Chiang Mai J. Sci. 2017; 44(4) : 1279-1294
http://epg.science.cmu.ac.th/ejournal/
Contributed Paper

Genetic Divergence of Austroasiatic Speaking Groups in the Northeast of Thailand: A Case Study on Northern Khmer and Kuy

Panthipa Chantakot Ψ [a], Pittayawat PittayapornΨ [b], Kanokpohn Srithongdaeng [c], Suparat Srithawong [c] and Wibhu Kutanan* [c]

[a] Forensic Science Program, Faculty of Science, Khon Kaen University, Khon Kaen, T40002, Thailand.

[b] Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, 10330, Thailand.

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

* Author for correspondence; e-mail: wibhu@kku.ac.th

Ψ These authors contributed equally.

Received: 5 October 2015
Accepted: 21 April 2016

ABSTRACT

The Austroasiatic speaking people are regarded as the native inhabitants in Thailand, predating the coming of Kra-Dai-speaking groups from Southwestern China. Two of the largest Austroasiatic speaking groups in Northeastern Thailand are the Northern Khmer and the Kuy. Previous genetic surveys of these two ethnic groups mostly used mitochondrial DNA polymorphisms, therefore, the present study intended to utilize the variability of autosomal microsatellites to deeply investigate the genetic structure of the Northern Khmer and the Kuy. Thirty-one and forty-seven samples of unrelated Northern Khmer and Kuy, respectively, were genotyped for 15 microsatellites. Distance based and model based clustering methods were employed to elucidate the evolutionary relationship between the studied populations and ten other linguistically and geographically diverse comparable populations in Thailand. Analysis of Molecular Variance and Factorial Correspondence Analysis revealed a genetic heterogeneity in Austroasiatic populations but an opposite trend was observed in the genetically homogeneous Kra-Dai populations. STRUCTURE result shows that the Northern Khmer contributes approximately 31% of their genes to the gene pool of their neighbor, Lao Isan, indicating a genetic exchange among them. The extreme genetic divergence of the Kuy from other populations seems to be much higher than the Northern Khmer. A genetic admixture of the Northern Khmer and with their neighboring Lao Isan was detected and this is consistent with archaeological evidence.

Keywords: autosomal microsatellites, northeastern Thailand, northern Khmer, Kuy, genetic divergence

1. INTRODUCTION

Northeastern Thailand or Isan is geographically located on the Khorat Plateau and shares the border with Laos and Cambodia. The ethnic and linguistic diversity among the populations that inhabit this region is a result of historical migrations and the resettlement of populations. The 18 ethno-linguistic groups of the region speak languages that belong to two major language families, namely Austroasiatic and Kra-Dai (also known as Tai-Kadai). While the Austroasiatic-speaking groups are considered to be indigenous, the arrival of Kra-Dai-speaking groups into the area is due to continuous Tai migrations from Southern China that started in the later centuries of the first millennium [1-2]. The lower parts of the region, which is comprised of the provinces of Surin, Sri Saket, and Buriram is of special interest to population genetics. In particular, interesting questions regard whether genetic distinction or genetic admixture might have occurred, as the three major ethnic populations have co-existed for centuries.

The Lao Isan are the largest of all ethnolinguistic groups of northeastern Thailand, numbering around 14 million people [3]. They first came from the territory of present-day Lao People's Democratic Republic (Lao PDR) to occupy northeastern Thailand the mid 14th century A.D. Historical records show that they had reached the lower Isan region by the late18th century A.D. [4-5]. Prior to the arrival of the Lao, the Isan region was primarily inhabited by people who were part of the ancient Khmer civilization. Numerous archeological sites clearly attest to the presence of Khmer culture since around the 6th century A.D. [6-7]. Separated from Cambodia to the South by natural border called the Dongrak escarpment (Figure 1), the lower part of the Isan region is home to 1.4 million Khmer ethnics [3]. These "Northern Khmer" populations speak a dialect quite divergent from the "Southern Khmer" in Cambodia [8-10]. Another Austroasiatic-speaking group in lower part of Isan is the Kuy, often called by the Thai as "Suay" [11]. About 400,000 Kuy speakers now reside in the provinces of Surin and Srisaket [3]. Their first exodus from southern Laos took place in the later part of the 17th century A.D. This was followed by sporadic migrations until the later 18th century A.D. when a mass re-settlement occurred [5, 12]. The present-day Kuy in Thailand are trilingual, speaking both Lao and Northern Khmer in addition to their native language [9, 13].

Our previous study on maternal genetic variation among ten northeastern Thai ethnicities, including the Northern Khmer and Kuy, indicated that geography was an influential factor [14]. The two groups exhibited maternally genetic differentiation from other populations whereas a certain degree of maternally genetic resemblance was detected. However, the previous study only employed mitochondrial DNA (mtDNA) as the genetic marker, thus providing a picture of their genetic relationship on the maternal side only. In the current study, we therefore expanded our investigation to include bi-parentally inherited genetic markers. This set of markers is composed of 15 autosomal microsatellites or short tandem repeats (STRs) which have been proven to be powerful genetic markers for inferring the genetic relationship of populations within a regional scale [15-18]. In particular, we examined the statistical distribution of allele frequency at 15 STRs of the Northern Khmer and Kuy from the Northeast of Thailand. In addition, linguistically and geographically diverse populations in Thailand from earlier studies were included in multiple genetic analyses to reconstruct their genetic structure and genetic relationships, as well as to consider variance components contributed by linguistic and geographic differences.

2. MATERIAL AND METHODS

2.1 Samples

Each genotyped individual was

interviewed to ensure that they were unrelated by at least three generations and that all four grandparents had been born within the village. Informed consent was obtained after an interview and a buccal swab was collected using a brush embedded in a Gentra Puregene Buccal Cell Kit (Qiagen, Hilden, Germany). The use of human subjects for this study was approved by Ethics Committee for Human Research of Khon Kaen University, Thailand. Both the Northern Khmer and Kuy (*n*=31 and *n*=47, respectively) samples reside in Sangkla District and Samrongtap District, respectively, in the area of Surin Province, Thailand. In order to execute detailed comparative statistics, the reference data from the surrounding populations were selected based on historical evidence which stated that the native Austroasiatic people inhabited the Thai territory before the arrival of the Kra-Dai speaking populations. In addition, because the statistical analyses in this study included both distance based and model based clustering methods that need the raw genotypic data, the selected representatives were from previous studies in which the raw genotypic data of fifteen STRs were available. Therefore, in total ten neighboring populations from our previous studies [19-20], i.e., Lawa1, Lawa2, Mon, Lue1, Lue2, Yuan, Yong, Khuen, Shan, and Lao Isan were utilized in the population comparison (Figure 1 and Table 1).

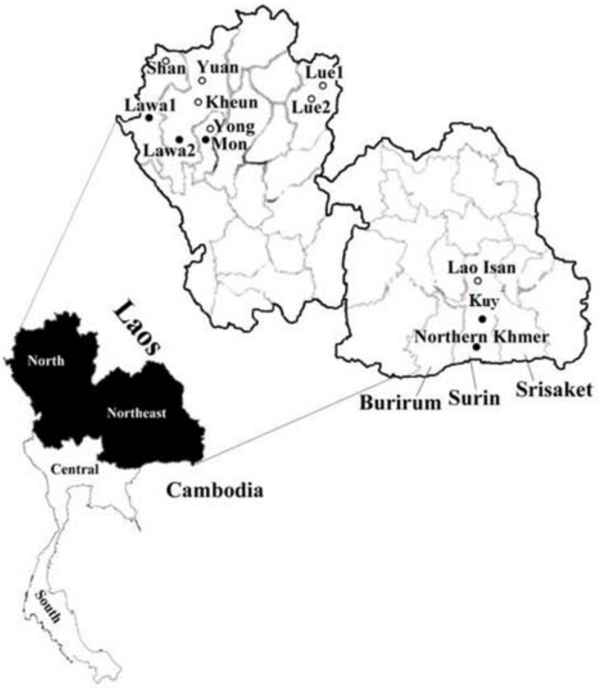

Figure 1. Map of Thailand showing the studied populations and comparable populations. Filled circles and empty circles represent populations speaking Austroasiatic and Kra-Dai language, respectively.

Table 1. General information and genetic diversities of the studied and comparable populations.

Population	Sample size (*n*)	Total Alleles	H_o	Gene diversity +/- S.D.	Location (District, Province)	Linguistic classification	References
Northern Khmer	31	116	0.7849	0.7730 +/- 0.3966	Sangkha, Surin	Austroasiatic	Present study
Kuy	47	112	0.7611	0.7539 +/- 0.3820	Samrongtap, Surin	Austroasiatic	Present study
Lao Isan	45	123	0.7686	0.7600 +/- 0.3850	Kaset Wisai, Roi Et	Kra-Dai	[20]
Lawa1	50	114	0.78	0.7669 +/- 0.3879	Hod, Chiang Mai	Austroasiatic	[19]
Lawa2	47	103	0.7643	0.7510 +/- 0.3806	Mae La Noi, Mae Hong Son	Austroasiatic	[19]
Mon	36	113	0.7703	0.7900 +/- 0.4004	Pa Sang, Lamphun	Austroasiatic	[19]
Yuan	87	126	0.7662	0.7806 +/- 0.3929	Mae Taeng and SanSai, Chiang Mai	Kra-Dai	[19]
Lue1	51	112	0.7817	0.7652 +/- 0.3871	Pua, Nan	Kra-Dai	[19]
Lue2	41	104	0.7756	0.7618 +/- 0.3863	Tha Wang Pha, Nan	Kra-Dai	[19]
Yong	55	125	0.7684	0.7757 +/- 0.3918	Pa Sang, Lamphun	Kra-Dai	[19]
Kheun	48	114	0.7402	0.7585 +/- 0.3841	Mae Wang and San Pa Tong, Chiang Mai	Kra-Dai	[19]
Shan	44	117	0.7484	0.7829 +/- 0.3961	Pang Ma Pa, Mae Hong Son	Kra-Dai	[19]

2.2 DNA Extraction and STR Typing

According to the manufacturer's specifications, genomic DNAs were isolated from the buccal samples using the same kit as buccal swab collection. Multiplex PCR reaction was performed using a commercial AmpF!STR Identifiler kit (Applied Biosystem, Foster City, CA, USA) according to manufacturer's protocol but using a total reaction volume per sample of 12.5 µl. Fifteen autosomal STR loci: *D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, vWA, TPOX, D18S51, D5S818, FGA, D19S433* and *D2S1338* were amplified. Amplicons were genotyped by multicapillary electrophoresis in an ABI 3100 DNA sequencer (Applied Biosystem). The results were further analyzed by Gene Mapper software v.3.2.1 (Applied Biosystem).

2.3 Statistical Analyses

Several statistics on genetic variation within population, that is, STR allele frequencies estimated by gene counting following a test of Hardy-Weinberg equilibrium, observed heterozygosity (H_O) and a number of alleles, and gene diversity (GD) were calculated by ARLEQUIN 3.5 software [21]. Significance level for Hardy-Weinberg *P*-values was adjusted according to the sequential Bonferroni correction (α = 0.05/15 or 0.0033) [22]. Since this set of markers is commonly used in forensic genetics, a forensic parameters that is a power of discrimination, matching probability, polymorphic information content, power of exclusion and typical paternity index were determined by Powerstats program (www.promega.com/geneticidtools/powerstats).

To investigate the amount of genetic variation due to differences at three hierarchical subdivisions i.e. within individuals of a population, among populations within a group, and among groups of populations according to linguistic classification (Austroasiatic and Kra-Dai groups) and geographic region (Northern and Northeastern Thailand), analysis of molecular variance (AMOVA) [23] as implemented in ARLEQUIN 3.5 was employed. Spatial analysis of molecular variance (SAMOVA) algorithm was applied to reveal in more detail the genetic structure of the populations. SAMOVA maximizes the differentiation between geographically homogeneous groups of populations on the basis of autosomal genotyping data (SAMOVA v.1.0; [24]).

The relationship between populations was investigated by pairwise F_{st} distances based on the number of different allele as well as the statistical significance using 1,000 permutations by ARLEQUIN 3.5. To further detect population relatedness, factorial correspondence analysis (FCA) implemented in GENETIX v.4.05.2 was employed [25]. Further investigation of population substructure was performed using the Bayesian clustering method implemented in STRUCTURE 2.3.2 [26-28] using the admixture with correlation between allele frequencies across cluster and LOCPRIOR model [28]. The number of cluster (*K*) was set from 1 to 12, and for each *K*, five independent replicates were performed with a MCMC chain burn-in length of 50,000 iterations followed by 100,000 iterations for estimate clustering. STRUCTURE Harvester [29] was used to calculate a posterior probability ((lnPr(X|K)) [24] and a second order rate of change logarithmic probability between subsequent *K* values (delta *K*) [30], to identify the optimal *K* in the data. Outputs from STRUCTURE were graphically modified by DISTRUCT [31].

A Mantel test was utilized to examine the correlations and partial correlations between two pairs of matrices of genetic and geographic distances as well as genetic and linguistic distances [32-33]. A matrix of genetic distance (F_{st}) was calculated by ARLEQUIN while geographic distances in kilometers between the approximate locations of each population were computed as great-circle distances calculated from their latitudinal and longitudinal coordinates. Pairwise linguistic distances were defined as following [34]. All matrices used for the Mantel test are shown in Table 5.

The Mantel test was also performed to correlate between the matrices of F_{st} from autosomal STRs and mtDNA sequences. Mitochondrial DNA (mtDNA) hypervariable region I (HVS-I) sequences were retrieved from previous literatures [14, 34]. Pairwise genetic distances among populations based on pairwise differences of mtDNA sequence were calculated by ARLEQUIN (data not shown).

3. Results

3.1 Genetic Variation within Population

Among all twelve populations, the average H_o was greatest in the studied Northern Khmer population (0.7849), whereas the values of GD (0.7730), and total allele (116) were intermediate (Table 1), indicating a rather high diversity in the Northern Khmer group. The Kuy also exhibited rather high values of genetic diversity parameters, i.e., H_O (0.7611), GD (0.7539) and total allele (112). After applying Bonferroni correction, deviation from Hardy-Weinberg equilibrium was not detected. The overall combined matching probability in the Northern Khmer and Kuy was 3.822×10^{-15} and 2.236×10^{-15}, respectively while the combined power of exclusion was 0.9998596 in the Northern Khmer and 0.999996 in the Kuy. The most polymorphic locus for the Northern Khmer was *FGA*, reflected by the highest H_o (Table 2 and Table 3). As expected, the most polymorphic locus was highly discriminating, as evidenced by the relatively high PD values (Table 2 and Table 3). All forensic parameters showed that this set of loci was useful for forensic identification.

Table 2. Allele frequencies and statistical parameters of genetic and forensic interests for the Northern Khmer.

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
6						7.80							3.10	1.60	
7			8.60			26.60	38.20	1.50						9.40	
8			8.60	1.50		15.60	7.40	13.20				68.80			
9						29.70						14.10		4.70	
9.3	16.10		21.40	13.60		14.10	10.30	7.40							
10	9.70		40.00	42.40		6.30	20.60	35.30		1.60		3.10		18.80	
11	8.10		17.10	34.80			20.60	19.10		1.60		14.10	10.90	26.60	
12	16.10		2.90	6.10	1.60		1.50	23.50		8.10			6.30	18.80	
13										3.20			1.60	18.80	
13.2	16.10			1.50			1.50			45.20			25.00		
14										6.50	18.30				
14.2	16.10		1.40		25.80					6.50			21.90		
15					1.60					27.40					
15.2	14.50				35.50								26.60		
16					1.60						16.70				
16.2	3.20				24.20				9.40				1.60		
17					8.10				3.10		33.30				
18					1.60				21.90		15.00		1.60		
19									14.10		13.30		1.60		10.00
20											3.30				1.70
20.2									21.00						1.70
21									4.70						20.00
22															16.70
22.2									10.90						3.30
23		2.40							20.30						11.70
24		2.40													16.70
24.2									4.70						1.70
25															8.30
25.2															1.70
26															5.00
27		28.60													1.70
30		2.40													
30.2		23.80													
31		9.50													
31.2		4.80													
32		19.00													
32.2		7.10													
33.2	0.87	0.77	0.77	0.68	0.80		0.81	0.68	0.90	0.65			0.93		
H_o	0.83	0.84	0.76	0.69	0.72	0.77	0.77	0.78	0.87	0.74	0.72	0.55	0.80	0.93	0.93
H_e	0.74	0.69	0.70	0.17	0.07	0.81	0.06	0.65	0.22	0.00	0.80	0.49	0.73	0.83	0.88
HWE	0.06	0.08	0.10	0.20	0.14	0.11	0.15	0.10	0.07	0.17	0.56	1.00	0.109	0.33	0.59
MP	0.94	0.92	0.90	0.80	0.86	0.10	0.85	0.90	0.93	0.83	0.089	0.295	0.891	0.096	0.058
PD	0.84	0.78	0.71	0.62	0.70	0.90	0.72	0.72	0.84	0.66	0.911	0.705	0.77	0.904	0.942
PIC	0.74	0.62	0.45	0.42	0.61	0.76	0.64	0.39	0.81	0.35	0.76	0.45	0.808	0.79	0.86
PE	3.88	2.63	1.75	1.65	2.58	0.56	2.83	1.55	5.33	1.41	0.482	0.248	5.33	0.872	0.795
TPI	87.10	81.00	71.40	69.70	80.60	2.29	82.40	67.60	90.60	64.50	1.88	1.14	90.60	8.00	5.00
Het						78.10					73.30	56.30		93.80	90.00

H_o, observed heterozygosity. H_e , expected heterozygosity. HWE, Hardy-Weinberg *p* values. MP, matching probability. PD, power of discrimination. PIC, polymorphic information content. PE, power of exclusion. TPI, paternity index. Het, percent heterozygote

Table 3. Allele frequencies and statistical parameters of genetic and forensic interests for the Northern Kuy.

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
6						21.90								2.10	
7			21.90			20.80	53.10	1.00							
8			10.40			3.10	7.30	15.60				55.20		3.20	
9						46.90						10.40			
9.3	8.30		38.50	10.60		6.30	8.30	16.70						16.00	
10	13.50		24.00	43.60		1.00	6.30	20.80				10.40	3.10	30.90	
11	6.30		4.20	40.40			13.50	24.00		6.30		24.00	6.30	30.90	
12	14.60		1.00	1.10	1.00		11.50	14.60		21.90			20.80	13.80	
13										5.20					
13.2	10.40			4.30	4.20			7.30		37.50			18.80	1.10	
14										6.30	46.90				
14.2	33.30				18.80					10.40			21.90		
15										9.40	6.30				
15.2	11.50				37.50				1.10				18.80	2.10	
16										3.10	6.30				
16.2	2.10				26.00				13.80				2.10		
17					11.50				4.30		15.60				4.20
18					1.00				22.30		13.50		1.00		2.10
19									7.40		7.30				3.10
20									5.30		4.20		3.10		1.00
21									1.10				4.20		26.00
22									29.80						22.90
23									10.60						9.40
24									3.20						19.80
25									1.10						7.30
26		5.00													4.20
28		43.80													
30		18.80													
31		5.00													
31.2		1.30													
32		15.00													
32.2		6.30													
33.2	0.87	0.83	0.70	0.55	0.83		0.64	0.87	0.78	0.74			0.83	0.85	
Ho	0.82	0.77	0.74	0.65	0.75	0.64	0.69	0.83	0.83	0.79	0.77	0.64	0.84	0.77	0.87
He	0.58	0.66	0.28	0.40	0.08	0.69	0.05	0.35	0.95	0.86	0.74	0.63	0.36	0.94	0.83
HWE	0.07	0.12	0.12	0.18	0.14	0.66	0.16	0.08	0.05	0.08	0.51	0.82	0.07	0.11	0.90
MP	0.93	0.89	0.88	0.82	0.86	0.14	0.84	0.92	0.95	0.92	0.11	0.19	0.93	0.89	0.07
PD	0.79	0.71	0.69	0.56	0.70	0.86	0.64	0.80	0.80	0.75	0.89	0.81	0.81	0.73	0.93
PIC	0.74	0.60	0.44	0.22	0.66	0.64	0.32	0.74	0.58	0.51	0.70	0.57	0.62	0.66	0.80
PE	4.00	2.50	1.71	1.07	3.00	0.35	1.33	4.00	2.35	2.00	0.51	0.32	2.67	2.94	0.74
TPI	87.50	80.00	70.80	53.20	83.30	1.41	62.50	87.5060	78.70	75.00	2.00	1.33	81.30	83.00	4.00
Het						64.60					75.00	62.50			87.50

H_o, observed heterozygosity. H_e, expected heterozygosity. HWE, Hardy-Weinberg p values. MP, matching probability. PD, power of discrimination. PIC, polymorphic information content. PE, power of exclusion. TPI, paternity index. Het, percent heterozygote

3.2 Genetic Variation Among Population

Geography and language have often been reported as influential factors which can limit gene flow and maintain genetic distinction among populations in both local and regional geographic frameworks [35-37]. In Thailand, several studies indicate geography [14] but some report language [34] depending on the set of studied populations. One method to quantify the effect of language and geography on the genetic relationship among populations is AMOVA. This analysis assesses the proportion of genetic variation within and between geographically and linguistically groups (Table 4). Non-significant F_{ct} statistics ($P > 0.01$) were observed, indicating that the genetic structure was neither influenced by geographic separation nor linguistic diversification. AMOVA also specified that higher genetic divergence between populations in the Austroasiatic group than the Kra-Dai group was observed, indicating that the Austroasiatic group was more genetically structured than the more homogeneous Tai group. In addition, the Northern Thai populations exhibited higher genetic variation than the Northeastern Thai populations which reflected genetic heterogeneity.

However, it should be noted that the AMOVA analysis might be not completely straightforward, because differences of genetic differentiation in both linguistic and geographic groups were detected. Another method, the Mantel test, was performed to clarify geographic and linguistic factors in shaping genetic variation patterns. The results from the Mantel test indicated that genetic *versus* linguistic distances were correlated (*r*=0.2759, *P*<0.01) and partially correlated (*r*=0.25179, *P*<0.01) whereas absent statistical significance of correlation coefficients of genetic *versus* geographic distances was detected (*r*=0.3063, *P*>0.01 for correlation; *r*=0.2852, *P*>0.01 for partial correlation).

Table 4. Analysis of molecular variance (AMOVA) results according to geographic and linguistic classification.

			% of variance (Fixation indices)		
	No. of groups	No. of populations	Within populations F_{st}	Within groups F_{sc}	Among groups F_{ct}
All samples	1	12	98.34 (**0.0166**)	1.66	
Geography					
Northeast	1	3	97.97 (**0.0203**)	2.03	
North	1	9	98.56 (**0.0144**)	1.44	
Northeast/North	2	12	98.13 (**0.0187**)	1.53 (**0.0154**)	0.33 (0.0033)
Language					
Austroasiatic	1	5	97.33 (**0.0267**)	2.67	
Kra-Dai	1	7	99.08 (**0.0092**)	0.92	
Austroasiatic/Kra-Dai	2	12	98.22 (**0.0178**)	1.52 (**0.0153**)	0.26 (0.0026)

Bold letter indicated statistically significant at $P < 0.01$

3.3 Genetic Differentiation and Population Structure

Among 66 pairwise F_{st} comparisons to test for genetic differences between the studied populations and other compared populations, fifty-seven pairwise differences (86.36%) were statistically significant after applying Bonferroni correction (*P*<0.000758) (Table 5). The Northern Khmer and Kuy exhibited significant genetic differences to all populations reflecting a high genetic divergence (Table 5). This pattern was also noticed in other Austroasiatic speaking groups from Northern Thailand (Lawa1, Lawa2, and Mon). Nine non-significant genetic differences were observed between the closely related Kra-Dai speaking populations (Table 5).

Population relationships were visualized by FCA as shown in Figure 2. The Northern Khmer and Kuy was segregated from all populations on Axis 1, which explained 15.69% of the distance matrix while the Lawa2 was separated in an upward position on Axis 2, which explained 14.27 % of the variation. Lawa1, Lue2 and Khuen were dispersed further in an upward position on Axis 2 and Axis 3 (11.45 % variation). The segregated populations, i.e., Northern Khmer, Kuy, Lawa1, Lawa2, Lue2 and Khuen mirrored genetic differentiation of these populations from the other six populations (Mon, Lue1, Yuan, Shan, Yong and Lao) which occupied an intermediate position in the plot.

Table 5. Genetic distance (F_{st}) between population based on number of different allele (below the diagonal) and geographic distance matrix as well as linguistic distance matrix as expressed in parentheses (above the diagonal).

	Northern Khmer	Lao Isan	Kuy	Lawa1	Lawa2	Mon	Yuan	Lue1	Lue2	Yong	Khuen	Shan
Northern Khmer		122 (3)	48 (2)	785 (2)	719 (2)	681 (2)	746 (3)	613 (3)	614 (3)	697 (3)	710 (3)	809 (3)
Lao Isan	**0.01521**		82 (3)	695 (3)	626 (3)	580 (3)	642 (2)	496 (2)	499 (2)	601 (2)	613 (2)	705 (2)
Kuy	**0.02618**	**0.0195**		768 (2)	700 (2)	658 (2)	721 (3)	579 (3)	581 (3)	676 (3)	689 (3)	784 (3)
Lawa1	**0.02513**	**0.0146**	**0.0303**		69 (1)	143 (2)	138 (3)	330 (3)	319 (3)	105 (3)	101 (3)	140 (3)
Lawa2	**0.02349**	**0.0259**	**0.0325**	**0.0243**		83 (2)	110 (3)	272 (3)	262 (3)	41 (3)	46 (3)	143 (3)
Mon	**0.01892**	**0.0119**	**0.0318**	**0.0229**	**0.0261**		71 (3)	189 (3)	179 (3)	41 (3)	41 (3)	131 (3)
Yuan	**0.01365**	0.0033	**0.0176**	**0.0115**	**0.0202**	**0.0099**		207 (2)	197 (2)	84 (2)	69 (2)	62 (2)
Lue1	**0.01849**	**0.0103**	**0.0279**	**0.0189**	**0.0235**	**0.0164**	0.0032		10 (1)	230 (2)	229 (2)	256 (2)
Lue2	**0.02905**	**0.0180**	**0.0326**	**0.0173**	**0.0286**	**0.0165**	**0.0109**	**0.0114**		220 (2)	219 (1)	247 (2)
Yong	**0.01821**	0.0049	**0.0241**	**0.0168**	**0.0267**	**0.0139**	0.0028	**0.0084**	**0.0192**		16 (2)	133 (2)
Khuen	**0.02646**	**0.0172**	**0.0361**	**0.0206**	**0.0257**	**0.0148**	**0.0071**	0.0057	0.0099	**0.0100**		116 (2)
Shan	**0.01963**	0.0086	**0.0193**	**0.0150**	**0.0239**	**0.0140**	0.0051	**0.0167**	**0.0206**	0.0091	**0.0165**	

Bold letter indicated $P<0.000758$ (after applying Bonferroni correction)

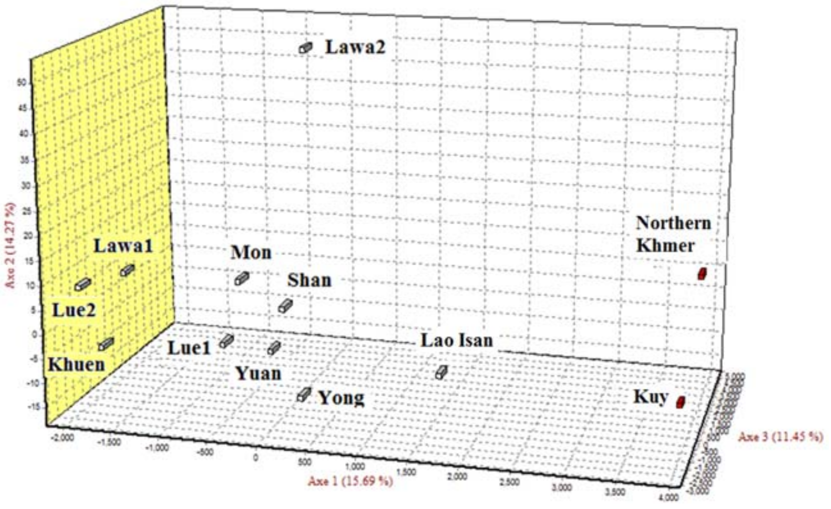

Figure 2. The Factorial Correspondence Analysis (FCA) representing population relatedness. Each population was plotted into 3-dimensional plot. Axis 1, 2 and 3 accounts for 15.69%, 14.27%, and 11.45%, respectively.

A model-based clustering method was also performed to clarify population substructures. Due to the low level of genetic divergence among the populations in which F_{st} ranged from 0.00283 to 0.03606, STRUCTURE analysis was executed using a sampling location as the *priori*, which provided more information [38]. The highest output posterior probability averaged over five replicates [26] as well as *ad hoc* statistic delta *K* which was calculated from the rate of change in the log probability of data between successive *K* values [30] were used to determine the most appropriate configuration. This was observed at *K*=8 (Supplementary 1). The population substructure of representative runs at *K*=2 to 8 are shown in Figure 3. No clear separation of distinct population was detected at *K*=2, while the red-colored component initially emerged in the Lawa2 at *K*=3. When *K* was continuously released to 4, 5, 6 and 7, the green, orange, purple, and white components were predominantly in Kuy, Lawa1, Mon, and Northern Khmer, respectively. Until *K*=8 which was the best value of number of clusters that represent the structure of the data, weakly distinguished cluster was formed in the Lue1 represented by black color. Therefore, at *K*=8, the Northern Khmer, Kuy, Lawa1, Lawa2 formed different genetic clusters from each other and from the rest. The fifth cluster mainly formed by Lao, Yuan, Yong and Shan although each population show some intermixture with different clusters, for examples, the influence of Northern Khmer cluster in the Lao and the introgression of Lue cluster in the Yuan. The Lue1, Lue2, and Khuen were assigned to the sixth cluster. The seventh cluster mainly existed in the Mon (purple) albeit the Mon was mixed by the yellow and blue clusters. The last extremely weak cluster showed in Lue1 where around 21.7% consisted in this cluster.

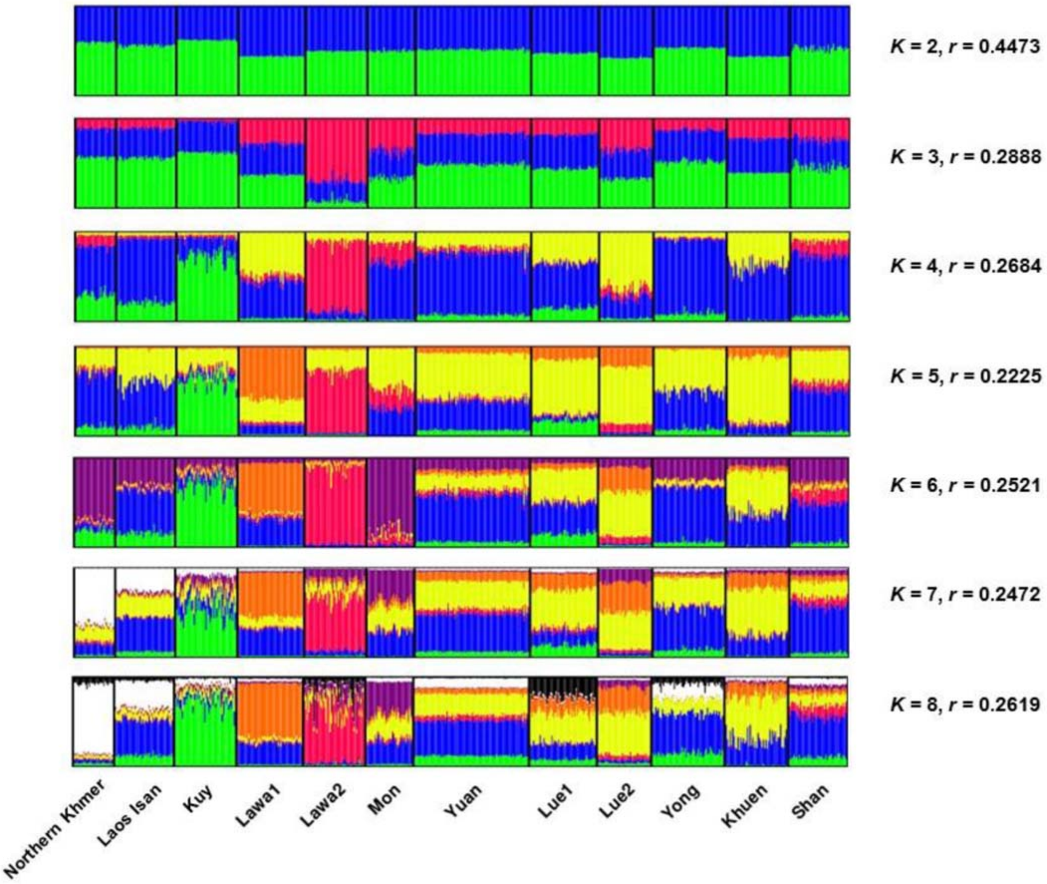

Figure 3. Clustering assignments depending on Bayesian method under an admixture with correlation between allele frequencies and LOCPRIOR models obtained by STRUCTURE. Each individual is represented by a single column that is divided into segments whose size and color correspond to the relative proportion of a particular cluster. Populations are separated by black lines. The *r* value to the right of each DISTRUCT plot is lower than 1 indicating the informativeness of sampling location in these analyses.

According to SAMOVA analysis which has been used to elucidate further detailed genetic structure and differentiation (Table 6), when the putative number of populations was increasing from 2-groups until 7-groups category, the Kuy, Lawa2, Northern Khmer, Lawa1, Lue2, Mon were partitioned from the other populations, respectively (Table 6). The former separated population in lower number of populations, the higher genetic differentiation. Although the maximization of the genetic differentiation among groups was achieved for ten population clusters (1.408%, *P*<0.01), the major increase on percent variation among group occurred for 8 groups, with value only increasing slightly thereafter SAMOVA result were broadly consistent with STRUCTURE and FCA which reflect the largely genetic divergence of the Kuy and Lawa2 than the Northern Khmer.

Table 6. Groups of populations and fixation indices as inferred by SAMOVA.

No. of group	Population member											F_{at}
2	Kuy	Northern Khmer, Lao, Lawa1, Lawa2, Mon, Yuan, Lue1, Lue2, Yong, Khuen, Shan										0.01151
3	Kuy	Lawa2	Northern Khmer, Lao, Lawa1, Mon, Yuan, Lue1, Lue2, Yong, Khuen, Shan									0.01267
4	Kuy	Lawa2	Northern Khmer	Lao, Lawa1, Mon, Yuan, Lue1, Lue2, Yong, Khuen, Shan								**0.01249**
5	Kuy	Lawa2	Northern Khmer	Lawa1	Lao, Mon, Yuan, Lue1, Lue2, Yong, Khuen, Shan							**0.01214**
6	Kuy	Lawa2	Northern Khmer	Lawa1	Lue2	Lao, Mon, Yuan, Lue1, Yong, Khuen, Shan						**0.01222**
7	Kuy	Lawa2	Northern Khmer	Lawa1	Lue2	Mon	Lao, Yuan, Lue1, Yong, Khuen, Shan					**0.01263**
8	Kuy	Lawa2	Northern Khmer	Lawa1	Lue2	Mon	Lue1, Khuen	Lao, Yuan, Yong, Shan				**0.01324**
9	Kuy	Lawa2	Northern Khmer	Lawa1	Lue2	Mon	Lue1, Khuen	Shan	Lao, Yuan, Yong			**0.01368**
10	Kuy	Lawa2	Northern Khmer	Lawa1	Lue2	Mon	Lue1	Khuen	Shan	Lao, Yuan, Yong		**0.01408**
11	Kuy	Lawa2	Northern Khmer	Lawa1	Lue2	Mon	Lue1, Yuan	Khuen	Shan	Lao	Yong	0.01371

Bold letter indicated $P < 0.01$

F_{at} = Fixation indices among groups of population

4. Discussion

The results from all our analyses agree in revealing a genetic divergence of the Kuy and the Northern Khmer populations from the other ethnolinguistic groups in northeastern Thailand. This finding is highly compatible with the linguistic classification of the Austroasiatic and Kra-Dai language family. As shown by Axis 1 of the FCA (Figure 2), the Kuy and the Northern Khmer populations in this study show a close genetic relationship to each other but are genetically quite distant from the other Austroasiatic groups. The clustering of Kuy and Northern Khmer neatly reflects the fact that the languages of the two groups are often classified together in the same major branch of the Austroasiatic family [39-40]. The genetic distance between the Kuy and the Northern Khmer on one hand, and the other populations studied on the other is in sharp contrast to the homogeneity among the Kra-Dai groups. Again, this close genetic relationship is expected given that the languages of all populations included in the current study all belong to the same subgroup within the Tai branch of the Kra-Dai family [41-43].

In addition to the agreement between genetic and linguistic groupings, this study also reveals a genetic admixture coherent with the history of our populations. When cryptic population substructure was explored by STRUCTURE, the Northern Khmer seems to share some genetic components with the Lao Isan, the majority in northeastern Thailand. The admixed components in Lao Isan, composing of green, blue, yellow, and white components belong to the Kuy, (Yuan-Yong-Shan), (Lue2-Khuen), and Northern Khmer, respectively. The result indicated that the Northern Khmer was one of the parental populations who contributed genes (approximately 31%) to the Lao Isan from Roi-Ed Province. Gene flow between these two ethnicities has also been previously reported [44]. The admixture between these two populations is consistent with the history of the region. More specifically, biculturalism comprising the Northern Khmer and the Lao Isan cultures has existed in the lower part of Northeastern Thailand for a long time [10]. Moreover, the numerous monuments of the Khmer civilization in the area suggest a presence of a Khmer population prior to the mass migration of the Lao from Laos during from the 14th to 18th century A.D. [4]. This suggests a long-lasting co-existence of the native Khmer and the immigrant Lao that resulted in a high degree of population mixing.

The genetic drift which accounted for losing maternal genetic diversity in the studied Northern Khmer from Surin Province has been previously reported [45]. Such circumstance was erased by this study with autosomal marker. A high value of all genetic variation indices was observed (Table 1). Since mtDNA markers have 1/4 the effective sample size of autosomal markers carrying the information from both parents, mtDNA has been particularly subject to the effects of random genetic drift. Therefore, we prove that the panel of STRs in the forensic study was an informative genetic marker in distinguishing Austroasiatic populations in Thailand. This agrees with previous study [46] which reported that this marker set is highly informative in separating more recently diverged Asian minority populations. However, other modes of inheritance are needed to investigate genetic structure and reconstruct population history. Again, to get more insight into Northeastern Thai populations, the paternal genetic marker is required for investigation.

One puzzle is the lack of admixture between Kuy and either Lao Isan or Northern Khmer, which indicates a lack of gene flow from or into the studied Kuy population. This finding is unexpected given the extensive language contact between the two groups reported in the literatures [9, 13]. Theoretically speaking, both languages and genetics can be transmitted from parents to their offspring. As such, concordance between genetics and language is expected for cases in which genes and language have the same evolutionary process. To understand the lack of gene flow in the face of extensive language contact, at least two explanations are possible. The first one has to do with history of language shift and maintenance in the lower part of northeastern Thailand. Although the Kuy language in Thailand is still relatively vibrant, its territory has shrunk significantly as many communities have been shifting to speak Lao or Northern Khmer [9, 11, 47]. If intermarriage accelerates language shift as it has often been reported [48, 49], a Kuy population that has maintained their original language despite a widespread shift is possibly less open to marriage to individuals from other ethnolinguistic groups. If this was indeed the case, we would not expect to see clear traces of gene flow in the Kuy population in our study. To test this hypothesis, data from other Kuy-speaking populations from neighboring provinces from Thailand and LPRD are needed.

Another possible explanation involves patterns of residence after marriage. The Kuy are one of the few remaining matrilocal residence societies in Thailand and Laos. Matrilocal residence means that after the marriage the females remain in their natal villages and the males move to their wife's village. Marital residence pattern has been demonstrated as one of the cultural factors affecting genetic variation in several populations both outside and inside Thailand [50-51]. Large genetic distances between groups exist in matrilocal groups. The genetic uniqueness of the Kuy revealed by autosome was similar to our previous studies which used mitochondrial DNA as genetic markers [14]. We also tested the correlation between the matrices of F_{st} from the autosomal STRs and mtDNA sequences which showed a strong correlation between distance matrices of these two different markers in their mode of inheritance (r = 0.5718, P<0.01). The tremendous genetic differentiation of the Kuy exhibited by our past and present studies might be affected by matrilocal residence, however, future studies on Y chromosomal variation is needed to confirm our assumption.

In conclusion, our results demonstrate extreme genetic divergence among the Kuy and moderate genetic differentiation among the Northern Khmer. Both studied populations also have genetic distinction though their geographic proximity. The Mantel test seems to support language as the predominant factor in shaping genetic variation in overall populations including the studied and comparable populations, however, AMOVA exhibited contrary results indicating that neither language nor geography was concerned in determining patterns of genetic variation. This trend might be the result of the very heterogeneous genetic structure of Austroasiatic groups. However, genetic homogeneity existed in the Kra-Dai populations, therefore, inconclusive address was proposed for whether geography or language was more important in this study. However, to our knowledge, it seems likely that genetic divergence of the Kuy might be due to their history of language shift and maintenance in the area or the matrilocal tradition practiced in Kuy society. Language adoption but without genetic introgression from the Northern Khmer to the Kuy or *vice versa* might be the possible reason. Little fraction of the Kuy component prevailed in the Lao Isan genes, reflecting minor contribution from the Kuy to Lao Isan. On the contrary, the Northern Khmer have contributed a high genetic fraction to the Lao Isan gene pool, indicating a genetic exchange between these two linguistically different populations concordant with archaeological and anthropological records.

ACKNOWLEDGEMENTS

We would like to thank all volunteers who donated their buccal cells and village chiefs for their participation. PP was funded by the Ratchadaphiseksomphot Endowment Fund of Chulalongkorn University (RES560530179-HS). WK was supported by Khon Kaen University (Grant number 590043).

DECLARATION OF INTEREST

The authors report no declarations of interest.

REFERENCES

[1] O'Connor R.A., *J. Asian Stud.*, 1995; **54**: 968-996.

[2] Pittayaporn P., *J. Humanities*, 2014; **20**: 47-68.

[3] Premsrirat S., Deepadung S., Buasuang A., Suwanket E., Choosri I., Srijampa S., *et al.*, *Ethnolinguistic Maps of Thailand.* Institute of Language and Culture for Rural, Development: Mahidol University, 2004.

[4] Keyes C., *Isan: Regionalism in Northeastern Thailand*, Department of Asian Studies,

Southeast Asia Program, Cornell University, New York, 1967.

[5] Wiphakphotchanakit T., *Isan History*, The Social Science Association of Thailand, Bangkok, 1999.

[6] Coedes G., *The Indianized States of Southeast Asia*, East-West Center, Honolulu, 1968.

[7] Vallibhotama S., 'Srisaket: the area of the "backward Cambodians"', *RuangBoran*, 1989; **15(4)**: 27-50.

[8] Thomas D., *J. Lang. Cult.*, 1990; **9**: 98-106.

[9] Smalley W., *Linguistic Diversity and National Unity: Language Ecology in Thailand*, University of California Press, Berkeley, 1994.

[10] Vail P., *Asian Ethnicity*, 2007; **8**: 111-130.

[11] Van Der Haak F. and Woykos B., *Mon-Khmer Studies*, 1990; **16-17**: 109-142.

[12] Sa-ard O., *Phrase to Sentence in Kuay (Surin)*, Master Thesis, Mahidol University, Thailand, 1984.

[13] Premsrirat S., *Mon-Khmer Stud.*, 1997; **27**: 129-136.

[14] Kutanan W., Ghirotto S., Bertorelle G., Srithawong S., Srithongdaeng K., Pontham N. and Kangwanpong D., *J. Hum. Genet.,* 2014b; **59**: 512-520.

[15] Crossetti S.G., Demarchi D.A., Raimann P.E., Salzano F.M., Hutz M.H. and Callegari-Jacques S.M., *Am. J. Hum. Biol.*, 2008; **20**: 704-711.

[16] Gaibar M., Esteban E., Moral P., Gomez-Gallego F., Santiago C., *et al.*, *Ann. Hum. Biol.*, 2010; **37**: 253-266.

[17] Babiker H.M.A., Schlebusch C.M., Hassan H.Y. and Jakobsson M., *Invest. Gen.*, 2011; **2**: 12.

[18] Kraaijenbrink T., Van der Gaag K.J., Zuniga S.B., Xue Y., Carvalho-Silva D.R., et al., *PLoS ONE.*, 2014; **9**: e91534.

[19] Kutanan W., Kampuansai J., Colonna V., Nakbunlung S., Lertvicha P., Seielstad M., Bertorelle G. and Kangwanpong D., *J. Hum. Genet.*, 2011b; **56**: 130-137.

[20] Srithawong S., Srikummool M., Pittayaporn P., Ghirotto S., Chantawannakul P., Sun J., Eisenberg A., Chakraborty R. and Kutanan W., *J. Hum. Genet.*, 2015, **60**: 371-380.

[21] Excoffier L. and Lischer H.E., *Mol. Ecol. Resour.,* 2010; **10**: 564-567.

[22] Rice W.R., *Evolution*, 1989; **43**: 223-225.

[23] Excoffier L., Smouse P.E. and Quattro J.M., *Genetics*, 1992; **131**: 479-491.

[24] Dupanloup I., Schneider S. and Excoffier L., *Mol. Ecol.*, 2002; **11**: 2571-2581.

[25] Belkhir K., Borsa P., Chikhi L., Goudet J. and Bonhomme F., *Genetix 4.00 WindowsTM software for sample genetics*, Laboratoire Génome, Populations, Interactions, University of Montpellier, France, 2004.

[26] Pritchard J., Stephens M. and Donnelly P., *Genetics,* 2000; **155**: 945-959.

[27] Falush D., Stephens M. and Pritchard J.K., *Genetics,* 2003; **164**: 1567-1587.

[28] Hubisz M., Falush D., Stephens M. and Pritchard J., *Mol. Ecol. Resour.*, 2009; **9**: 1322-1332.

[29] Earl D.A. and VonHoldt B.M., *Conserv. Genet. Resour.,* 2011; **4**: 359-361.

[30] Evanno G., Regnaut S. and Goudet J., *Mol. Ecol.,* 2005; **14**: 2611-2620.

[31] Rosenberg N.A., *Mol. Ecol. Notes*, 2003; **4**: 137-138.

[32] Mantel N., *Cancer Res.,* 1967; **27**: 209-220.

[33] Smouse P.E. and Long J.C., *Yearbook Phys. Anthropol.*, 1992; **35**: 187-213.

[34] Kutanan W., Kampuansai J., Fuselli S., Nakbunlung S., Seielstad M., Bertorelle

G. and Kangwanpong D., *BMC Genet.*, 2011a; **12**: 56.

[35] Zerjal T., Beckman L., Beckman G., Mikelsaar A.V., Krumina A., Kucinskas V., Hurles M.E. and Tyler-Smith C., *Mol. Biol. Evol.*, 2001; **8**: 1077-1087.

[36] Helgason A., Yngvadottir B., Hrafnkelsson B., Gulcher J. and Stefánsson K., *Nat. Genet.*, 2004; **37**: 90-95.

[37] Pardiñas A.F, Roca A., García-Vazquez E. and López B., *PLoS ONE*, 2012; **7**: e50206.

[38] Latch E., Dharmarajan G., Glaubitz J. and Rhodes O.Jr., *Conserv. Genet.*, 2006; **7**: 295-302.

[39] Diffloth G., The Contribution of Linguistic Palaeontology to the Homeland of Austro-asiatic, The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics; in Sagart L., Blench R. and Sanchez-Mazas A., eds., *Curzon: Routledge*, 2005.

[40] Sidwell P., Issues in Austroasiatic Classification. Language and Linguistics Compass, 2013; **7**: 437-457.

[41] Li F.K., *Handbook of Comparative Tai*, University of Hawai'i Press, Honolulu, 1977.

[42] Edmondson J.A. and Solnit D.B., Introduction; In Edmondson J.A and Solnit D.B., eds., Comparative Kadai: the Tai branch. Arlington, TX: The Summer Institute of Linguistics and the University of Texas at Arlington, 1997: 1-26.

[43] Pittayaporn P., *Phonology of Proto-Tai*, Ph.D, Thesis, Cornell University, New York, 2009.

[44] Kutanan W., Srithawong S., Kamlao A. and Kampuansai J., *Adv. Anthropol.*, 2014; **4**: 7-12.

[45] Boonsoda P., Srithawong S., Srikuka S. and Kutanan W., *Thai J. Genet.*, 2013; **6**: 40-48 (in Thai).

[46] Listman J.B., *Biases in Study Design Affecting the Inference of Evolutionary Events and Population Structure in Closely-related Human Populations*, New York University, 2009.

[47] Seidenfaden E., *J. Siam Society*, 1952; **39**: 144-180.

[48] Castonguay C., *Can. J. Sociol.*, 1982; **7**: 263-277.

[49] Stevens G., *Am. Sociological Rev.*, 1985; **50(1)**: 74-83.

[50] Kumar V., Langstieh B.T., Madhavi K.V., Naidu V.M., Singh H.P., Bisward S., Thangaraj K., Singh L. and Reddy M., *PLoS Genet.,* 2006; **2**: e53.

[51] Besaggio D., Fuselli S., Srikummool M., Kampuansai J., Castrí L., Tyler-Smith C., Seielstad M., Kangwanpong D. and Bertorelle G., *BMC Evol. Biol.*, 2007; **7(2)**: S12.