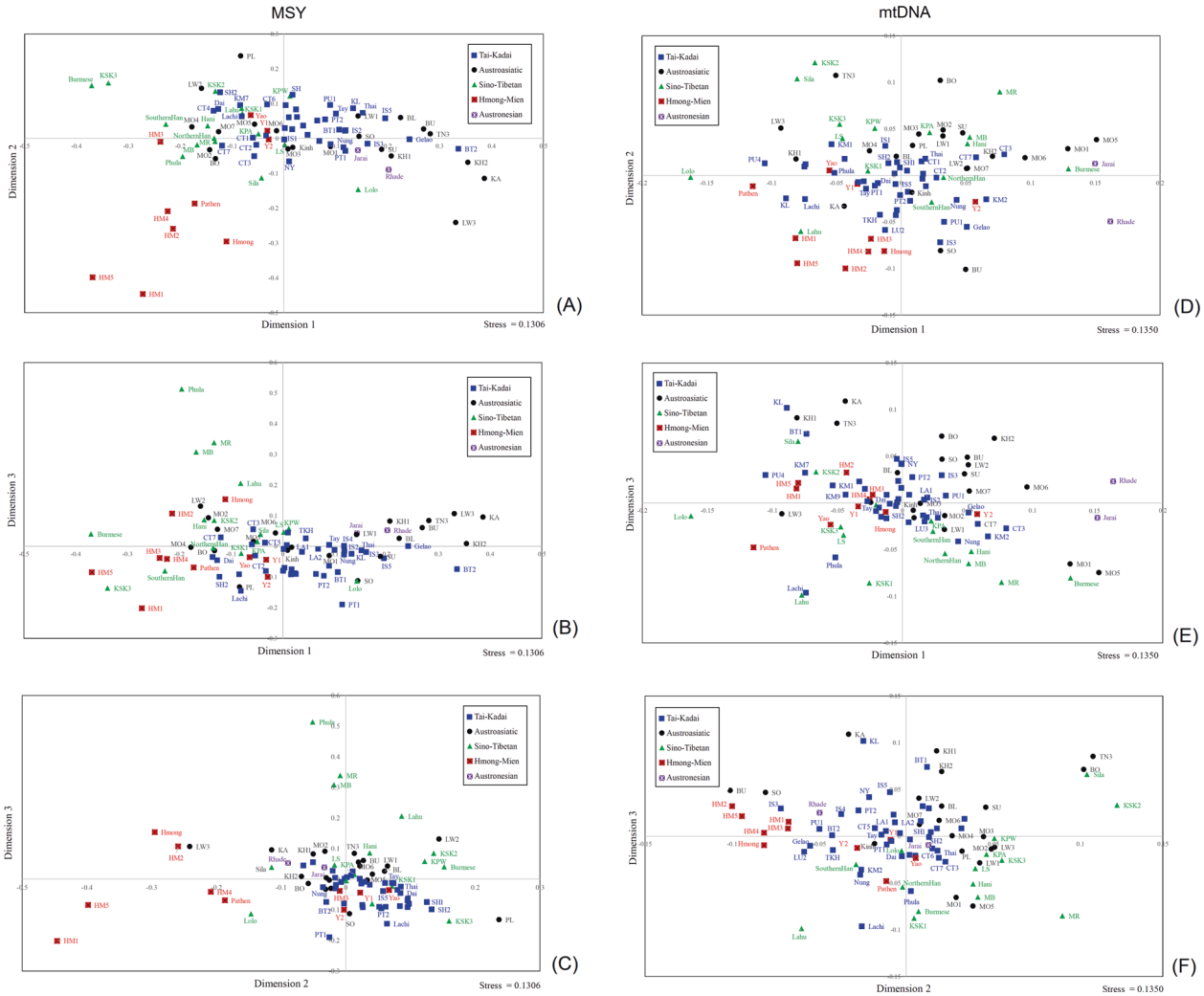

Fig. 5 The three-dimensional MDS plot based on the Φ_{st} distance matrix for 88 SEA populations. (**a**-**c**) MSY and (**d**-**f**) mtDNA. The stress values are 0.1306 for MSY and 0.1350 for mtDNA.

The CA analysis based on mtDNA haplogroup frequencies (Supplementary Fig. S7) further confirms the distinctiveness of the HM groups based on several haplogroups, i.e., B5a1c1a, B5a1c1a1, B4a5, C7a, D4e1a3, F1g1, F1g2, N9a10 (16311C), and M74a. The Lahu and MO5 were distinguished in the third dimension, reflecting haplogroups B4e and D4j1a1. In the fourth dimension several groups are distinguished via many specific lineages, including all of the Karen groups and two Mon groups from the border between Thailand and Myanmar.

Bayesian skyline plots

MSY

The BSPs of population size change (N_e) over time were constructed for each ethnicity. For the MSY, different trends were observed for different groups (Fig. 6). The N_e of the Hmong gradually increased since ~30 kya and then declined ~2–3 kya, while for the Lahu the N_e remained stable for a long period of time and then was sharply reduced around ~1 kya. The Karen, Shan, and Phutai showed a similar trend: the N_e gradually increased, and then decreased ~5 kya, with sharp increases ~2–3 kya, followed by another decrease ~1 kya. The N_e for the IuMien slightly increased, and then decreased ~2–3 kya. In general, we see similar demographic changes at 5 kya, 2–3 kya and 1 kya in both the new groups studied here and in our previous studies of AA Mon, Khmer and Htin groups, and central Thai TK groups [12]. The first two changes may reflect male-specific expansion during the Neolithic period and the Bronze/Iron Age that are characteristic of modern AA and TK groups respectively, as discussed previously [12].

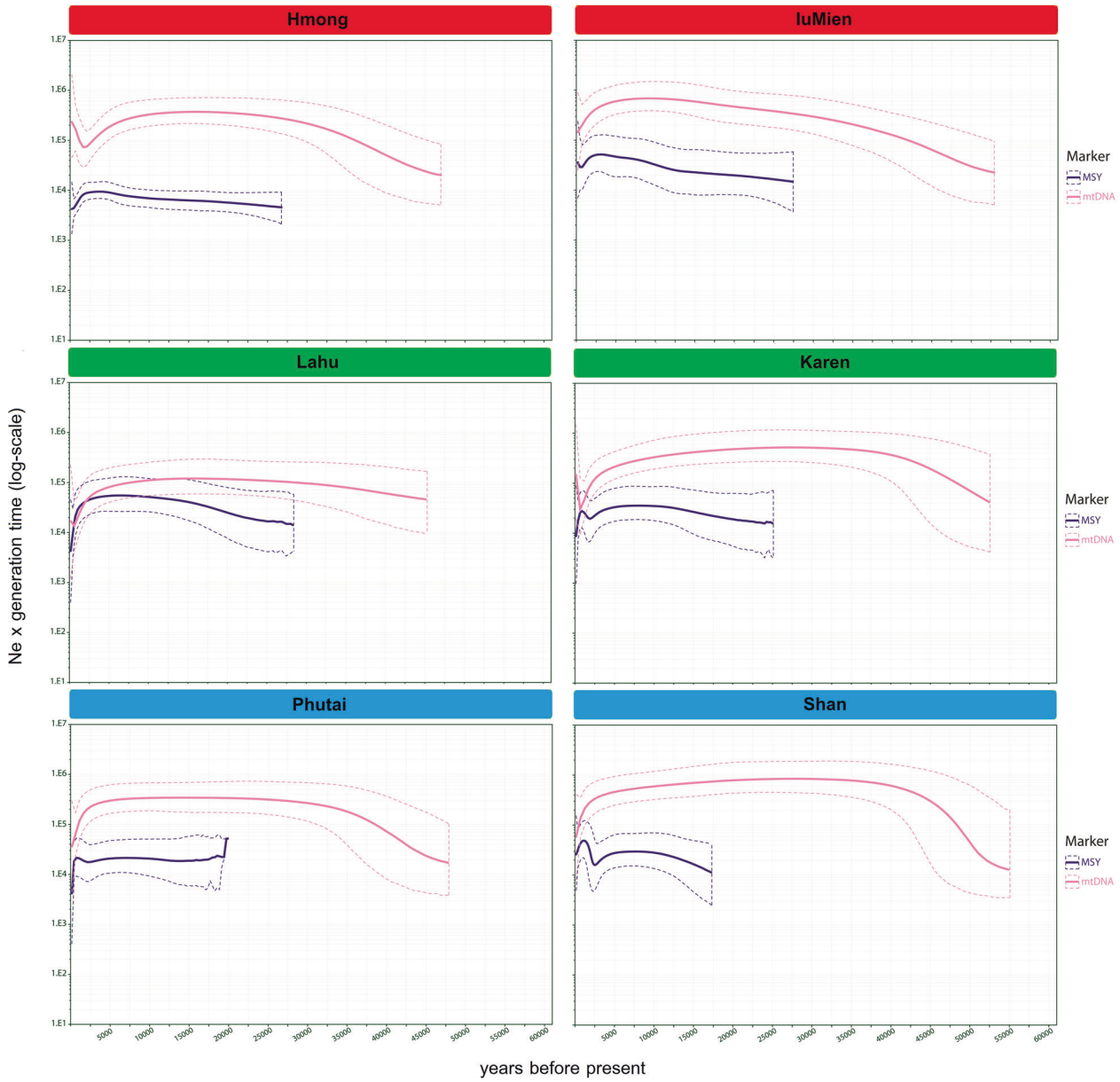

Fig. 6 Bayesian skyline plots (BSPs). The BSPs based on the MSY and mtDNA for the Hmong, IuMien, Lahu, Karen, Shan and Phutai groups. Solid lines are the median estimated effective population size (y-axis) through time from the present in years (x-axis). The 95% highest posterior density limits are indicated by dotted lines.

mtDNA

The BSPs for each ethnicity show that several groups, i.e., Lahu, Shan, Phutai, show a common trend of N_e increasing 30–50 kya, and then stable until a decline ~2–5 kya (Fig. 6; BSPs for each individual population are in Supplementary Fig. S8). However, the Hmong and Karen showed a different pattern, namely a decrease in N_e ~5 kya followed by rapid growth ~2.0–2.5 kya for the Hmong, and ~1.0 kya for the Karen. In general, the BSP plots for the AA (Mon, Htin, Lawa, and Khmer) and TK populations (Yuan, Phuan, and Lue) [12] are similar to the BSPs observed for most of the groups in this study. The major TK groups (Khon Muenag, Central Thai, and Lao Isan) also showed a population increase ~10 kya in our previous study [12]. Interestingly, the distinct BSP plot of the Hmong, showing population increase during the Bronze/Iron Age, indicates different demographic changes in the maternal vs. paternal side and supports genetic differences of the Hmong from the other populations indicated by other results (Figs. 3, 4).

Patrilocal vs. matrilocal genetic variation

There are nine official hill tribes in Thailand: the AA-speaking Lawa, Htin, and Khmu; the HM-speaking Hmong and IuMien; and the ST-speaking Karen, Lahu, Lisu and Akha. The Lahu, Karen, and Htin are matrilocal (i.e., the husband moves to the residence of the wife after marriage) whereas the others are patrilocal. Our previous study [12] had investigated four hill tribes (Lawa, Htin, Khmu, and Karen); here we add data from four additional hill tribes (Hmong, IuMien, Lahu, and Lisu) for a total of 23 populations belonging to eight hill tribes. Moreover, although the Palaung is not officially recognized as a hill tribe group, we include them in the analysis because they are a minority people from the same mountainous region of northern Thailand.

The Hmong (HM1–HM5), IuMien (Y1–Y2), Lisu (LS), Khmu (KA), Lawa (LW1-LW3), and Palaung (PL) groups practice patrilocality, whereas the Htin (TN1–TN3), Karen (KSK1–KSK3, KPA, and KPW) and Lahu (MR and MB) are matrilocal. If genetic variation was influenced by residence pattern, then lower within-population genetic diversity coupled with greater genetic heterogeneity among populations is expected for matrilocal groups than for patrilocal groups for the mtDNA, whereas the opposite pattern is expected for the MSY [11]. However, the MSY *h* and MPD values do not differ significantly between patrilocal and matrilocal groups (Supplementary Table S9 and Fig. S9). For mtDNA, genetic diversity values are significantly higher for patrilocal than for matrilocal groups for *h* but the differences are not statistically significant for MPD (Supplementary Table S9). Notably, the patrilocal groups HM1, HM4, and LS exhibit higher than average MPD values for the MSY (78.65–99.37, compared with the average of 51.74) and some matrilocal groups, e.g., KSK3 and TN1-TN2, show much lower MPD (6.92–20.69) than average (33.99) for mtDNA (Supplementary Table S9). Furthermore, the genetic diversity of all Htin groups are much lower than the other groups (Supplementary Fig. S9).

For genetic differences between populations revealed by AMOVA (Table 1), the MSY genetic variation among patrilocal populations is much higher than that for mtDNA (MSY: 27.43%, $P < 0.01$; mtDNA: 6.36%, $P < 0.01$), while for the matrilocal groups the mtDNA genetic variation is much higher, but still less than that for the MSY (MSY: 24.49%, $P < 0.01$, mtDNA: 16.46%, $P < 0.01$). Much stronger contrasting between-group variation is seen in two patrilocal groups, i.e., Hmong (MSY: 18.13%, $P < 0.01$, mtDNA: 0.53%, $P > 0.05$) and Lawa (MSY: 35.65%, $P < 0.01$, mtDNA: 7.78%, $P < 0.01$) and one matrilocal Htin group (MSY: 9.85%, $P < 0.01$, mtDNA: 25.71%, $P < 0.01$) (Supplementary Fig. S10). In contrast to our previous study [12], with the inclusion of the new KSK3 population, the matrilocal Karen shows more differentiation for the MSY than for mtDNA (MSY: 9.11%, $P < 0.01$, mtDNA: 5.85%, $P < 0.01$). When Hmong, IuMien, and Lahu from Thailand and Vietnam were combined, contrasting patterns of genetic variation between the MSY and mtDNA were still in accordance with expectations based on residence pattern, albeit the IuMien show only slightly higher between-group differentiation for the MSY than for mtDNA (Table 1).

Another potential effect of patrilocality vs. matrilocality is on the shared haplotypes between populations. If recent contact between populations is influenced by residence pattern, one would expect more MSY haplotype sharing among matrilocal groups than among patrilocal groups, and more mtDNA sharing among patrilocal groups than among matrilocal groups. However, the results for haplotype sharing between populations within matrilocal and patrilocal groups do not show a strong effect (Supplementary Table S10). Haplotype sharing for the MSY is slightly lower on average for patrilocal groups (0.15) than for matrilocal groups (0.18), in accordance with expectations, but haplotype sharing for mtDNA is also lower on average for patrilocal groups (0.15) than for matrilocal groups (0.22), which is not in accordance with expectations based on residence pattern.

Discussion

Our previous studies have focused on the genetic ancestry of the TK and AA groups in Thailand and Laos, here we investigate the less well-studied HM and ST speaking groups from Thailand, to gain more insights into the genetic history of MSEA. We sequenced ~2.3 mB of the MSY and complete mtDNA genomes of the HM and ST groups who are regarded as hill tribes from northern Thailand, as well as additional TK groups from northern and northeastern Thailand. Although we focus on the HM and ST groups, we note that the previously-observed general pattern of overall genetic homogeneity of Thailand TK groups [12] continues to be maintained with these additional TK groups, consistent with the idea that the TK language family spread via demic diffusion [13]. However, an additional insight arises when we compare the Thai TK data to similar mtDNA and MSY data from Vietnamese TK groups: [33] some of the TK-speaking groups from Vietnam (i.e., Nung, Tay, and Thai) are quite similar to the TK populations from northeastern Thailand (Phutai, Kalueang, and Black Tai) (Fig. 5). This is in agreement with historical evidence for a migration of the ancestors of the Phutai, Kalueang, and Black Tai from Vietnam through Laos during last 200 years [6, 34].

Genetic differences between the Hmong and IuMien groups and their origins

Previous studies of HM groups have reported sequences of the mtDNA hypervariable region 1 with some diagnostic coding SNPs to define haplogroups [35], and Y-STR variation and genotypes for Y chromosomal bi-allelic loci [36]. Here we analyze complete mtDNA and partial MSY sequences from five Hmong and two IuMien populations from Thailand; strikingly, we find significant differences between Hmong and Mien populations in Thailand, with the IuMien more similar to other populations (Figs. 3, 4, Supplementary Fig. S6, S7, Table 1), while the Hmong show genetic distinction that was not previously documented in Thai/Lao populations (Figs. 3, 4, Supplementary Figs. S6, S7 and Table 1).

Apart from the genetic distinction from their linguistic relatives, the IuMien, the Hmong in Thailand are genetically distinct from almost all other groups (Fig. 3). There are no shared mtDNA haplotypes between HM populations and other Thai/Lao populations, and only a few shared MSY haplotypes (Fig. 3a), moreover, they do not overlap with other groups in the MDS analysis (Fig. 4), suggesting that they add unique genetic profiles that were not found in the previous studies of Thai/Lao AA, TK, and ST groups [12]. This striking genetic divergence of Hmong populations in Thailand may reflect cultural isolation. Hmong communities have strong connections and prefer to marry with other Hmong groups rather than with non-Hmong groups [6]. In contrast, the IuMien have shared haplotypes and closer genetic relatedness with several TK-speaking groups, indicating more contact with other groups. These results may reflect the pronounced IuMien culture for adoption. Based on ethnographic accounts from the 1960s, around 10–20% of adult IuMien were adopted from other ethnic groups (both highland and lowland), in order to increase the size of their household and their family's influence [6, 9, 37]. Another factor behind the genetic similarity of IuMien with other East Asian populations could be admixture, as suggested by sharing of features between IuMien (but not Hmong) and Sinitic languages [38].

Although the proto-HM groups were suggested to have originated in central and southern China during the Neolithic Period [35], their greatest ethnolinguistic diversity is found between the Yangtze and Mekong rivers today. In general, the ages of mtDNA and MSY haplogroups characteristic for HM groups are during the Holocene to the late Neolithic Period (Supplementary Figs. S2, S3). This is consistent with archaeological and historical evidence that the proto-HM group might be linked with the Neolithic cultures in the Middle Reach of the Yangtze River in southern China, namely the Daxi culture (5300–6400 kya) and the Qujialing culture (4600–5000 kya) [35]. Our results are also consistent with a recent study of HM groups from Húnán, China [39], which identified lineages within MSY haplogroup O-N5 as specific to Hmong (and dated to ~2.33 kya) and mtDNA haplogroup B5a1c1a as correlating with Pahng and IuMien (and dated to ~9.80 kya). However, we do find B5a1c1a* and B5a1c1a1 specific to Hmong, but not IuMien, in Thailand. Overall, the coalescent ages of both MSY and mtDNA lineages are in the same range.

The origin of the Sino-Tibetan groups

It has been proposed that Neolithic people living at least 6 kya [40] in northwestern China were probably the ancestors of modern ST populations. It has also been suggested that ST languages originated among millet farmers, located in North China, around 7.2 kya [41] or 5.9 kya [42]. Linguistic evidence then suggests differentiation between Sinitic and Tibeto-Burman languages, and also between Tibetan and Lolo-Burmese languages, and southward and westward expansions of ST groups [41, 42].

Although an MSY lineage (O-M122*) was proposed to be characteristic of all modern ST populations [43], subsequent studies have found further differentiation, e.g., haplogroup O2a1c-002611, which is at high frequency in Han Chinese but found at very low frequencies in Tibeto-Burman populations [44, 45]. Also, autosomal STR genotypes differentiate Tibetan and Lolo-Burmese speaking groups [46].

ST-speaking groups in Thailand have not been studied in the same detail as those in China; here we analyzed three groups: Lisu, Lahu (Mussur), and Karen. Lisu and Lahu speak Lolo-Burmese languages, while the Karen languages belong to a different branch, Karenic [6]. Historical evidence indicates that Lisu and Lahu migrated from southern China through Myanmar to northern Thailand about 100–200 years ago [6]. The Karen claim to be the first settlers in Myanmar who migrated from southern China before the arrival of Mon and Burmese people, and the Karen groups in Thailand have been migrating from Myanmar started around 1750 A.D. due to the growing influence of the Burmese [47].

The genetic distances between the Lisu and Lahu are significantly different from zero for both mtDNA and the MSY, and they also share both mtDNA and MSY haplotypes (Fig. 3a, b), indicating recent contact and/or shared ancestry. The two Lahu populations (Black Lahu (MB) and Red Lahu (MR)) are genetically similar to one another and both are genetically distinct from the other populations (Figs. 3b, 4b–d). However, the Lisu do not differ significantly from many AA and TK populations (Fig. 3b), suggesting interactions between the Lisu and other populations. For the Karen, we have added an additional Karen population (Skaw (KSK3)) to the previously studied Karen populations; KSK3 has very low haplogroup diversity and

MPD values for the MSY (Fig. 2), suggesting strong genetic drift that has in turn increased their divergence from the other Karen populations (Fig. 4). This is in keeping with historical information: according to their oral history, KSK3 was founded ~60 years ago by just 18 households in a very remote region that is isolated from other Karen villages. The other Karen groups are genetically similar to several populations, and also share many basal mtDNA M haplogroups (M21a, M* and M91a) with neighboring Austroasiatic populations, especially the Mon, suggesting significant admixing (Supplementary Table S8). Previous studies based on autosomal STRs and SNPs also support the relatedness of the Karen and other AA groups in Thailand [47, 48].

The estimated coalescent ages of the predominant lineages in ST populations provide an upper bound for their divergence/contact from other groups. MSY haplogroup O2a2b1a1-Page23, equivalent to O-M117, which was previously reported to be abundant in TB groups in southwestern China and in Han Chinese [45], was dated to around ~2.41 kya in this study. However, this MSY lineage also occurs in HM, TK, and AA populations, reflecting recent shared ancestry and/or contact (Supplementary Fig. S2). The ages of mtDNA haplogroups prevalent in Lahu and Lisu, namely A13, B4e, D4j1a1, and G1c, are dated to ~6.74, ~2.21, ~2.49, and ~2.20 kya, respectively (Supplementary Fig. S3). Thus, the coalescent ages of many MSY and mtDNA lineages prevalent in ST groups are around the time of the Han expansion (~2.5 kya) [49].

Contrasting paternal and maternal genetic variation in patrilocal vs. matrilocal groups

Previously, postmarital residence pattern has been shown to influence genetic variation in the hill tribes of Thailand [9, 11, 12]. Our previous study had investigated four hill tribes: Lawa, Htin, Khmu and Karen [12]. Here we added data from four additional hill tribes (Hmong, IuMien, Lahu, and Lisu) for the most detailed investigation to date, comprising a total of 23 populations belonging to eight hill tribes. The Hmong, IuMien, Lisu, Lawa and Khmu are patrilocal (i.e., the wife moves to the residence of the husband after marriage) whereas the others are matrilocal. If postmarital residence pattern is having an influence on patterns of genetic variation, we would expect larger between-group differences and smaller within-group diversity for patrilocal groups for the MSY, and the same trends for matrilocal groups for mtDNA [11].

In general, the within-population genetic diversity values were not in agreement with expectations (Supplementary Table S9 and Fig. S10) whereas genetic differentiation between populations did go in the direction predicted by postmarital residence pattern (Table 1). However, when focusing on genetic differentiation within individual groups, the patrilocal Hmong and Lawa and the matrilocal Htin did fit with expectations, i.e., higher genetic differentiation among populations for the MSY than for mtDNA for the Hmong and Lawa, and the opposite for the Htin (Table 1). However, the matrilocal Karen show higher differentiation for the MSY than for mtDNA (Table 1), contrary to expectations and contrary to previous results based on four Karen populations [12]. The addition of the KSK3 population increases the between-population MSY genetic variance from 2.3 to 9.1%, while the between-population mtDNA genetic variance is relatively unchanged. The low MSY MPD value (Fig. 2c) and outlier position in the MSY MDS plots (Fig. 4), as well as their oral history, indicate a strong effect of genetic drift on MSY variation in KSK3, which might then mitigate any influence of postmarital residence pattern on MSY vs. mtDNA variation. Interestingly, overall we find less contrast between matrilocal and patrilocal groups than found previously for the hill tribes [9, 11, 12]. Presumably this is because of both more detailed sampling and higher-resolution analysis of the mtDNA and MSY genomes. And, this is not unexpected because while some studies find an impact of residence pattern on mtDNA/MSY variation, others do not [50, 51]. Many different factors can influence mtDNA/MSY variation, e.g., micro-evolutionarily factors such as genetic drift (as seen with the TN1, LW2 and KSK3 population), physical landscape, and other human cultural patterns, e.g., adoption in IuMien and cultural isolation and intermarriage in Hmong and Lawa [9, 12, 52, 53]; these can dilute or erase any potential impact of residence pattern.

Nonetheless, one striking pattern remains in our data, and that concerns NEA vs. SEA ancestry. Previous genetic studies supported a north-south division in East Asian peoples and with some spread of northern ancestry to the south [35, 49]. There are distinct differences in the mtDNA and MSY lineages of NEA vs. SEA populations [48, 49, 54], and here we also find a higher frequency of both mtDNA and MSY lineages of SEA origin than of NEA origin in most of the studied populations. In general, the SEA specific maternal lineages (B5*, F1a*, M7b* and R9b*) are at an average frequency of 38.28%, while NEA mtDNA lineages (i.e., A*, D* and G*) have an average frequency of 9.38% (Supplementary Table S8). The MSY haplogroups also show major SEA lineages (O1b*) predominating at an average frequency of 45.35%, and minor NEA lineages (C2e*, D-M174 and N*) at an average frequency of 8.33% (Supplementary Table S7).

However, the HM and ST groups are a dramatic exception to this general pattern of higher SEA than NEA ancestry for both paternal and maternal lineages (Fig. 1). The estimated NEA maternal ancestry of the HM groups is 11.94%, comparable to that of other Thai/Lao populations (average = 9.11%), while the average frequency of NEA

paternal lineages in HM groups is 24.72% (compared with the average frequency of 6.59% for other Thai/Lao populations). Conversely, in the ST groups we detect an average of 24.09% NEA maternal ancestry, which is much higher than the average NEA maternal ancestry for other Thai/Lao groups (7.57%), while the NEA paternal ancestry in ST groups is comparable to that in other Thai/Lao groups (11.95% vs. 7.88%).

Given that both HM and ST groups originated from southern China or northwestern China, it is likely that the ancestral HM and ST groups both had relatively high levels of NEA ancestry for both the MSY and mtDNA, as this has been reported for contemporary Chinese populations (Supplementary Table S5) [35, 39, 54, 55]. We suggest that there was subsequent contact with SEA groups as their ancestors migrated southward, with HM populations incorporating more SEA maternal than paternal lineages, and ST populations incorporating more SEA paternal than maternal lineages. This could be explained if the ancestral HM group was patrilocal (as all HM populations are today), and so subsequent interactions between the HM ancestors and SEA groups incorporated more SEA mtDNA lineages than MSY lineages into HM populations. Conversely, if the ancestral ST group was matrilocal (as the Karen and Lahu are today), subsequent interactions between ST ancestors and SEA groups would have incorporated more SEA MSY lineages than mtDNA lineages into ST populations. Matrilocality for ancient ST groups has also been suggested based on linguistic evidence [56]. The fact that some ST populations are now patrilocal (e.g., Lisu) while still exhibiting higher frequencies of NEA maternal lineages may then reflect recent changes from matrilocality to patrilocality.

However, the interaction between NEA and SEA groups was undoubtedly a complex admixture process that started in prehistoric times and has continued into historic times, and there are many factors in addition to residence pattern that could have led to sex-biased admixture [55]. Additional studies of both other ethnolinguistic groups (e.g., Akha, who show strong cultural practice preservation) as well as populations sampled from different villages of Lahu, Lisu, and IuMien from northern Thailand and Hmong and Karen (Skaw and Kayah) from western and central Thailand, are necessary to further investigate this pattern.

Conclusion

We have carried out the most extensive study to date, using high-resolution methods, of the maternal and paternal lineages in HM and ST speaking groups of northern Thailand. We find unexpected differences between the Hmong and IuMien, which may reflect different cultural practices, and genetic heterogeneity among ST groups. Compared with previous studies, we find less contrast in genetic diversity and differentiation between matrilocal and patrilocal groups among the hill tribes. However, a novel finding of this study is the contrast between HM and ST groups, both assumed to have origins in southern China, in frequencies of NEA maternal and paternal lineages. We suggest that this striking difference reflects ancestral patrilocality for HM groups vs. ancestral matrilocality for ST groups. Overall, our results further attest to the impact of cultural practices on patterns of mtDNA vs. MSY variation in human populations.

Acknowledgements We would like to thank all participants who donated their biological samples and coordinator, namely Siriluck Kanthasri, Sukhum Ruangchai and Nattapol Poltham. We thank Roland Schröder, Enrico Macholdt and Leonardo Arias for technical assistance. This study was supported by the Max Planck Society. WK and RS were also supported by the Thailand Research Fund (RSA6180058 and RTA6080001). JK was supported by Chiang Mai University. Open Access Funding Provided by Projekt DEAL.

Funding The Max Planck Institute for Evolutionary Anthropology; The Thailand Research Fund (RSA6180058 and RTA6080001)

Author contributions WK and MS conceived and designed the project; WK, RS MSr and JK collected samples; WK and TS generated data; WK, SS and AH carried out the data analyses; WK and MS wrote the paper with input from all coauthors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Eberhard DM, Simons GF, Fennig CD. Ethnologue: languages of the World. 23rd edn. Dallas: SIL International; 2020.
2. Higham C. Early Mainland Southeast Asia: from first humans to Angkor. Bangkok: River Books Press; 2014.

3. McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, et al. The prehistoric peopling of Southeast Asia. Science 2018;361:88–92.
4. Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science. 2018;361:92–5.
5. Blench R. Reconstructing Austroasiatic prehistory. In: Sidwell P, Jenny M editors. Handbook of Austroasiatic. Canberra: Pacific Linguistics; 2015.
6. Schliesinger J. Ethnic groups of Thailand: non-Tai-speaking peoples. Bangkok: White Lotus Press; 2000.
7. Penth H, Forbes A. The people of mountaintops. In: Penth H, Forbes A, editors. A brief history of Lan Na and the peoples of Chiang Mai. Chiang Mai: Chiang Mai City Arts and Cultural Centre Chiang Mai Municipality; 2004. pp. 247–54.
8. Ratliff MS. Hmong-Mien language history. Canberra: Pacific Linguistics; 2010.
9. Besaggio D, Fuselli S, Srikummool M, Kampuansai J, Castri L, Tyler-Smith C, et al. Genetic variation in Northern Thailand Hill Tribes: origins and relationships with social structure and linguistic differences. BMC Evol Biol. 2007;7(Suppl 2):S12.
10. Wang WSY. Three windows of the past. In: Mair VH, editor. The bronze age and early iron age peoples of Eastern Central Asia. Philadelphia: University of Pennsylvania Museum Publications; 1998. pp. 508–34.
11. Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet. 2001;29:20–1.
12. Kutanan W, Kampuansai J, Srikummool M, Brunelli A, Ghirotto S, Arias L, et al. Contrasting paternal and maternal genetic histories of Thai and Lao Populations. Mol Biol Evol. 2019;36:1490–506.
13. Kutanan W, Kampuansai J, Srikummool M, Kangwanpong D, Ghirotto S, Brunelli A, et al. 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. 2017;136:85–98.
14. Kutanan W, Kampuansai J, Changmai P, Flegontov P, Schröder R, Macholdt E, et al. Contrasting maternal and paternal genetic variation of hunter– gatherer groups in Thailand. Sci Rep. 2018a;8:1536.
15. Kutanan W, Kampuansai J, Brunelli A, Ghirotto S, Pittayaporn P, Ruangchai S, et al. New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. Eur J Hum Genet. 2018b;26:898–911.
16. Srikummool M. X-,Y-chromosomal and mitochondrial DNA variations of the Karen, Hmong and Iu Mien in the upper northern part of Thailand. PhD Thesis, Chiang Mai University, Chiang Mai, Thailand, 2005.
17. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010:pdb.prot5448.
18. Maricic T, Whitten M, Pääbo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 2010;5:e14004.
19. Excoffier L, Lischer H. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 2010;10:564–7.
20. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2016). http://www.R-project.org/
21. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. 2016. https://doi.org/10.1101/088716.
22. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, et al. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat. 2011;32:25–32.
23. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30:E386–94.
24. Drummond AJ, Suchard MA, Xie D, Rambaut A. A Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73.
25. Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015;25:459–66.
26. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 2016;538:201–6.
27. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 2012;9:772.
28. Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, Jagadeesan A, et al. The Y-chromosome point mutation rate in humans. Nat Genet. 2015;47:453–7.
29. Baele G, Lemey P, Bedford T, Rambaut A, SuchardMA, Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol. 2012;29:2157–67.
30. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol Evol. 2013;30:239–43.
31. 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. 2009;84:740–59.
32. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, et al. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet. 2012;90:675–84.
33. Macholt E, Arias L, Duong T, Ton N, Phong N, Schröder R, et al. The paternal and maternal genetic history of Vietnamese populations. Eur J Hum Genet. 2019;28:636–45.
34. Schliesinger J. Tai group of Thailand. Bangkok: White Lotus Press; 2001.
35. Wen B, Li H, Gao S, Mao X, Gao Y, Li F, et al. Genetic structure of Hmong-mien speaking populations in East Asia as revealed by mtDNA lineages. Mol Biol Evol. 2005;22:725–34.
36. Cai X, Qin Z, Wen B, Xu S, Wang Y, Lu Y, et al. Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. PLoS ONE 2011;6:e24282.
37. Jonsson H. Thailand Mien relations: mountain people and state control in Thailand. New York: Cornell University Press; 2005.
38. Blench R. Stratification in the peopling of China: how far does the linguistic evidence match genetics and archaeology? In: Alicia SM, Blench R, Ross MD, Peiros I, Marie L, editors. Human migrations in continental East Asia and Taiwan. Matching archaeology, linguistics and genetics. London: Routledge; 2008. pp. 105–32.
39. Xia ZY, Yan S, Wang CC, Zheng HX, Zhang F, Liu YC, et al. Inland-coastal bifurcation of southern East Asians revealed by Hmong-Mien genomic history. 2019. https://doi.org/10.1101/730903
40. Yang X, Wan Z, Perry L, Lu H, Wang Q, Zhao C, et al. Early millet use in northern China. Proc Natl Acad Sci USA 2012;109:3726–30.

41. Sagart L, Jacques G, Lai Y, Ryder RJ, Thouzeau V, Greenhill SJ, et al. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proc Natl Acad Sci USA 2019;116:10317–22.
42. Zhang M, Yan S, Pan W, Jin L. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. Nature 2019;569:112–15.
43. Su B, Xiao C, Deka R, Seielstad MT, Kangwanpong D, Xiao J, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet. 2000;107:582–90.
44. Wang CC, Yan S, Qin ZD, Lu Y, Ding QL, Wei LH, et al. Late Neolithic expansion of ancient Chinese revealed by Y-chromosome haplogroup O3a1c-002611. J Syst Evol. 2013;51:280–86.
45. Wang CC, Wang LX, Shrestha R, Zhang M, Huang XY, Hu K, et al. Genetic structure of Qiangic populations residing in the western Sichuan corridor. PLoS ONE. 2014;9:e103772.
46. Yao HB, Wang CC, Wang J, Tao X, Shang L, Wen SQ, et al. Genetic structure of Tibetan populations in Gansu revealed by forensic STR loci. Sci Rep. 2017;7:41195.
47. Kutanan W, Srikummool M, Pittayaporn P, Seielstad M, Kangwanpong D, Kumar V, et al. Admixed origin of the Kayah (Red Karen) in Northern Thailand Revealed by Biparental and Paternal Markers. Ann Hum Genet. 2015;7:108–22.
48. Xu S, Kangwanpong D, Seielstad M, Srikummool M, Kampuansai J, Jin L, et al. Genetic evidence supports linguistic affinity of Mlabri-a hunter-gatherer group in Thailand. BMC Genet. 2010;11:18.
49. Wen B, Li H, Lu D, Song X, Zhang F, He Y, et al. Genetic evidence supports demic diffusion of Han culture. Nature 2004b;431:302–5.
50. Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S, et al. Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. PLoS Genet. 2006;2:e53.
51. Arias L, Schröder R, Hübner A, Barreto G, Stoneking M, Pakendorf B. Cultural innovations influence patterns of genetic diversity in Northwestern Amazonia. Mol Biol Evol. 2018b;35:2719–35.
52. Wilkins JF, Marlowe FW. Sex-biased migration in humans: what should we expect from genetic data? Bioessays 2006;28:290–300.
53. Heyer E, Chaix S, Pavard S, Austerlitz F. Sex-specific demographic behaviours that shape human genomic variation. Mol Ecol. 2012;21:597–612.
54. Xue Y, Zerjal T, Bao W, Zhu S, Shu Q, Xu J, et al. Male demography in East Asia: a north-south contrast in human population expansion times. Genetics 2006;172:2431–9.
55. 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. 2004a;74:856–65.
56. van Driem. The diversity of the Tibeto-Burman language family and the linguistic ancestry of Chinese. Bull Chin Linguist. 2007;1:211–70.