Molecular Genetics and Genomics (2020) 295:579–589
https://doi.org/10.1007/s00438-019-01644-x

ORIGINAL ARTICLE

Paternal genetic history of the Yong population in northern Thailand revealed by Y-chromosomal haplotypes and haplogroups

Jatupol Kampuansai[1,2] · Wibhu Kutanan[3] · Eszter Dudás[4] · Andrea Vágó-Zalán[4] · Anikó Galambos[4] · Horolma Pamjav[4]

Received: 16 May 2019 / Accepted: 26 December 2019 / Published online: 13 January 2020

Abstract

We have determined the distribution of Y-chromosomal haplotypes and haplogroups in the Yong population, one of the largest and well-known ethnic groups that began migrating southward from China to Thailand centuries ago. Their unique mass migration pattern provided great opportunities for researchers to study the genetic links of the transboundary migration movements among the peoples of China, Myanmar and Thailand. We analysed relevant male-specific markers, such as Y-STRs and Y-SNPs, and the distribution of 23 Y-STRs of 111 Yong individuals and 116 nearby ethnic groups including the Shan, Northern Thai, Lawa, Lua, Skaw, Pwo and Padong groups. We found that the general haplogroup distribution values were similar among different populations; however, the haplogroups O1b-M268 and O2-M112 constituted the vast majority of these values. In contrast with previous maternal lineage studies, the paternal lineage of the Yong did not relate to the Xishuangbanna Dai people, who represent their historically documented ancestors. However, they did display a close genetic affinity to other prehistoric Tai-Kadai speaking groups in China such as the Zhuang and Bouyei. Low degrees of genetic admixture within the populations who belonged to the Austroasiatic and Sino-Tibetan linguistic families were observed in the gene pool of the Yong populations. Resettlement in northern Thailand in the early part of the nineteenth century AD, by way of mass migration trend, was able to preserve the Yong's ancestral genetic background in terms of the way they had previously lived in China and Myanmar. Our study has revealed similar genetic structures among ethnic populations in northern Thailand and southern China, and has identified and emphasized an ancient Tai-Kadai patrilineal ancestry line in the Yong ethnic group.

Keywords Y haplotypes and haplogroups · Yong ethnic groups in Thailand · Human demographic history

Introduction

Northern Thailand is geographically located on the crossroads of human migration among 4 Southeast Asian countries; China, Myanmar, Laos and Thailand. This region, in general, consists of fertile plains rimmed by mountain ranges. These fertile plains help to support a large number of the region's inhabitants. Archaeological and historical records indicate that this area has been occupied by various ethnic groups ever since the prehistoric period. According to a recent analysis of ancient genomes, the northern part of Thailand and nearby areas were involved in three important waves of human migration that occurred through Southeast Asia: the arrival of the hunter gatherers about 45,000 years ago, the Neolithic expansion of farmers from China approximately 4500 years ago, and the Bronze Age migration from

Communicated by Stefan Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-019-01644-x) contains supplementary material, which is available to authorized users.

✉ Horolma Pamjav
phorolma@hotmail.com

1 Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

2 Center of Excellence in Bioresources for Agriculture, Industry and Medicine, Chiang Mai University, Chiang Mai, Thailand

3 Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand

4 Department of Reference Sample Analysis, Institute of Forensic Genetics, Hungarian Institute for Forensic Sciences, Budapest, Hungary

China that occurred about 3000 years ago (Lipson et al. 2018).

Yong people are the native Tai speaking inhabitants of Mueang Yong (which literally means Yong city). This city was founded in the 13th–14th century AD by the Dai people who came from Jinghong (Chiang Rung), the capital of the Xishuangbanna Autonomous Prefecture of China. In the past, the city was located in the Chinese territory, but now belongs to the Shan state of the Union of Myanmar (Penth and Forbes 2004) (Fig. 1). The largest influx of the migration of the Yong people into Northern Thailand occurred in 1805 AD. After a period of Burmese decolonization, the King of Siam (Thailand) ordered the northern Thai rulers to launch several military raids to the north and west into Myanmar and Southern China to recruit inhabitants for the purposes of re-populating the abandoned cities of Thailand (Ongsakul 2005). About 10,000 Yong people moved their entire households to resettle in Lamphun Province of Northern Thailand. The Yong have now become the majority ethnic group of Lamphun Province and still maintain their culture, traditions and dialects, which are uniquely different from other ethnic groups.

The study of the mitochondrial DNA hypervariable region I (HVI) revealed that the Yong are in fact of the same ethnicity as the Lue people of northern Thailand. Notably, both the Yong and the Lue shared Xishuangbanna Dai genetic ancestry. However, the Yong, who have traditionally practiced mass migration, exhibited closer genetic affinity to the Dai than to the Lue people. The founder effects revealed their impact through a sudden reduction of effective population size and by shaping the genetic differentiations that exist among the fragmented Lue population of Northern Thailand. The maintenance of the ancestral Dai maternal genetic structure among the Yong people might be due to their mass migration, as well as to other related scenarios in which populations moved from one place to another within a short period of time (Kampuansai et al. 2016). By identifying the unique mass migration pattern that can preserve ancestral genetic information, the Yong can provide researchers with expanded opportunities to study the genetic links of a population who has migrated across geographic borders from China and Myanmar to Thailand. We, therefore, aim to present an exhaustive paternal genetic history of the Yong population in Northern Thailand using Y-chromosomal genetic markers.

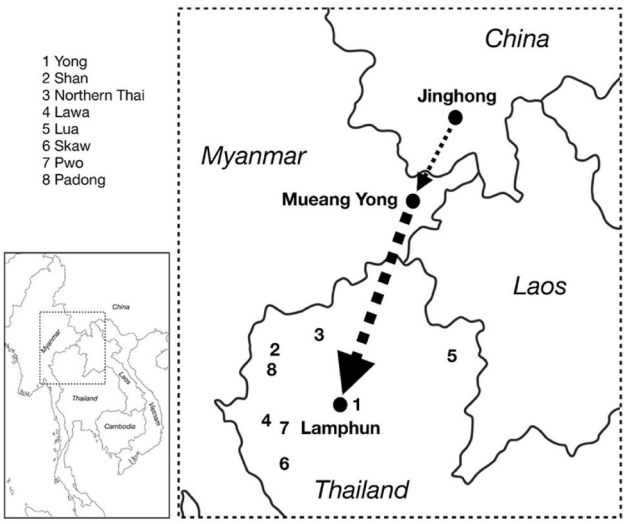

Fig. 1 Locality of sampling populations and the historic migration route of the Yong people. The size of the arrow is consistent with the estimated number of immigrates

The human Y chromosome provides a unique system of haplotypes and haplogroups. Since most of its region is non-recombining, it can avoid the effects of meiotic recombination. Therefore, combinations of allelic states of markers along the Y chromosome are transmitted directly from father to son and from generation to generation. Data of the Y chromosome are generally gathered from single nucleotides polymorphisms (SNP) and short tandem repeat (STR) loci. The biallelic SNP markers have low mutation rates, which offer the advantage of robust haplogroup classification, while STRs are more rapidly mutated that can support the phylogeny of haplotype reconstruction. Thus, the entire Y chromosome preserves a simpler record of male history and how they have uniparentally inherited their lineage. This can actually benefit the study of modern human origins, migrations, and population differentiations (Underhill et al. 2000). This is why we have focused on the patrilineal lineage of the Yong, one of the largest and most well-known ethnic groups who have migrated southward from China to Thailand. Through a comparison of the Y-chromosomal haplotypes and haplogroups of several ethnic populations, an exhaustive genetic history of the Yong people and nearby ethnic populations in northern Thailand would be revealed.

Materials and methods

Materials

A total of 111 unrelated male samples acquired from 6 Yong villages in Lamphun Province of Northern Thailand were studied. We also collected samples from nearby populations for comparison including those of the Shan, Northern Thai, Lawa, Lua, Skaw, Pwo and Padong people (Table 1). Note that the Skaw, Pwo and Padong people are subgroups of the Karen ethnic tribe. White blood cell lysate solutions of each sample were obtained from previous studies (Kutanan and Kangwanpong 2010; Kampuansai et al. 2016; Lithanatudom et al. 2016). Informed consent was obtained from all individuals included in the study. The number of samples

Table 1 Language family, number of samples and genetic diversity of the studied populations

Population	Language family	*n*	*k*	*h*	MPD
Yong	Tai-Kadai	111	102	0.998 ± 0.002	15.512 ± 6.978
Shan	Tai-Kadai	12	12	1.000 ± 0.034	16.197 ± 7.774
Northern Thai	Tai-Kadai	14	14	1.000 ± 0.027	15.396 ± 7.322
Lawa	Austroasiatic	12	9	0.909 ± 0.080	9.030 ± 4.475
Lua	Austroasiatic	26	15	0.945 ± 0.026	9.431 ± 4.473
Skaw	Sino-Tibetan	24	23	0.996 ± 0.013	15.094 ± 6.993
Pwo	Sino-Tibetan	15	13	0.981 ± 0.031	14.352 ± 6.818
Padong	Sino-Tibetan	13	11	0.962 ± 0.050	13.885 ± 6.669

n Number of samples, *k* number of observed haplotypes, *h* haplotype diversity, *MPD* mean number of pairwise difference

for each population was dependent upon the number of volunteers who were over 20 years old and those who had no ancestors from any other ethnic group for at least the last 3 generations. Prior to sample collection, information on linguistic, cultural aspects and family pedigree were obtained using form-based oral interviews.

According to the relevant policy concerning the publication of forensic population genetic data, all newly obtained male samples acquired from Northern Thailand published herein were sent to the YHRD database for external evaluation prior to publication. The samples then received the following YHRD accession numbers: YA004161-YA004168. The data on these new populations can be accessed at www.yhrd.org by population name, contributor name or accession number.

Methods

Testing of Y-STR and Y-SNP markers

Genomic DNA was extracted from white blood cell lysate solutions using the inorganic salting out protocol (Seielstad et al. 1999). The samples were quantified using the ABI 7500 Real-time PCR System (Life Technologies, Foster City, CA, USA). DNA was amplified using the PowerPlex Y23 amplification kit including 23 Y-STR loci according to the manufacturer's instructions. Fragment sizes and allele designations were determined using a Genetic Analyzer ABI3130 (Applied Biosystems, Foster City, CA, USA) and GeneMapper ID-X v.1.2 software.

When testing Y-SNP markers, amplifications of 3–5 ng genomic DNA were performed in ABI 7500 and GeneAmp 9700 thermal cyclers, with *TaqMan* probes using the programs designed by ABI. The relative degree of fluorescence of the PCR products was analysed on an ABI 7500 Real-time PCR System using its SDS software, as was described in the ABI manufacturer's manual. A complete list of primers and *TaqMan* assays for SNP markers has been described elsewhere (Pamjav et al. 2017). The nomenclature of haplogroups followed the ISOGG 2014 Y-DNA haplogroup tree due to recent and new additions to the tree that had been uncovered by YCC (Karafet et al. 2008).

Phylogenetic study

Haplotype and haplogroup frequencies and their diversity values were calculated according to the previously described method (Nei 1973). To examine the STR variations within the haplogroups, relational Median-Joining (MJ) networks were constructed using the Network 5.0.0.0 program (Bandelt et al. 1999). Data of 10 STR loci, including DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439, were included in this network analysis. Repeats of the DYS389I locus were subtracted from the DYS389II locus according to common practice. Within the network program, the rho statistic was used to estimate the time to the most recent common ancestor (TMRCA) of haplotypes within the haplogroups (Bandelt et al. 1999). We used a STR mutation rate of 6.9×10^{-4}/locus/25 years, as was described by Zhivotovsky et al. (2004).

Genetic relationships

Pairwise *Fst* genetic distances were calculated based on haplogroup frequencies using Arlequin 3.5 software (Excoffier and Lischer 2010). In addition, 27 published populations (Online Resource 1, ESM_1), who were made up of various linguistic families that shared geographic and historical features with the group being studied, were selected for comparison (Hammer et al. 2006; Zhong et al. 2010; Kim et al. 2011; Brunelli et al. 2017). For these purposes, haplogroups were combined into the following groups; ABDF*, C, E, G, H, I, J, K*, L, N, O, Q, R, and T, so that the published sources could be used for the basis of comparison.

Y-STR-based pairwise *Rst*-genetic distances were computed from haplotype frequencies using the online AMOVA program available on YHRD.org (release 60) (Willuweit and

Roewer 2015). Y-STR haplotype data of 23 populations that had been genotyped with PowerPlex Y23 and submitted to the YHRD database (Online Resource 2, ESM_2) were integrated into this analysis. Kruskal's non-metric multidimensional scaling (MDS) plots, based on the genetic distance matrix, were constructed with the XLSTAT 6.502 trial version (Addinsoft Inc. NY, USA). Hierarchical genetic variances of the groups of the studied populations according to their linguistic families were investigated using the AMOVA method as implemented with Arlequin 3.5 software (Excoffier and Lischer 2010).

Results

Genetic diversity and haplogroup distribution

The results of the haplotypes of the 23 Y-STR loci and haplogroups of 227 males are available at the Online Resource 3 (ESM_3). One hundred and ninety-eight distinct Y-STR haplotypes were observed. Among them, 197 types were unique to each population, whereas one haplotype was shared by the Skaw and Pwo populations. Of the 197 unique haplotypes, 19 were shared by two or more individuals within populations, whereas the remaining 178 haplotypes were only observed in one individual. Both haplotype diversity (h) and the mean number of pairwise differences (MPD) were found to be the lowest in the Lawa population (0.909 ± 0.080 and 9.030 ± 4.475, respectively), while the highest values were observed in the Shan population (1.000 ± 0.034 and 16.197 ± 7.774, respectively) (Table 1). The overall haplotype diversity for the 8 populations was 0.9984 ± 0.0007. However, the average haplotype diversity of the populations who belonged to the Tai-Kadai (0.9989 ± 0.0011) and the Sino-Tibetan (0.9940 ± 0.0056) linguistic families were relatively higher than those of the Austroasiatic (0.9659 ± 0.0150) group.

The frequencies of haplogroups in each studied population are presented in Table 2. Overall, the large haplogroups were O1b-M268 (former name O2-M268) and O2-M122 (former name O3-M122), while other haplogroups had frequencies of less than 5%. The haplogroup O1b-M268 was present in all of the studied populations and ranged from 29.2% in Skaw to 100% in Lua. Interestingly, all 15 Y-STR haplotypes observed in the Lua were classified in the haplogroup O1b-M268. Nevertheless, it should be noted that the observed haplotype percentages of some populations were based on small sample sizes (less than 20) and/or on genetic drifts, thus the findings should be further interpreted with precaution. Haplogroups C, C3, H, R1* and R2 were only distributed within the Tai-Kadai speaking populations, while the G1 was specific to the Sino-Tibetan group. Notably, the haplogroup N-M231 (xN1, N2, N3) was observed within the Yong, Shan and Skaw populations.

Phylogenetic study

Based on Y-STR haplotypes, networks were constructed for the haplogroups O1b-M268 and O2-M122, which were exhibited at the highest frequency in the studied populations. The MJ network of 116 O1b-M268 samples was generated. However, we found that this haplogroup was very diverse and produced several overlapping reticulations. Thus, we used the post-processing MP calculation method within the network-building algorithms. A reconstructed MJ network of O1b-M268 is presented in Fig. 2. Although, most of the studied samples did not share common haplotypes, there was one haplotype that was shared by 3 populations (1 Yong, 1 Shan and 1 Pwo). We marked this as a common haplotype. Surprisingly, we found a specific founder haplotype for the Lawa population that was defined as DYS391 = 11, which

Table 2 Haplogroup distribution in the studied populations from northern Thailand

Haplogroup	Mutation	No. of samples (%)								
		Yong	Shan	N. Thai	Lua	Lawa	Skaw	Pwo	Padong	Total
C	M216	1 (0.9)		1 (7.1)						2 (0.9)
C3	M217	5 (4.5)								5 (2.2)
D	M174	3 (2.7)	2 (16.7)	1 (7.1)		1 (8.3)			3 (23.1)	10 (4.4)
G1	M285						4 (16.7)	4 (26.7)		8 (3.5)
H	M52	1 (0.9)								1 (0.4)
N	M231	5 (4.5)	1 (8.3)				3 (12.5)			9 (4.0)
O1a	M119	8 (7.2)	1 (8.3)				1 (4.2)			10 (4.4)
O1b	M268	47 (42.3)	6 (50.0)	9 (64.3)	26 (100.0)	10 (83.3)	7 (29.2)	5 (33.3)	6 (46.2)	116 (51.1)
O2	M122	41 (36.9)	2 (16.7)	1 (7.1)		1 (8.3)	9 (37.5)	6 (40.0)	4 (30.8)	64 (28.2)
R1*	M173		1 (7.1)						1 (0.4)	
R2	M124			1 (7.1)						1 (0.4)

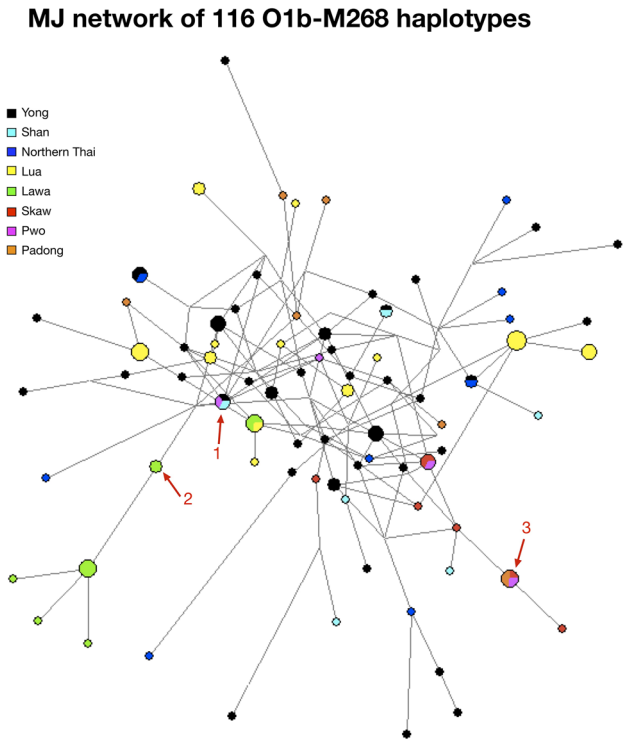

Fig. 2 Median-Joining Network (MJ) of 116 O1b-M268 haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual. (Arrow 1: common haplotype; Arrow 2: Lawa founder haplotype; Arrow 3: Sino-Tibetan founder haplotype)

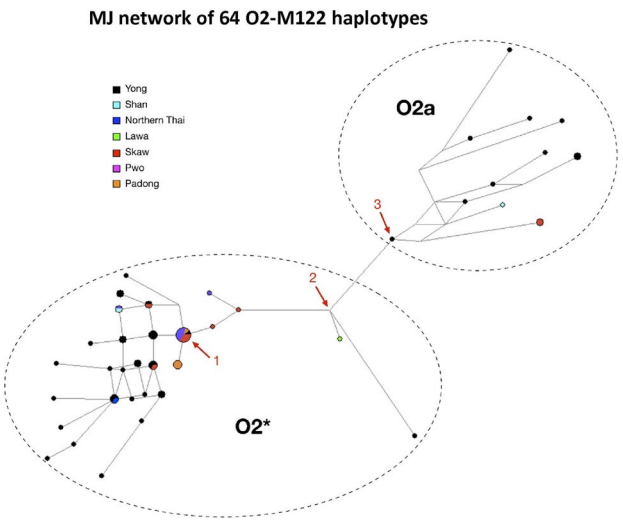

Fig. 3 Median-Joining Network (MJ) of 64 O2-M122 haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual. (Arrow 1: O2* founder haplotype; Arrow 2: ancestral haplotype; Arrow 3: O2a founder haplotype)

was two molecular steps away from the DYS391 = 9 of the founder haplotype. The age of the accumulated STR variations within the O1b-M268 lineage for 116 samples was estimated to be 13,242 ± 2341 ybp, considering the common haplotype to be the founder haplotype (indicated by arrow 1 in Fig. 2).

The MJ network of the 64 O2-M122 samples was generated and is presented in Fig. 3. Two clusters were clearly separated within the network. We did not test our samples deeper than sub-haplogroup O2-M122 because of a lack of downstream SNPs; however, we included some individuals in the analysis with known O2-M122 haplotypes that were taken from the data published by Kim et al. (2011). Based on the relevant network, we identified the clusters as subgroups O2* and O2a. The founder haplotype of the sub-haplogroup O2* was shared by 9 samples (1 Yong, 3 Skaw, 4 Pwo, and 1 Padong), which were almost exclusively restricted to the Sino-Tibetan linguistic group. The founder haplotype of the sub-haplogroup O2a was a Yong male. Interestingly, all derived haplotypes were derived from the Yong samples, except for one Shan and two Skaw males. For this network, we chose a median node that was present in the centre of the network (arrow 2 in Fig. 3) as the ancestral haplotype for TMRCA calculations, although there was no sample at this node. The age of the accumulated STR variations within the O2-M122 lineage for 64 samples was estimated to be 27,059 ± 6735 ybp, while the TMRCAs for the sub-haplogroups O2* and O2a were 26,000 ± 8225 ybp and 31,213 ± 7672 ybp, respectively.

Genetic relationship

Fst-genetic distances and *p* values based on haplogroup frequencies among 27 populations were calculated as is shown in the Online Resource 1 (ESM_1) and presented as an MDS plot (Fig. 4). The MDS plot revealed that the Yong people were closely related to the Tai-Kadai and Austroasiatic speaking populations, especially those of the Lawa2, Shan2 and Zhuang people. In general, populations who belonged to the Tai-Kadai, Austroasiatic and Sino-Tibetan linguistic groups were clustered together on the right side of the plot, while the Altaic speaking populations were differentiated from this group on the top of the MDS plot. Significant *p* values of *Fst*-genetic distances between each pair of Altaic populations were also reflected in the genetic heterogeneity within this group. The Lawa3 people of Chiang Mai Province in Northern Thailand was considered an outlier as they were determined to be extremely different from the others, which reflected their unique genetic structure. The influence of the genetic drift within this Lawa people had been reported previously (Kutanan et al. 2011a).

We also constructed non-metric MDS analysis based on Y-chromosomal haplotypes that consisted of 23 STR loci that were available from 23 populations (http://www.yhrd.

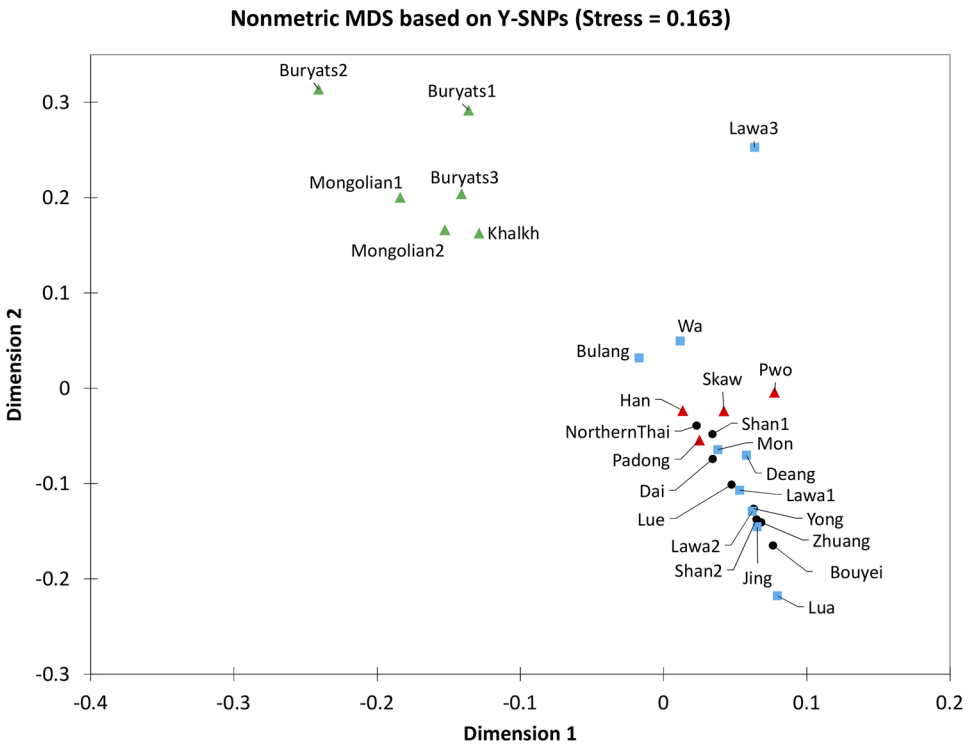

Fig. 4 Multidimensional scaling (MDS) plot constructed based on the *Fst* genetic distances of the Y haplogroup frequencies of 27 populations compared. (Stress = 0.163). Symbols represent their languages as followings: black circles: Tai-Kadai, blue cubes: Austroasiatic, red triangles: Sino-Tibetan, green triangles: Altaic (color figure online)

org). As shown in Fig. 5, the Yong were closely related to most of the Tai-Kadai speaking populations, which were concentrated in the centre of the plot. The non-significant *Rst*-genetic distances between the Yong and the Northern Thai, Shan, and Zhuang populations (Online Resource 2, ESM_2) were in correspondence with the MDS genetic affinity. Some Sino-Tibetan and Hmong-Mien populations exhibited close relationships to the Tai-Kadai group as they were scattered around and they appeared an outlier of Tai-Kadai cluster. Two studied Austroasiatic populations, Lawa

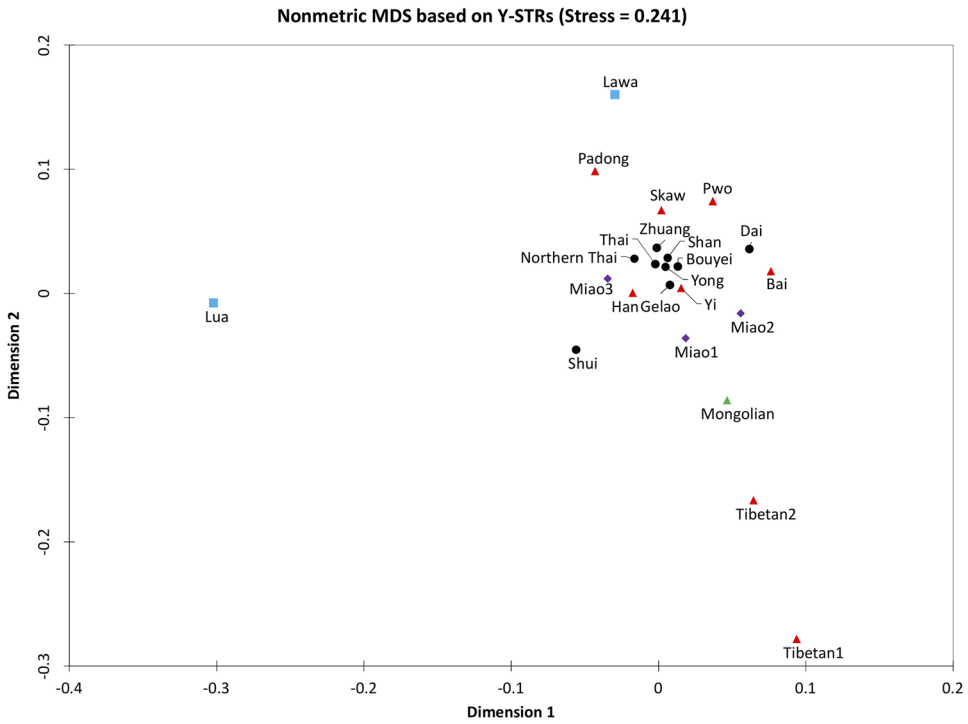

Fig. 5 Multidimensional scaling (MDS) plot constructed based on the *Rst*-genetic distances of 23 STRs-based Y haplotype frequencies of 23 populations compared (Stress = 0.24). Symbols represent their languages as followings: black circles: Tai-Kadai, blue cubes: Austroasiatic, red triangles: Sino-Tibetan, green triangles: Altaic, purple diamonds: Hmong-Mien (color figure online)

and Lua, displayed highly divergent positions from the other populations as they exhibited significant genetic differences.

The hierarchical genetic structure of the studied populations was investigated by AMOVA. When all samples were combined, 90.16% of the genetic variations were found within the populations, whereas 9.84% of the variations were present among them. The genetic variations between populations belonged to the Tai-Kadai linguistic family (0.42%) was smaller than those of the Austroasiatic (24.20%) or Sino-Tibetan (2.12%) people. The proportion of genetic variations attributed to the differences among three groups amounted to 7.40% and was considered significant ($p < 0.01$) (Table 3).

Discussion

The main objective of this study was to reveal the paternal genetic history of the Yong population and to compare the findings to various ethnic groups located in Northern Thailand. Although, we included a small sample size (less than 20 samples) of some ethnic groups involved in this study, the overall pattern of haplogroup distributions in our studied populations were similar, with haplogroups O1b-M268 and O2-M122 appearing at the highest frequencies (Table 2). Both of these O lineages predominantly range across China and Southeast Asia. These regions are known to be the original homeland of several O lineages and sub-lineages such as O1b-M268, O2-M122, O2a1-M95 (Yan et al. 2011; Zhang et al. 2015; Poznik et al. 2016). The observed O1b-M268 and O2-M122 haplogroups were suggested to have different origins of expansion, namely O1b-M268 in Chinese people and O2-M122 in Southeast Asian. Since our studied populations include the native people of East/Southeast Asia and those who reside in the buffer zone between mainland China and nearby Southeast Asian countries, a high frequency of the O haplogroup was observed as expected and was almost equal in some populations such as those of the Yong, Skaw and Pwo. Appearance of the R1* and R2 sub-haplogroups, which are usually found in Europe, Central Asia and South Asia (Underhill et al. 2010), was restricted to the northern Thai population. Although, there was no directly historical connection between the northern Thai population and people of the haplogroup R predominant area, the genetic admixture might be responsible for this scenario. The northern Thai people, who are locally known as the Khon Mueang people, have been reported to be admixed descendants of different ethnic groups (Kutanan et al. 2011b). Thus, the gene flow of different lineages into the northern Thai population, including the R haplogroups, was not an unexpected event (Table 2).

We also detected the haplogroups C-M216, C-M217, and D-M174, which accounted for 7.49% of our samples. Haplogroup C was observed in the Yong and northern Thai populations, while haplogroup D was found among the Lawa, Yong, Shan and northern Thai populations. Although, downstream SNPs of the C subclades had not been genotyped in our study, there is the possibility that distribution of the haplogroup C in our samples was due to the Mongolian empire expansion, especially in the 13th century AD after the rise of Genghis Khan. Varying degrees of the Mongolian patrilineal genetic imprint had been observed in several populations living in Europe, South/Central Asia and the Indian subcontinent (Zhong et al. 2010; Bai et al. 2014). The area comprised of Southern China and Northern Myanmar was subdued by Mongolians originating from 1292 AD until the middle of the 14th century, when Genghis Khan's power over China finally collapsed (Schliesinger 2001). The paternal lineage of ethnic groups who formerly resided in this region, like the Yong and northern Thai people, might have been influenced by the Mongolian admixture and contributed to the presence of C haplogroups in these populations, because this was the most well-known historical process. Haplogroup D-M174 is believed to have originated from Asia/Central

Table 3 Analysis of Molecular Variance (AMOVA) results based on 23 Y-STRs data

Groups	No. of population in each group	Within populations		Among populations within groups		Among groups	
		Variance (%)	*Fst*	Variance (%)	*Fsc*	Variance (%)	*Fct*
All samples	8	90.16	0.0984*	9.84			
TK	3	99.58	0.0042	0.42			
AA	2	75.80	0.2420*	24.20			
SN	3	97.88	0.0212	2.12			
TK/AA	3/2	84.29	0.1571*	4.80	0.0539*	10.91	0.1091
TK/SN	3/3	96.15	0.0385*	1.01	0.0104	2.84	0.0284
AA/SN	2/3	76.79	0.2321*	7.59	0.0900*	15.62	0.1562
TK/AA/SN	3/2/3	88.64	0.1136*	3.96	0.0428*	7.40	0.0740*

TK Tai-Kadai, *AA* Austroasiatic, *SN* Sino-Tibetan

*Statistical significant at $p < 0.01$

Asia some 60,000 years before the present time (Shi et al. 2008; Karafet et al. 2008). It has been identified in high frequencies today among populations living in Japan and on the Tibetan Plateau, while low-to-moderate frequencies were identified among Chinese ethnic groups, for example the Zhuang in China and among several minority populations of Sichuan and Yunnan that speak Tibeto-Burman languages and reside in close proximity to the Tibetans (Karafet et al. 2008). D-M174 is also present in Mongolia and Southeast Asia (Hammer et al. 2006), while D-M174* is found in Central Asia (Karafet et al. 2001). Based on the outcomes of our previous study, we detected haplogroup D-M174 in the Uzbeg Madjar population (Central Asia) in 26% of the population samples, which was in accordance with our observations described above (Bíró et al. 2015). We also observed haplogroup N-M231 (xN1, N2, N3) in the Yong, Shan and Skaw at low frequencies. Based on the MJ network constructed from searching the YHRD ancestry information, these samples probably belong to haplogroup N-F2930 (Online Resource 4, ESM_4), which is known to be frequent in Chinese Han and Japanese populations. This observation is consistent with the results published by Ilumäe et al. (2016), because the Yong, Shan and Skaw ethnic groups had close contact with the Han Chinese when they were located in China hundreds of years ago.

We observed a mysterious event in the ethnic-specific haplogroup, G1-M285, which found in the Skaw and Pwo populations account for 16.7% and 26.7% of their genetic makeup, respectively. Haplogroup G1 is exhibited at a high frequency in West Asia, especially in the Madjar tribe of Kazakhstan and was marked as an ancient genetic link between the Iranic speakers of South-West Asia and populations located in the Central Asian steppes, where Iranian speakers predominated during the first and second millennia BC (Bíró et al. 2009; Balanovsky et al. 2015). Until today, the origin of the Karen ethnic groups, Skaw and Pwo, is somewhat obscure, because early historical records are lacking. They appeared to migrate along isolated mountain ranges and had left little or no traces in other ethnic histories. Various theories on the origin of the Karen have been proposed. The most striking theory associated with their origin is that they had originated from around the Gobi Desert or upper the Yellow River valley in China. However, their language and personal characteristics indicate that they probably originated from near the Tibetan Plateau, but none live there today. Nowadays, there are large numbers of Karen people living in Myanmar (more than 2 million), while some groups have migrated into Thailand within this century (Schliesinger 2000; Besaggio et al. 2007). To find out the best possible relationship between haplogroup G1 samples in Northern Thailand and those of Central/Inner Asia, data of G haplotypes were retrieved from our previous studies (Bíró et al. 2009, 2015) for the purposes of comparison. The MJ network of 82 G haplotypes (4 Skaw, 4 Pwo, 39 Kazakh Madjar, 6 Csángó, 7 Uzbek, 1 Mongolian, 4 Székely, 17 Hungarian) was constructed (Online Resource 5, ESM_5). As can be seen in the network, haplogroup G1 carriers among the Skaw and Pwo people are complete outliers from the Central and Inner Asian samples that are included in the study. The age of the accumulated STR variations within G1-M285 and G2-L156 lineages are estimated to be 1486 ± 695 ybp and $13{,}312 \pm 2948$ ybp, respectively. Thus, the G1 lineages observed in Northern Thailand had a very deep common ancestry with populations in the Central and Inner Asia regions, or they might be indicative of different genetic histories. Our G1 samples might have inherited this lineage from the descendants who practiced traditional methods of agriculture. These methods were probably associated with the wet rice cultivation process that was widespread during the Neolithic period and was later routinely practiced in Mainland Southeast Asia and the Indian subcontinent (Bellwood 2011; Fuller 2011). The observation of G1-M285 in the Skaw and Pwo populations would open investigations into a new issue on the origins and the genetic affinity of the Karen people; however, deep-resolution haplogroup analyses with more diverse population comparisons need to be conducted in future studies to draw more robust conclusions. However, until that time, due to a lack of data, the history of haplogroup D-M174 as a genetic legacy of the farmers remains a mystery.

With a focus on the Yong ethnic ancestry, the origin of the Yong people could be roughly divided into 3 periods according to historical records. The first period was the founding of the Mueang Yong in the 13th–14th centuries AD. Scholars all agree that the Yong and Xishuangbanna Dai peoples shared a close common ancestry. The overlapping of maternal mitochondrial haplotypes between the Yong and Dai people has confirmed that the Yong were the descendants of the Xishuangbanna Dai, a majority Tai-Kadai speaking people in Southern China (Kampuansai et al. 2016). In contrast with the maternal genetic history, our paternal SNP-, and STR-based MDS (Figs. 4, 5) revealed that, among the Tai-Kadai speaking populations, the Zhuang, Bouyei and Shan exhibited much closer genetic affinity to the Yong than did the Xishuangbanna Dai people. The Zhuang and Bouyei have been acknowledged as one of the oldest Tai-Kadai groups of China, and have been living in the area for more than 2000 years (Chaoxiong 2005). Thus, we suspect that the patrilineal lineages of the Yong should be linked with the ancient ancestry of the Tai-Kadai linguistic family. However, the genetic admixture within the Yong in the later ages should not be neglected and should be discussed further during the second period. It is important to note that, beside the majority Yong people, the first recognised indigenous inhabitants of Mueang Yong were the Lawa people who were documented to occupy this area since the prehistoric

era. The arrival of the Dai from Jinghong in the 14th century AD prompted the Lawa to be exiled to rural areas (Malasam 2001). Our study of Y-chromosomal lineages supports the belief that there was no assimilation between the Yong and Lawa in this period, as has been reflected in the significant *Rst*-genetic distances (Online Resource 2, ESM_2). A phylogenetic analysis also supported this assumption as no shared haplotypes were present between these two ethnic groups, while the closest Lawa founder O1b-M268 haplotype had been separated from the Yong since 2898 ± 1449 ybp (Fig. 2). This would have been earlier than the arrival of the Yong founders. Nevertheless, the Lawa had been reported to pass along a strong founder effect and fragmentation (Kampuansai et al. 2012), our Lawa samples might not cover the prehistoric group. Genetic affinity between the Yong and other Lawa populations, especially in Northern Myanmar, would need to be further investigated.

The second period involved the contact between various ethnic groups. After being founded in the 14th century AD, the city of Mueang Yong became associated with neighbourhoods within a short period of time because of its locality. It was situated along a trading intersection and along the marching army routes to several large cities in China, Myanmar, Laos and Thailand (Malasam 2001). The Yong potentially had contact with different ethnic peoples in the region such as the Shan, Dai, Han, Burmese and some Karen people. Intermarriages were an inevitable circumstance and this led to the genetic admixture that occurred among them. Close genetic relationships among the Tai-Kadai, Austroasiatic and Sino-Tibetan paternal lineages that had been revealed in our SNP-, and STR-based MDS plots (Figs. 4 and 5) probably occurred due to the admixture event during this period. Genetic homogeneity in the Tai-Kadai group, which was reflected by high haplotype diversity, tight clusters in the MDS, and a low degree of AMOVA genetic variance (Table 1 and ESM_4; Figs. 4, 5), hinted that male genes flowed within this linguistic family much more often than in the other groups.

The third period involved the mass migration of the Yong into northern Thailand (Fig. 1). Early in the 19th century, large groups of the Yong people were forced into migration and resettled in Lamphun Province of Thailand. High numbers of the Yong's specific Y haplotypes support the determination that the Yong in Northern Thailand have still maintained their unique genetic ancestry, with a low degree of genetic admixture with other ethnic groups. Previous mtDNA studies unveiled that the unique mass migration pattern of the Yong had shaped the different genetic structures of the Lue, their close relatives who practised founder migration trends (Kampuansai et al. 2016). Our Y-chromosomal results support the claim that the Yong and Lue in Northern Thailand share a deep common ancestry, which is affirmed by a close genetic affinity in SNP-based MDS and non-significant *Fst*-genetic distances (Fig. 4 and ESM_1). Unfortunately, there was no available Y-STRs data on the Lue in the YHRD database; therefore, the recent genetic history of the Yong and Lue in Northern Thailand during these past two centuries could not be compared. In fact, the Y-STR data of the ethnic groups in Southeast Asia that had been submitted to the YHRD database are very limited when compared with the data of numerous members of indigenous and migrant ethnic groups found in this region. The boosting of the Y-chromosomal data archive in Southeast Asia will significantly benefit the study of modern human origins and contribute to the preservation of the genetic history of certain ethnic groups in this important part of the world.

In conclusion, our Y-chromosomal data revealed that the ethnic populations in northern Thailand shared a similar paternal genetic structure with the people of China and Southeast Asia, with some specific unexplained lineages among the Karen subgroups. However, we emphasize that the sampling numbers of several ethnic groups in this study are small. Larger numbers of samples in future studies will be needed to clearly identify the genetic ancestry and migration of the ethnic groups of this region. Our highlighted Yong population had inherited the patrilineal ancient ancestry of the Tai-Kadai people in China, while the impact of mass migration pattern had helped to preserve their ancestral genetic background until today.

Acknowledgements The authors thank all volunteers and village chiefs for their participation in the sample collection process. This work was financially and technically supported by the former Network of Forensic Science Institutes, Institute of Forensic Medicine, DNA laboratory (2015), Budapest, Hungary. JK gratefully acknowledges the partial support provided by Chiang Mai University, Thailand, and Tempus Public Foundation, Hungary. WK was provided funding from the Thailand Research Fund (Grant No. RSA6180058).

Compliance with ethical standards

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

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Bai H, Guo X, Zhang D, Narisu N, Bu J, Jirimutu J, Liang F, Zhao X, Xing Y, Wang D, Li T, Zhang Y, Guan B, Yang X, Yang Z, Shuangshan S, Su Z, Wu H, Li W, Chen M, Zhu S, Bayinnamula B, Chang Y, Gao Y, Lan T, Suyalatu S, Huang H, Su Y, Chen Y, Li W, Yang X, Feng Q, Wang J, Yang H, Wang J, Wu Q, Yin Y, Zhou H (2014) The genome of a Mongolian individual reveals the genetic imprints of Mongolians on modern human populations. Genome Biol Evol 6(12):3122–3136

Balanovsky O, Zhabagin M, Agdzhoyan A, Chukhryaeva M, Zaporozhchenko V, Utevska O, Highnam G, Sabitov Z, Greenspan E, Dibirova K, Skhalyakho R, Kuznetsova M, Koshel S, Yusupov Y, Nymadawa P, Zhumadilov Z, Pocheshkhova E, Haber M, Zalloua P, Yepiskoposyan L, Dybo A, Tyler-Smith C, Balanovska E (2015) Deep phylogenetic analysis of haplogroup G1 provides estimates of SNP and STR mutation rates on the human Y-chromosome and reveals migrations of Iranic speakers. PLoS ONE 10(4):e0122968

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Bellwood P (2011) The checkered prehistory of rice movement southwards as a domesticated cereal-from the Yangzi to the equator. Rice 4:93–103

Besaggio D, Fuselli S, Srikummool M, Kampuansai J, Castrì L, Tyler-Smith C, Seielstad M, Kangwanpong D, Bertorelle G (2007) Genetic variation in Northern Thailand Hill Tribes: origins and relationships with social structure and linguistic differences. BMC Evol Biol 7(2):s12

Bíró AZ, Zalán A, Völgyi A, Pamjav H (2009) A Y-chromosomal comparison of the Madjars (Kazakhstan) and the Magyars (Hungary). Am J Phys Anthropol 139(3):305–310

Bíró A, Fehér T, Bárány G, Pamjav H (2015) Testing Central and Inner Asian admixture among contemporary Hungarians. Forensic Sci Int Genet 15:121–126

Brunelli A, Kampuansai J, Seielstad M, Lomthaisong K, Kangwanpong D, Ghirotto S, Kutanan W (2017) Y chromosomal evidence on the origin of northern Thai people. PLoS ONE 12(7):e0181935

Chaoxiong Z (2005) Study of the origin in Zhuang civilization. Guangxi People's Publishing House, Naning

Excoffier L, Lischer L (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Eco Res 10:564–567

Fuller DQ (2011) Pathways to Asian civilizations: tracing the origins and spread of rice and rice cultures. Rice 4:78–92

Hammer MF, Karafet TM, Park H, Omoto K, Harihara S, Stoneking M, Horai S (2006) Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet 51:47–58

Ilumäe AM, Reidla M, Chukhryaeva M, Järve M, Post H, Karmin M, Saag L, Agdzhoyan A, Kushniarevich A, Litvinov S, Ekomasova N, Tambets K, Metspalu E, Khusainova R, Yunusbayev B, Khusnutdinova EK, Osipova LP, Fedorova S, Utevska O, Koshel S, Balanovska E, Behar DM, Balanovsky O, Kivisild T, Underhill PA, Villems R, Rootsi S (2016) Human Y chromosome haplogroup N: a non-trivial time-resolved phylogeography that cuts across language families. Am J Hum Genet 99:163–173

Kampuansai J, Kutanan W, Phuphanitcharoenkul S, Kangwanpong D (2012) A suggested Khmuic origin of the hunter-gatherer Mlabri in northern Thailand: evidence from maternal DNA lineages. Thai J Genet 5(2):203–215

Kampuansai J, Kutanan W, Tassi F, Kaewgahya M, Ghirotto S, Kangwanpong D (2016) Effect of migration patterns on maternal genetic structure: a case of Tai-Kadai migration from China to Thailand. J Hum Genet 62:223–228

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet 69:615–628

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genom Res 18:830–838

Kim S, Kim K, Shin D, Jin H, Kwak K, Han M, Song J, Kim W, Kim W (2011) High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea. Investig Genet 2:10

Kutanan W, Kangwanpong D (2010) Genetic structure of the ethnic populations Lua and Htin in northern Thailand. Thai J Genet 3(2):160–171

Kutanan W, Kampuansai J, Fuselli S, Nakbunlung S, Seielstad M, Bertorelle G, Kangwanpong D (2011a) Genetic structure of the Mon-Khmer speaking groups and their affinity to the neighbouring Tai populations in Northern Thailand. BMC Genet 12:56

Kutanan W, Kampuansai J, Colonna V, Nakbunlung S, Lertvicha P, Seielstad M, Bertorelle G, Kangwanpong D (2011b) Genetic affinity and admixture of northern Thai people along their migration route in northern Thailand: evidence from autosomal STR loci. JHG 56(2):130–137

Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, Pryce TO, Willis A, Matsumura H, Buckley H, Domett K, Nguyen GH, Trinh HH, Kyaw AA, Win TT, Pradier B, Broomandkhoshbacht N, Candilio F, Changmai P, Fernandes D, Ferry M, Gamarra B, Harney E, Kampuansai J, Kutanan W, Michel M, Novak M, Oppenheimer J, Sirak K, Stewardson K, Zhang Z, Flegontov P, Pinhasi R, Reich D (2018) Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science 361(6397):92–95

Lithanatudom P, Wipasa J, Inti P, Chawansuntati K, Svasti S, Fucharoen S, Kangwanpong D, Kampuansai J (2016) Hemoglobin E prevalence among ethnic groups residing in malaria-endemic areas of northern Thailand and its Lack of association with *Plasmodium falciparum* invasion *in vitro*. PLoS ONE 11(1):e0148079

Malasam S (2001) On the migration and settlement of the Yong people in Lamphun, Thailand from 1805-1902. Thammasat University Press, Bangkok

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Ongsakul S (2005) History of Lanna. Silkworm Books, Chiang Mai

Pamjav H, Fóthi Á, Fehér T, Fóthi E (2017) A study of the Bodrogköz population in north-eastern Hungary by Y chromosomal haplotypes and haplogroups. Mol Genet Genomics 292(4):883–894

Penth H, Forbes A (2004) A brief history of Lanna and the peoples of Chiang Mai. O.S. Printing House, Bangkok

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, Chen Y, Banerjee R, Rodriguez-Flores JL, Cerezo M, Shao H, Gymrek M, Malhotra A, Louzada S, Desalle R, Ritchie GR, Cerveira E, Fitzgerald TW, Garrison E, Marcketta A, Mittelman D, Romanovitch M, Zhang C, Zheng-Bradley X, Abecasis GR, McCarroll SA, Flicek P, Underhill PA, Coin L, Zerbino DR, Yang F, Lee C, Clarke L, Auton A, Erlich Y, Handsaker RE, 1000 Genomes Project Consortium, Bustamante CD, Tyler-Smith C (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet 48(6):593–599

Schliesinger J (2000) Ethnic groups of Thailand: Non-Tai-Speaking peoples. White Lotus Press, Bangkok

Schliesinger J (2001) Tai groups of Thailand: volume 1 introduction and overview. White Lotus Press, Bangkok

Seielstad M, Bekele E, Ibrahim M, Toure A, Traore M (1999) A view of modern human origins from Y chromosome microsatellite variation. Genome Res 9:558–567

Shi H, Zhong H, Peng Y et al (2008) Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol 6:45

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonné-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26(3):358–361

Underhill PA, Myres NM, Rootsi S, Metspalu M, Zhivotovsky LA, King RJ, Lin AA, Chow CE, Semino O, Battaglia V, Kutuev I, Järve M, Chaubey G, Ayub Q, Mohyuddin A, Mehdi SQ, Sengupta S, Rogaev EI, Khusnutdinova EK, Pshenichnov A, Balanovsky O, Balanovska E, Jeran N, Augustin DH, Baldovic M, Herrera RJ, Thangaraj K, Singh V, Singh L, Majumder P, Rudan P, Primorac D, Villems R, Kivisild T (2010) Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur J Hum Genet 18(4):479–484

Willuweit S, Roewer L (2015) The new Y chromosome haplotype reference database. Forensic Sci Int Genet 15:43–48

Yan S, Wang CC, Li H, Li SL, Jin L, The Genographic Consortium (2011) An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet 19:1013–1015

Zhang X, Liao S, Qi X, Liu J, Kampuansai J, Zhang H, Yang Z, Serey B, Sovannary T, Bunnath L, Aun HS, Samnom H, Kangwanpong D, Shi H, Su B (2015) Y-chromosome diversity suggests southern origin and Paleolithic backwave migration of Austro-Asiatic speakers from eastern Asia to the Indian subcontinent. Sci Rep 5:15486

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74:50–61

Zhong H, Shi H, Qi XB, Xiao CJ, Jin L, Ma RZ, Su B (2010) Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. J Hum Genet 55:428–435