CHAPTER 4 RESULTS

4.1 Thai genetic structure comparing with global and Eurasian populations

4.1.1 Principal Component Analysis

The first ipPCA iteration displays the genetic variation between African and non-African populations, consistent with Xing *et al.* (2011). The second iteration of ipPCA separates South American populations from the rest of non-African populations, also consistent with Xing *et al.* (2011), leaving a relatively connected group of people in Eurasia (Figure 4.1).

The genetic variation in Eurasia corresponds with its geography (Figure 4.1). Populations from South Asia form a cline between the clusters of European populations and Southeast/East Asian populations. Kyrgyzstani and Buryat, Central Asian populations, position northwest of Southeast Asian populations on PC space. It is interesting that Nepalese individuals form a gradient which extends from the Indian Brahman cluster to that of Thai individuals. This not only indicates relatedness between some Thai and Nepalese individuals, but also suggests a higher level of relatedness of Thai people to Nepalese than to Indian. Another feature that is worth exploring is that Tongan and Samoan, populations from Polynesia, cluster with Southeast Asian populations (Thai, Khmer, Vietnamese and Iban) whereas the Chinese and Japanese individuals form a relatively separated cluster, rightmost on PC space.

Importantly, a remarkable feature which appears in PC analysis of individuals in Eurasia is the different levels of genetic variations in the populations in different Eurasian regions. European, Chinese and Japanese populations form tighter clusters while those from India, Nepal, Central Asia (Kyrgyzstani and Buryat) and mainland Southeast Asia (Thailand and Cambodia) produce loose scattered clusters. This indicates discreet substructures within these populations. It is evident how subsampling within Indian populations produces more uniformed clusters among the subpopulations. High levels of genetic diversity in Nepal and Central Asia has also been suggested (Xing *et al.*, 2010), explaining their gradient patterns on PC space. It is interesting how similar scattered PCA pattern are produced from mere three Thai populations (Ayutthaya and surround cities in Central region, Phuket and Moken) and one Khmer population. This underlines the need for additional genetic studies in mainland Southeast Asia with careful subsampling.

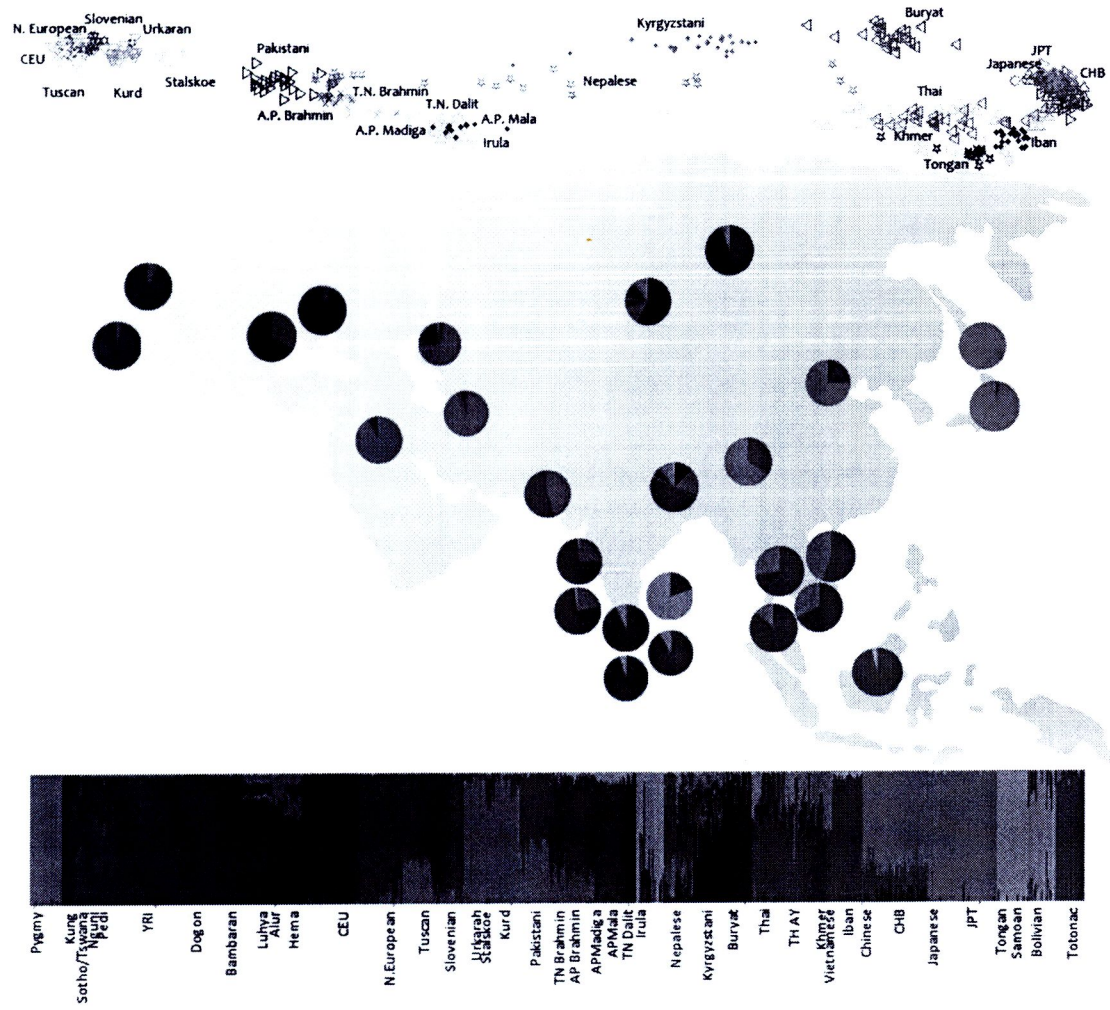


Figure 4.1 PCA and EM clustering (ADMIXTURE, K=13) of Thai (Ayutthaya and surrounding cities in Central area) and 40 populations from Xing *et al.*, 2010 which also include a Thai population from the South (Phuket and Moken). The pie charts on the map represent average admixture ratio of the population. The PCA results form a cline according to geography, placing European population to one side, Indian and Central Asian in the middle and Southeast Asian and Polynesian at the other end.

4.1.2 Genetic admixture

The program ADMIXTURE was used to assess the individual ancestry of each individual with an increasing number of ancestral components (K) (Figure 4.2).When K = 3, the three groups correspond to Africa, Europe and the rest of the world. The East Asian component is dominant in the Thai populations at a frequency of about 80%. This component decreases in a clinal pattern through Kyrgyzstani, Nepalese and Indian populations. When K = 4, South American individuals form a distinct group whose component appears in Central Asia (<20%) and East Asia (<10%) individuals while appears at very low frequency in mainland Southeast Asia (<2%) and Europe (<3%) and absent in Iban. When K = 5, the trace of South American component is still notable in Central Asian individuals but not Nepalese nor the Thai and other mainland Southeast Asian individuals. From K = 3-5, mainland Southeast Asian and the Polynesian individuals display similar admix

patterns, except that mainland Southeast Asian individuals display a little of European component (<10%) which is absent in the Polynesian but present at Nepalese and Kyrgyzstani at a relatively high frequency (20-40%).

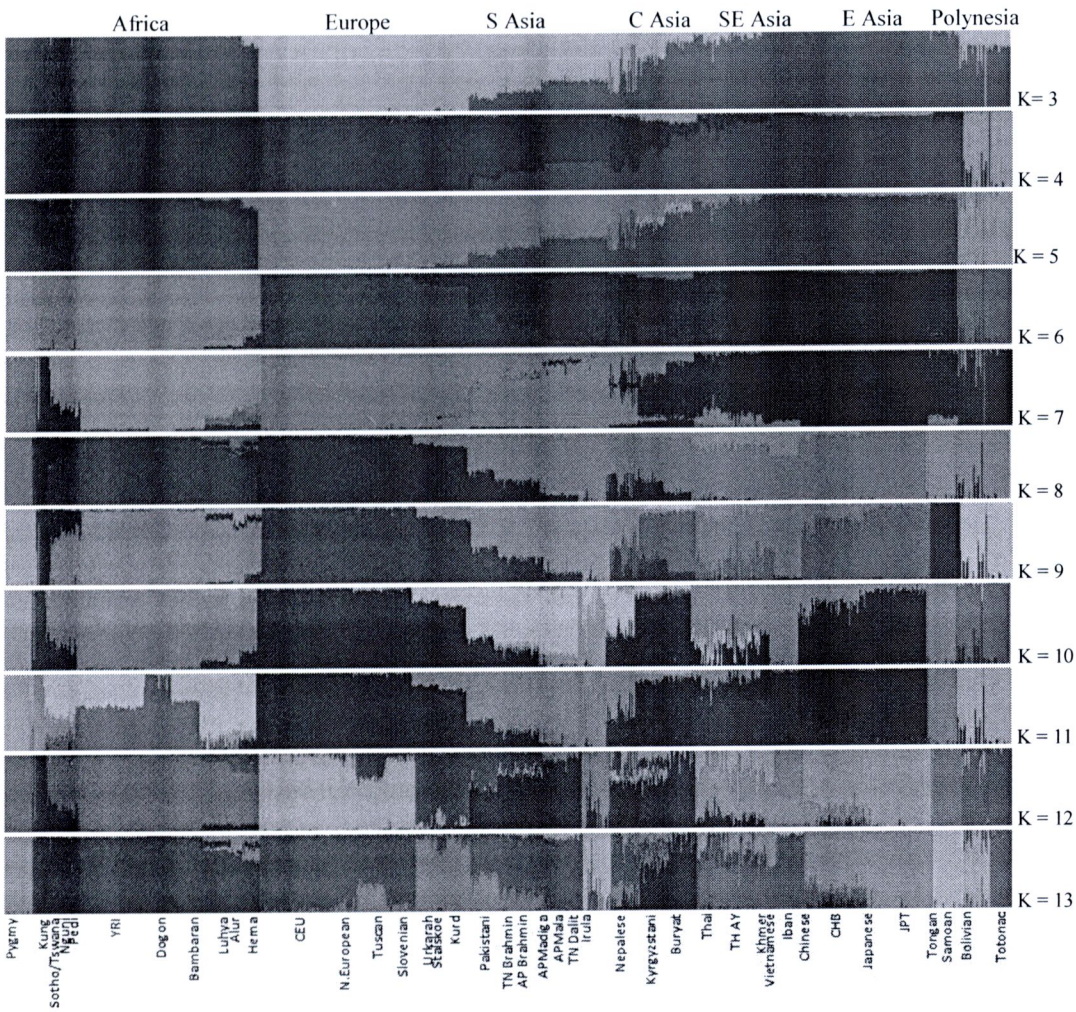


Figure 4.2 ADMIXTURE analysis of global individuals, including the new Thai samples from Ayutthaya and surrounding cities in Central area, from K = 3-13.
'Thai' means the Thai samples Thai Phuket and Moken, samples taken from *(Xing et al.*, 2010). 'TH AY' means Thai samples from Ayutthaya and surrounding cities in Central area. Other populations include those in Xing *et al.*, 2010 and HAPMAP populations.

When K =8, the Polynesian individuals form a distinct component which is shared at about 30% in Iban and 40% in Thai, Khmer and Vietnamese. This distinct Polynesian component also appears in minor proportion in Chinese individuals (<10%), suggesting that it decreases in a clinal pattern from island Southeast Asia, through mainland Southeast Asia and finally to China. At K > 8, there appears to be high variation in mainland Southeast Asian individuals. However, when K = 9, only a little of Polynesian component appears in some individuals of the Southern Thailand while absent altogether in other populations in mainland Southeast Asia. As all Iban individuals still contain a visible fraction of this component (>10%), it is possible that the Thai

individuals who have the Polynesian component at K = 9 are the Moken who live in coastal areas in Thailand.

Overall, admixture patterns in mainland Southeast Asia are consistent through most K's, having the shared components with those of Iban (~50%), Chinese (~30%), India (~20%) and Polynesia (~20%) (when they form discernible clusters), at lower K's and with those of Iban (~60%), Chinese (~30%) and India (~20%) at higher K's (Figure 4.3). It is interesting that, with more K's (K > 8), half of the Southern Thai individuals, Ayutthaya Thai and Vietnamese develop increasingly distinct patterns from one another.

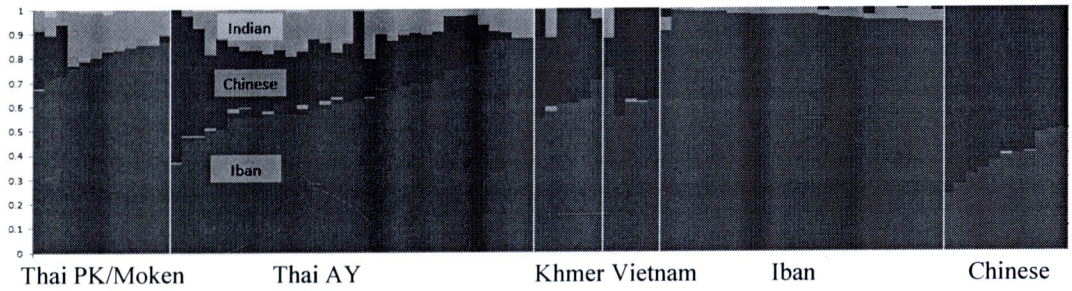
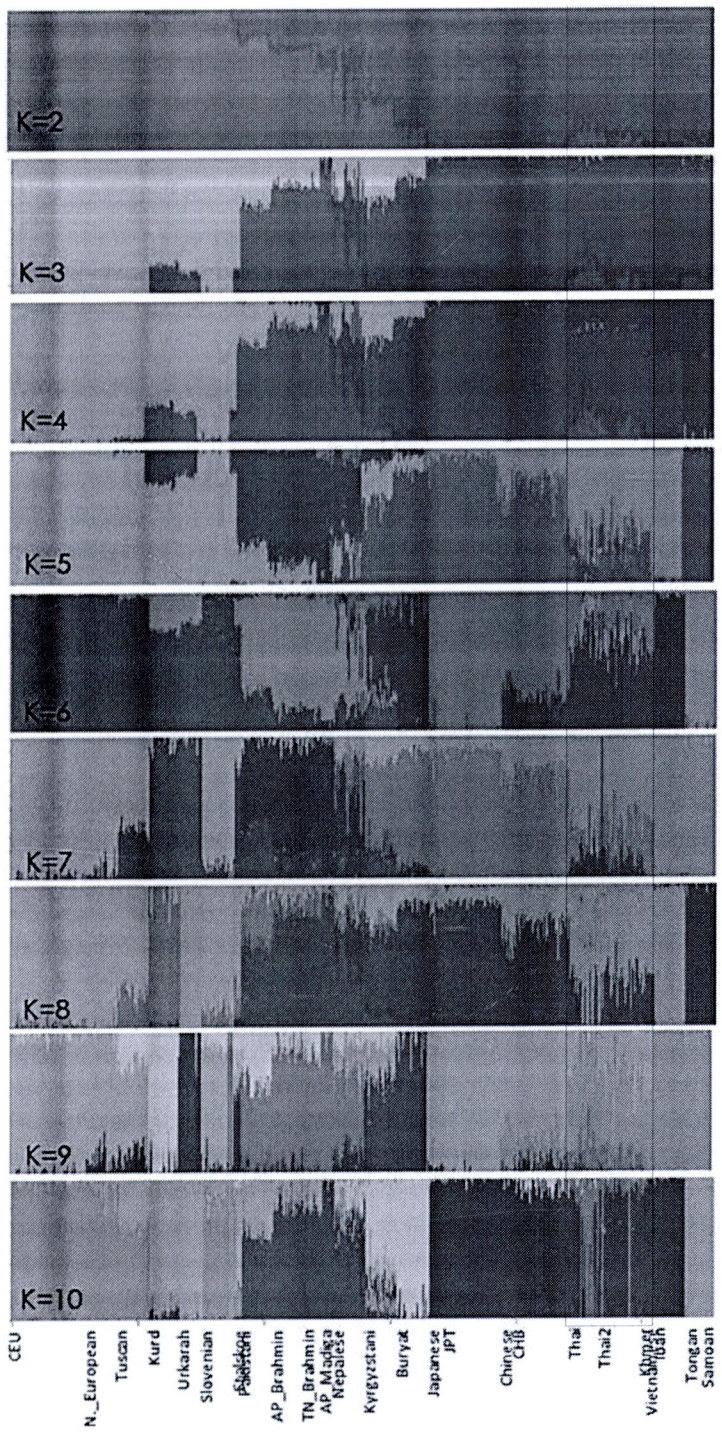


Figure 4.3 Zoom-in graph illustrating the admixture patterns of indigenous populations including Ayutthaya and surrounding cities in the Central region of Thailand.
Thai AY means the set of samples from Ayutthaya and other cities in the Central region. This is a part of a bigger set of ancestral estimation inferred from 226,412 SNPs using ADMIXTURE. The shared genetic components present in the Thai samples from Ayutthaya and surrounding cities in Central area are those of Iban (~50%), Chinese (~30%) and India (~20%).

ADMIXTURE analyses on Eurasian and Polynesian individuals were also performed exclusively (Figure 4.4) to examine their relationships with the Thai individuals in detail. Signals of shared ancestry components between Thai individuals and Iban, Chinese and Indian are consistent with the results of the larger global sets of samples that include individuals from other continents, except that all of the above mentioned patterns appear in lower K's. Furthermore, ADMIXTURE results from a pool of Eurasian populations show a distinctive cluster of populations from Southeast Asia (Thai from Ayutthaya and Phuket, Khmer and Vietnamese) early on since K=4. This indicates a close genetic distance between the populations in the cluster and how they are distinct from other populations. Interestingly, for K = 8, some Southern Thai individuals appear as an additional, separate cluster.



RESEARCH COUNCIL OF THAILAND
THE NATIONAL
RESEARCH LIBRARY

Figure 4.4 ADMIXTURE analysis (K = 2-10) of Eurasian and some Polynesian individuals from Xing *et al.* (2010) and the new indigenous Thai samples from the central part of Thailand.
The rectangle shows a distinct pattern of individuals in mainland Southeast Asia, represented by Thai individuals from Ayutthaya and surrounding cities (Thai2), Phuket and Moken (Thai) and Khmer and Vietnamese individuals.

4.1.3 Phylogenetic

Phylogenetic analysis was performed on the individual allele sharing distance (ASD) matrix of all populations. This ASD matrix captured the underlying genetic differences among all reported individuals. The software Mega4 (Smith *et al.*, 2009) was used to generate the neighbor-joining tree (Saito *et al.* 1987). The phylogenetic result is shown in Figure 4.5. It could be seen from the tree that many clusters are classified according to their geographical or ethnic groups, for example JPT, CHB and CHD are grouped together in one major branch. Particularly, the Thai individuals formed a dependent branch in the tree indicating a unique genetic pattern. The Thai genetic variants can be used to study the genetic diversity of Asian peopling and fill the variant spectrum of genetic pattern in Asia.

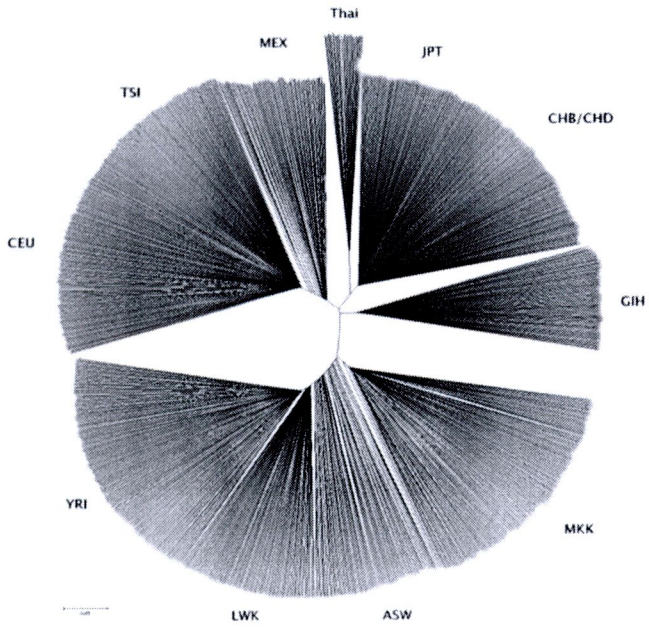


Figure 4.5 Individual phylogenetic tree constructed from allele sharing distance calculated from all individuals with neighbor joining method.

4.1.4 Heterozygosity

Expected heterozygosity was calculated from 299,837 overlapping SNPs with genotyping data from HapMap3 samples. Figure 4.6 shows the bar chart of expected heterozygosity of tested population, including Thai. The Caucasian descendants appeared to have highest values of expected heterozygosity while African and Asian descendants revealed lower values. However, the heterozygosity values do not tell us much about genetic relatedness among the populations derived from similar origins; for example, descendants from Africa do not share the similar expected heterozygosity values. It might need further investigation.

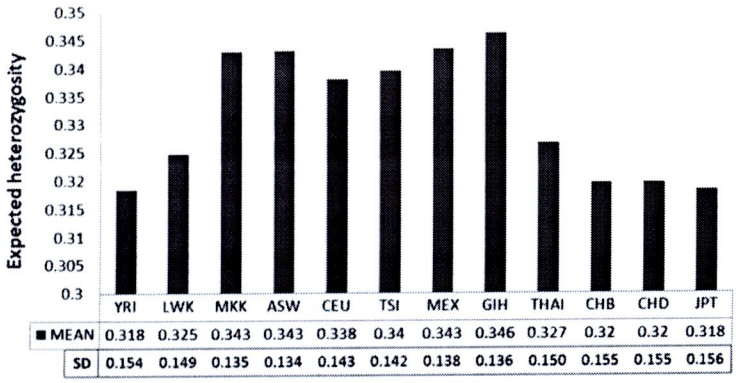


Figure 4.6 Expected heterozygosity calculated from genotype data from each population reported by ThaiSNPdb including 32 Thai data (THAI) and the data from 11 HapMap populations.

4.2 Thai genetic structure comparing with Southeast and East Asian populations

4.2.1 Principal Component Analysis

ipPCA was performed for both global and exclusive population sets. Higher iterations of ipPCA (Figure 4.7A, B) provide higher discriminative power while exclusive selection of relevant populations (Figure 4.8) provide clearer results with less noise. In global samples with ipPCA iteration 4 (Figure 4.7B), the majority of the genetic variation is found between the Polynesians (Tongan and Samoan) and the populations in Eurasia, as represented by the first principal component (PC1). PC2 reflects genetic variation in Central, East, Southeast Asia which form a cline corresponding to Central-East-Southeast Asian geography. However, Kyrgyzstani and Buryat (Central Asian) form discernible clusters from the more continuous clinal pattern produced by Southeast Asian and East Asian individuals. It is interesting that some Nepalese and individuals still position in proximity with Southeast Asian and East Asian individuals while Indian individuals are excluded entirely from the clinal cluster.

Additionally, Thai individuals form less defined clusters than do the rest of the populations. This indicates stratification within the sampled Thai individuals, with some showing a closer relationship to Chinese populations while others are closer to Iban, a hunter gatherer in Sarawak, Borneo. ipPCA results with exclusive population sample sets (Figure 4.8A) are also consistent with those from global sample sets with an evident separation within Thai individuals from Ayutthaya and the South. Thai Ayutthaya individuals are more closely related to Iban. Some Southern Thai individuals are closely related to Thai Ayutthaya while some are relatively very distant. It is also interesting that at higher ipPCA iteration, a number of Southern individuals form a separated cluster (not shown).

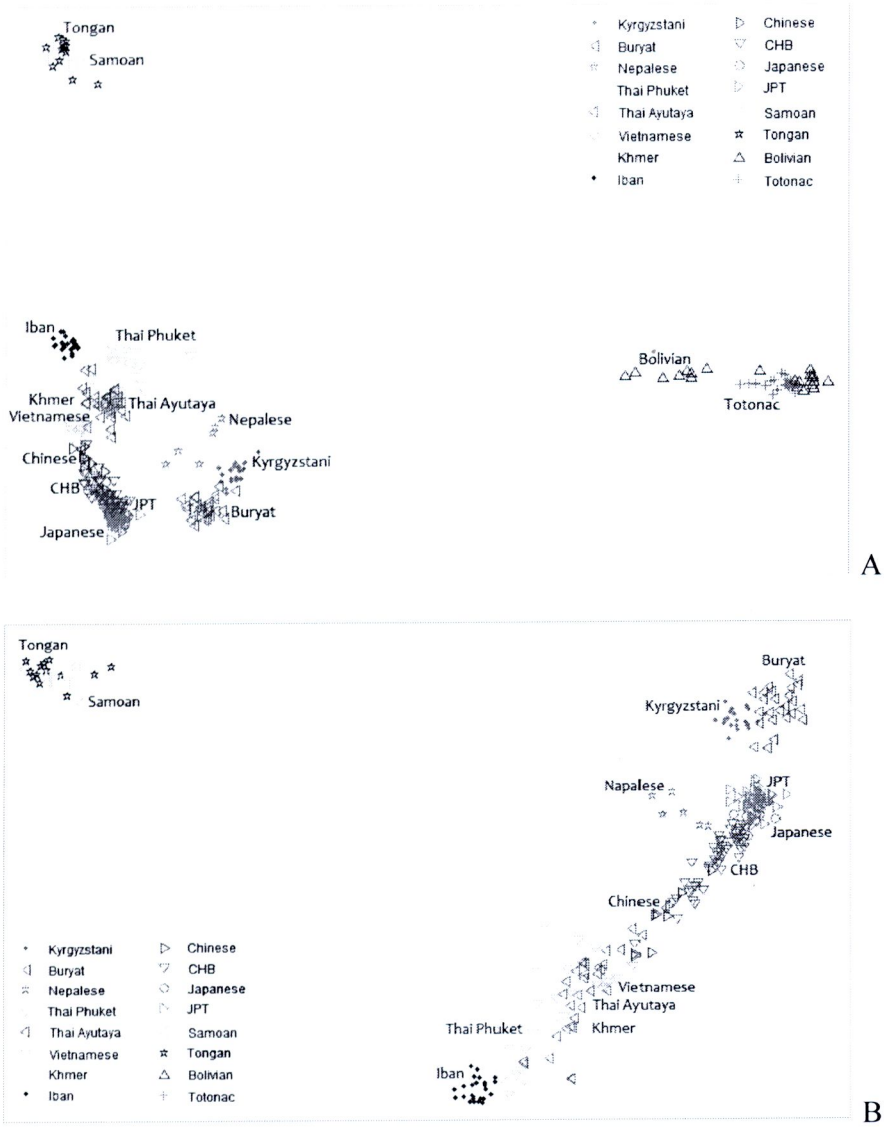


Figure 4.7 Non-parametric analysis—ipPCA iteration 3 (A) and iteration 4 (B)—of genetic variance in Southeast Asian populations.

Phylogenetic neighbor-joining tree constructed from allele sharing distance (ASD) (Figure 4.8B) also indicates that Thai Ayutthaya is closely related to Iban than the Southern Thai. Also, individuals from Thai Ayutthaya and Southern China form a mixed cluster, indicating high affinity between them. Indian individuals are separated first on the tree, followed by some Southern Thai individuals, Iban and all of Thai Ayutthaya. It is surprising that Iban and Chinese individuals are separated after Thai Ayutthaya, suggesting an older ancestry of the latter.

THE NATIONAL RESEARCH COUNCIL OF THAILAND
RESEARCH LIBRARY

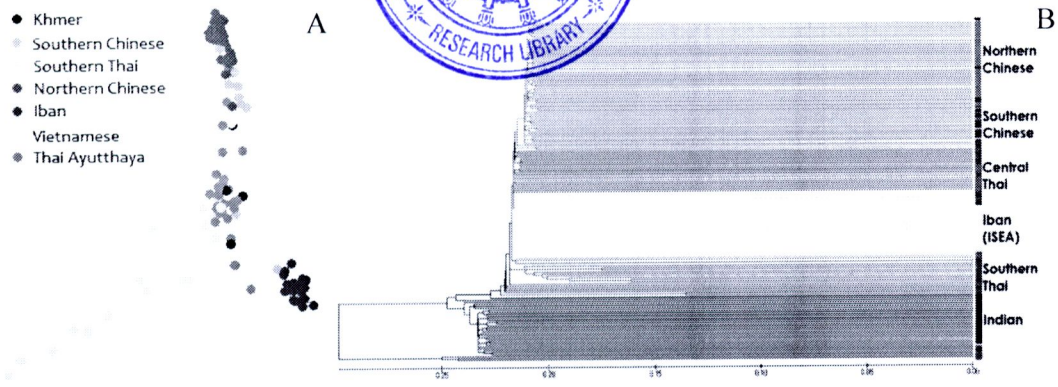
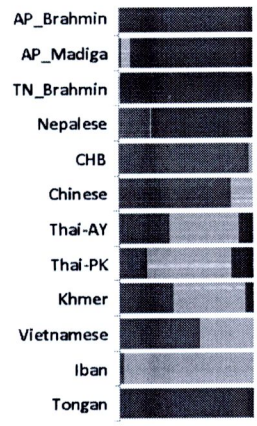


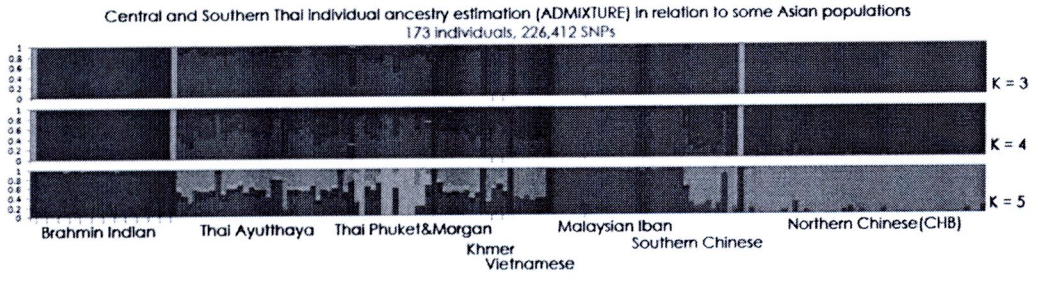
Figure 4.8 Principal component analysis of some genetic variation among populations whose global results show close genetic proximity to Thai populations (A) and phylogenetic Tree for 8 populations (2 Chinese, 2 Thai, Khmer, Vietnamese, Iban and Bhramin Indian) (B).
The light blue labels are Hapmap northern Chinese. dark blue labels are Southern Chinese. The red labels (Central Thai) are individuals from Ayutthaya and surrounding cities. Yellow is for Iban, which represents an Island Southeast Asian (ISEA) population. Green labels are Southern Thai individuals. Indian individuals are labeled in black

4.2.2 Genetic admixture

ADMIXTURE analyses are consistent with ipPCA results. When K = 2, most Indian and Chinese individuals cluster into distinct components while Nepalese, mainland Southeast Asian and Polynesian individuals have varying degrees of shared components from Indian and Chinese. When K = 3, Tongan forms a distinct component which is shared for approximately 30% in Iban, 20% in Thai and less than 5% in some Indian individuals. The Polynesian component is absent in Nepalese. Four ancestral components (K = 4) should best represent the major ancestral components in this set of populations. Tongan forms a distinct component, trace proportions of which appear in Iban (<3%) and some Indian, Chinese and Thai individuals (<1%). Iban and Chinese components are almost equally dominant in Thai individuals. Iban components are dominant in all Thai individuals while Chinese components are dominant only in all Central Thai. Importantly, for all K's, higher proportions of shared Chinese component are observed in Thai Ayutthaya individuals than do Southern Thai. Interestingly, when K = 5, some Southern Thai individuals appear as a distinct component. This Southern Thai component is shared for approximately 10-20% by the rest of the Thai individuals. It also appears in small degree in some Indian and Chinese individuals. With this set of samples, high variation within Thai populations become even more distinct, as shown in Figure 4.9A. Averaged ancestry estimates for each population (Figure 4.9B) indicate with consistency with results from global sample sets that Thai Ayutthaya, Southern Thai and Khmer populations have similar genetic patterns, composing of different degrees of shared components with Chinese, Iban and Indian populations.



A



B

Figure 4.9 Averaged (A) and Individual (B) ancestral estimation for Thai and populations which have shared ancestry in the larger set of samples.

4.3 Genetic structure within Thai population

4.3.1 Principal Component Analysis

PCAs of a smaller set of 455 recent Thai individuals using 39,049 autosomal SNP markers revealed a North to South cline of genetic variation with some Northeastern samples scattering toward the east of PC space, reflecting the geography of Thailand (Figure 4.10A). However, a larger set of 992 recent Thai individuals using 553,892 autosomal SNP markers revealed discernible patterns of genetic variation that reflects also distinction between individuals from different geographical regions in Thailand (Figure 4.10B). Different combinations of Thai populations could enhance the component signals between the selected populations (Figure 4.11B, C, D in comparison with Figure 4.11A). Southern and Northeastern Thai individuals formed trails of separated clusters concordant to their geographical directions (Figure 4.11A-D). The highly scattered PCAs of Southern and Northeastern samples indicate a high level of genetic diversity in these two populations. On the contrary, the Northern cluster was relatively more compact while the Central samples also formed a large scattered cluster.

Figure 4.10 PCA featuring a clinal pattern of genetic variation among Thai individuals in accordance to geography of Thailand (A) and PCA of a larger set of individuals and SNP markers featuring discernible patterns of genetic variation consistent with the four geographical regions of Thailand (B).

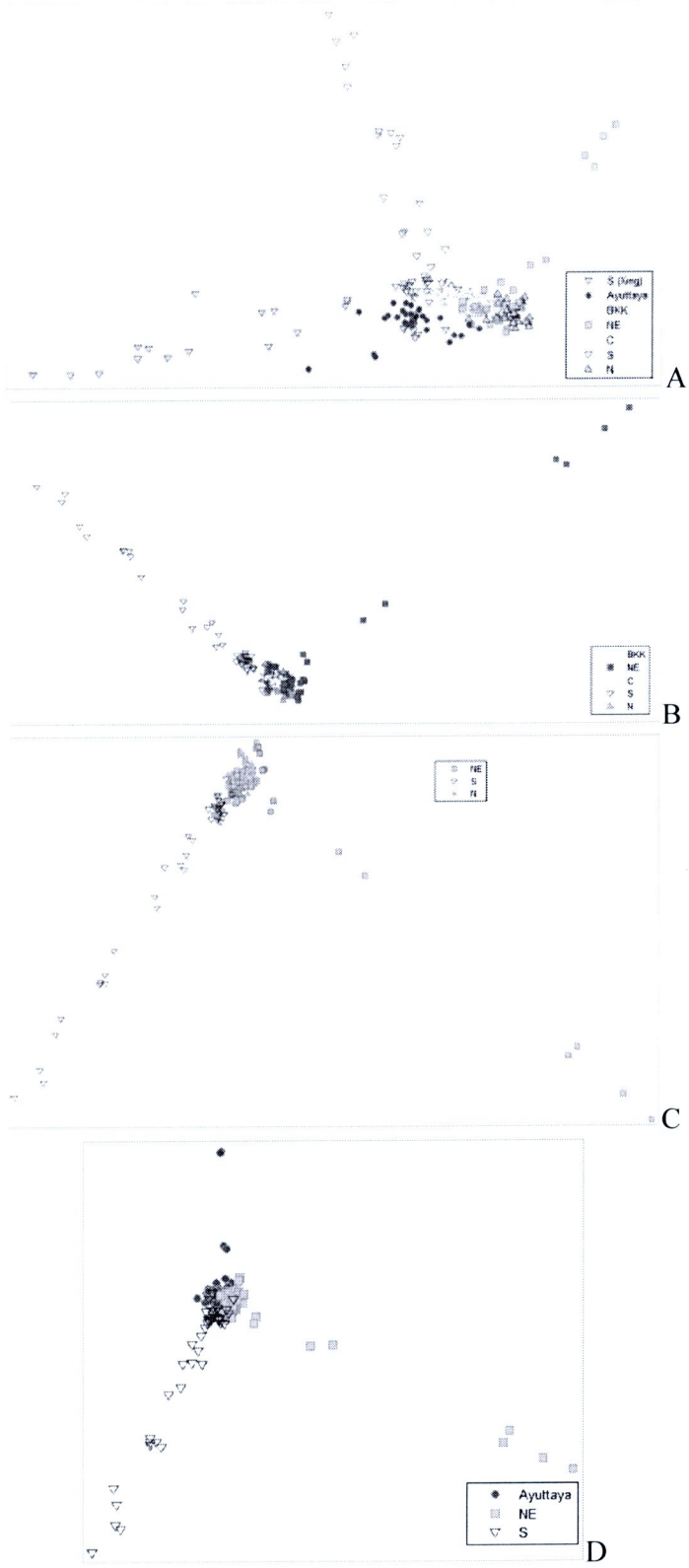


Figure 4.11 PCA with different sets of Thai populations: indigenous Thai population in central region (Ayutthaya) and southern region (S Xing) and recent populations in all regions of Thailand (N, NE, C, BKK, S).

4.3.2 Phylogenetic tree

The phylogenetic tree reconstructed from individual allele sharing distance from the Northern, Northeastern and Southern samples showed no distinct clades but placed individuals from the same geographical region in proximal branches (Figure 4.12). This confirms PCA findings of a genetic substructure among the three regions of Thailand. In a more secluded set of samples composing of individuals from Northern, Northeastern and Southern regions (Figure 4.13), the clades of Southern Thai individuals are separated first from the phylogenetic tree, indicating an older ancestry.

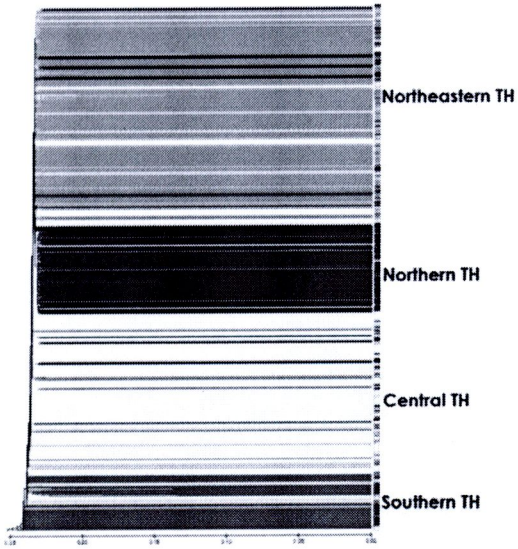


Figure 4.12 Phylogenetic tree reconstructed from individual allele sharing distance from recent Thai individuals from four regions of Thailand

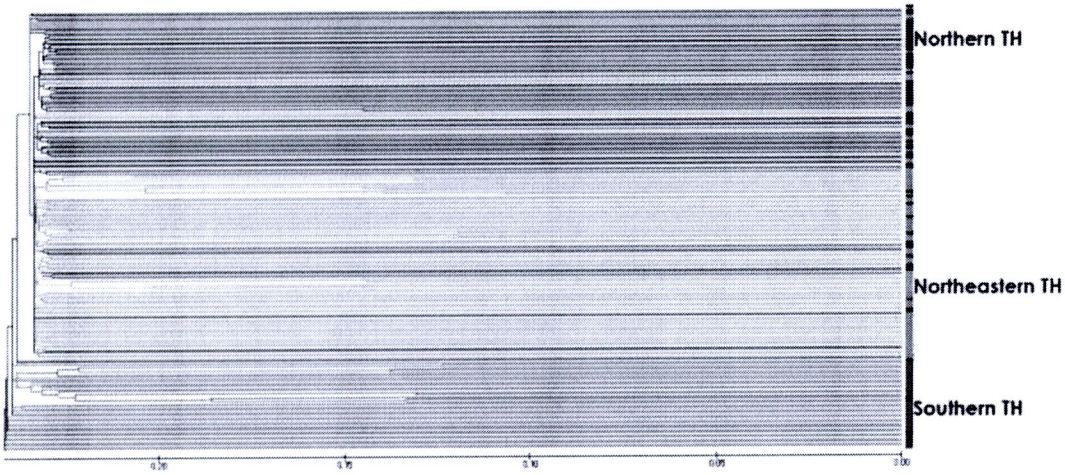


Figure 4.13 Phylogenetic tree reconstructed from individual allele sharing distance from the Northern, the Northeastern and the Southern Thai populations

When compared with other closely-related populations (Indian, Nepalese, Khmer, Vietnamese, Northern Chinese, Southern Chinese and Iban) (Figure 4.14), Thai Ayutthaya and the group of Thai Phuket and Moken are separated first after the oldest clades of Indian and Nepalese while the Northern Chinese are separated last. Northern

Thai and Mixed group 1, 2 and 3, which compose of individuals from four regions of Thailand, are relatively new compared to the group of Central and Southern Thai.

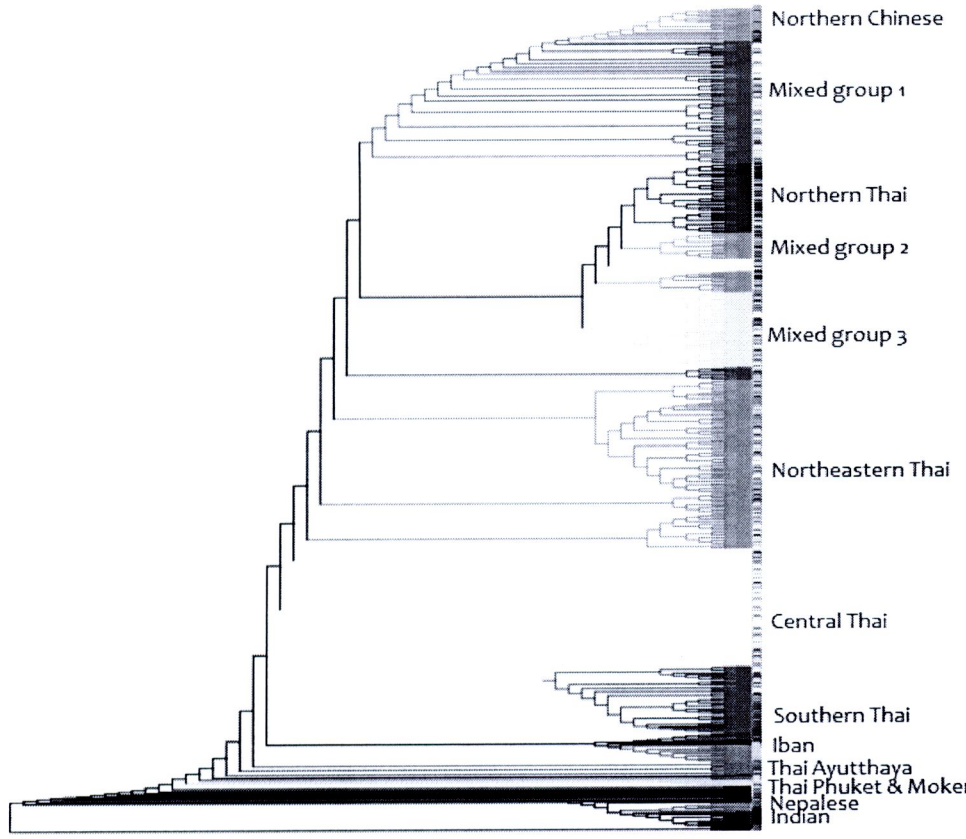
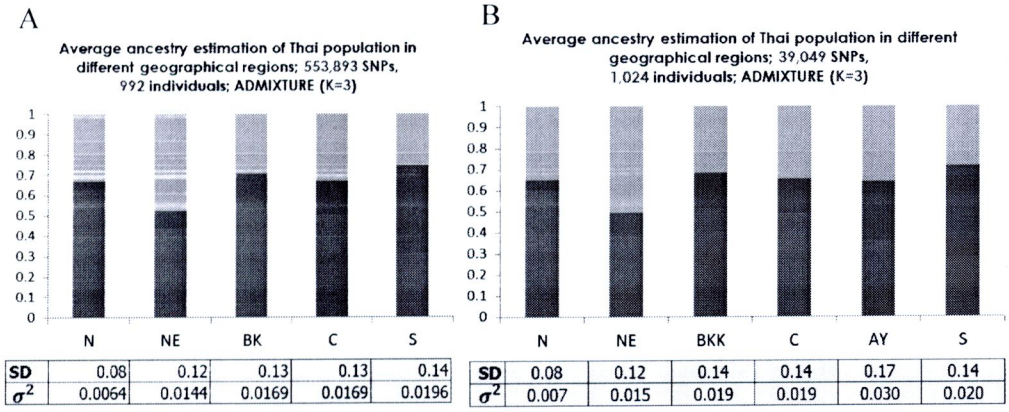


Figure 4.14 Phylogenetic tree reconstructed from individual allele sharing distance of recent Thai individuals from 4 regions of Thailand and individuals from other closely related populations in East and South Asia.
Mixed group 1, 2 and 3 contain mostly Thai individuals from the regions that are close to the mixed groups.

4.3.3 Genetic admixture

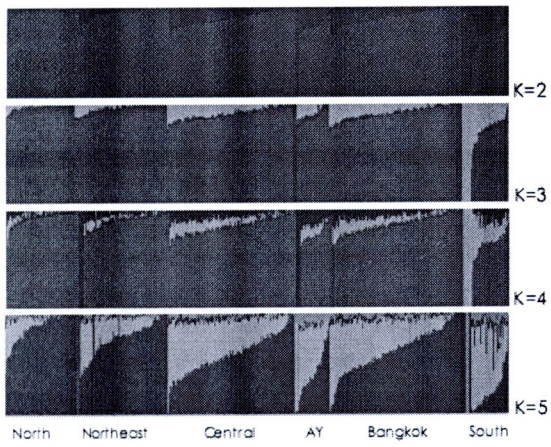
Ancestry estimation patterns of the population in each region of Thailand (K=3) strongly indicate that there are genetic substructures within the recent populations in Thailand (Figure 4.15A, B). Furthermore, individual ancestry estimations of both the small (Figure 4.16) and large (Figure 4.17) Thai datasets indicate that there is a degree of genetic variation among Thai individuals even in regional scale. To quantify the difference between Thai populations in each region, pairwise analysis of the norms of population average ancestry estimates were calculated (Table 4.1, Table 4.2). Angles between the norms of vectors representing the genetic components of Southern Thai population and other regions had low cosine values ($\cos \theta = 0.71$-0.85), indicating high degrees of difference between the Southern Thai population and the Thai in other regions. The highest cosine value (the least similar pattern of ancestry estimates) is that between Northern and Southern Thai populations ($\cos \theta = 0.71$) while Bangkok and Central Thai populations have the least cosine values as expected ($\cos \theta = 0.99$).



SD	0.08	0.12	0.13	0.13	0.14
σ^2	0.0064	0.0144	0.0169	0.0169	0.0196

SD	0.08	0.12	0.14	0.14	0.17	0.14
σ^2	0.007	0.015	0.019	0.019	0.030	0.020

Figure 4.15 Averaged ancestry estimation of Thai populations in Bangkok, the North, the Northeast, the South and the Central are of Thailand with pooled standard deviations and variance of the genetic component values among individuals in each region.
A) Result from a large dataset of 992 individuals and 553,893 SNP markers. B) Result from the previous dataset combined with Ayutthaya samples (32 individuals). Overlapping SNPs between two SNP chip platforms are only 39,049.



THE NATIONAL RESEARCH COUNCIL OF THAILAND
RESEARCH LIBRARY

Figure 4.16 Ancestry estimation of recent Thai individuals who live in the North, the Northeast, the Central area and the South of Thailand with two subsampling of the people in Bangkok and Ayutthaya and surrounding cities, ADMIXTURE K = 2-5, 455 individuals, 39, 049 SNPs.

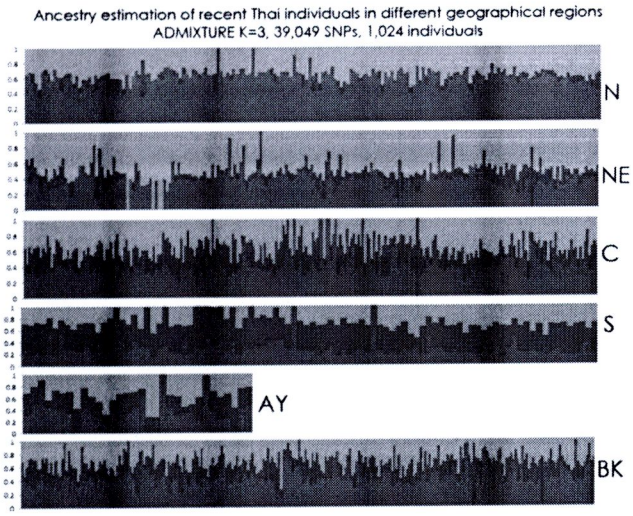


Figure 4.17 Ancestry estimation of recent Thai individuals in different geographical regions, ADMIXTURE K = 3 at 39,049 SNPs, 1,024 individuals.

Table 4.1 Pairwise analysis of the norms of ancestry estimates (vectors) in each population for the larger set of 992 Thai individuals and denser markers of 553,892 SNPs.

	N	NE	BKK	C	S
N					
NE	0.936030				
BKK	0.977911	0.940096			
C	0.972370	0.960359	0.997298		
S	0.716573	0.777026	0.845740	0.856048	

Note: Numbers represent cosine values of the angles between vectors, indicating the degrees of differences. The higher cosine values indicate higher degree of similarity.

Table 4.2 Pairwise analysis of the norms of ancestry estimates (vectors) in each population for a set of 1,024 Thai individuals (the dataset above plus 32 Thai Ayutthaya) and less dense overlapping markers of 39,049 SNPs.

	N	NE	BKK	C	AY	S
N						
NE	0.918459					
BKK	0.977861	0.932884				
C	0.972028	0.951024	0.998194			
AY	0.877786	0.939275	0.952684	0.965174		
S	0.698395	0.787728	0.831802	0.845429	0.946438	

Note: Numbers represent cosine values of the angles between vectors, indicating the degrees of differences. The higher cosine values indicate higher degree of similarity.

ADMIXTURE analyses (Figure 4.18) and subsequent pairwise analyses of the norms of genetic vectors (Table 4.3) also indicate that Thai Ayutthaya is most closely related to

Mon and some other Tai ethnic groups. On the contrary, the ancestry estimation profiles of Thai Ayutthaya are more distant to those of Northern hill tribal populations, *e.g.* Lawa and Hmong.

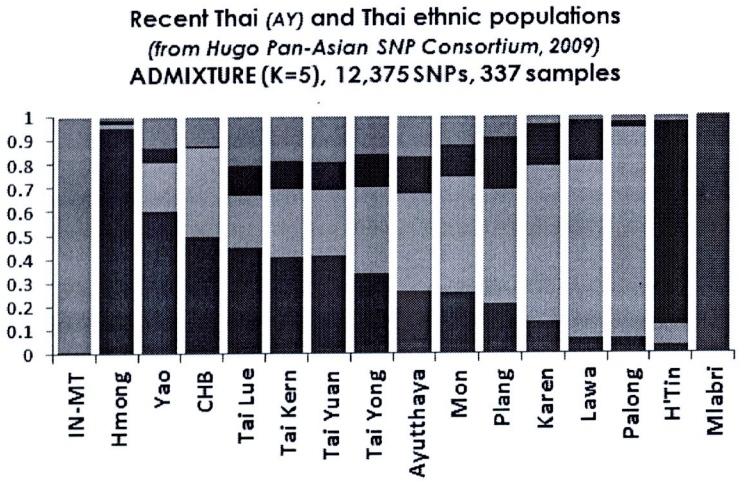


Figure 4.18 Averaged ancestry estimation Thai Ayutthaya, other ethnic populations in Thailand and one ethnic population from Indonesia (IN-MT).

Table 4.3 Pairwise analysis of the norms of ancestry estimates (vectors) in each ethnic population for a set of 337 individuals, 12,375 SNPs.

	Thai Ayutthaya
Thai Mon	0.990
Tai Yuan	0.983
Plang	0.969
Tai Kern	0.923
Tai Yong	0.913
Karen	0.907
CHB	0.870
Tai Lue	0.864
Lawa	0.860
Palong	0.824
Yao	0.748
Hmong	0.459
H'Tin	0.388
Indo-Mentawai	0.325
Mlabri	0.065

Note: Numbers represent cosine values of the angles between vectors, indicating the degrees of differences. The higher cosine values indicate higher degree of similarity. Lower cosine values indicate higher degree of genetic variation between the two populations.

4.3.4 F_{ST}

The F_{ST} between Thai populations are shown in Table 4.4. The highest F_{ST} is found between Southern population and Northern population at 0.0048. This indicates that there is a high degree of genetic difference between the two populations. The lowest F_{ST} values are those between Central and Northeastern and Central and Northern populations (0.0011 and 0.0012 respectively). There is relatively medium degree of genetic variation between Northern and Northeastern populations with pairwise F_{ST} at 0.0022.

Table 4.4 Pairwise F_{ST} of Thai populations from four regions of Thailand.

	S	C	NE	N
S	-	0.0017	0.0039	0.0048
C	-	-	0.0011	0.0012
NE	-	-	-	0.0022
N	-	-	-	-

4.4 Haplotype sharing

Haplotype inference and pairwise comparison between populations were performed in only 2 chromosomes (Table 4.5, 4.6) because the tests require expensive computational resources. Surprisingly, the higher numbers of haplotypes in chromosome 21 are those of Thai Phuket and Moken (as one mixed group) (1,381), Kyrgyzstani (1,378), Indian Brahmin (1,366) and Thai Ayutthaya (1,311), respectively. The highest numbers of haplotypes in chromosome 20 are those of Indian Brahmin (1,744), Kyrgyzstani (1,684), Thai Phuket and Moken (1,598) and Thai Ayutthaya (1,566), respectively. Even though the results for the first three populations (Indian, Kyrgyzstani and Thai Phuket and Moken) have variable highest numbers of haplotypes, they can be grouped together as older populations, as the higher numbers of haplotypes indicate longer time for recombination and thus older genetic composition. Thai Ayutthaya population is newer than these populations but is still relatively old when compared to Chinese, Japanese, Iban and Samoan populations.

Table 4.5 Number of haplotypes in chromosome 21 and haplotype sharing in populations with >20 individual samples.

Population A	Population B	Share AB	Num Hap in A
Thai Ayutthaya	Thai Phuket & Moken	711	1311
Kyrgyzstani	Thai Phuket & Moken	706	1378
JPT	Thai Phuket & Moken	699	1214
CHB	Thai Phuket & Moken	685	1241
Iban	Thai Phuket & Moken	685	1226
Indian Brahmin	Thai Phuket & Moken	649	1366
Thai Phuket & Moken	Tongan	619	1381
Samoan	Thai Phuket & Moken	596	1037

Table 4.6 Number of haplotypes in chromosome 20 and haplotype sharing in populations with >20 individual samples.

Population A	Population B	Share AB	Num Hap in A
Thai Phuket & Moken	Thai Ayutthaya	861	1598
Kyrgyzstani	Thai Ayutthaya	835	1684
JPT	Thai Ayutthaya	826	1431
CHB	Thai Ayutthaya	821	1458
Iban	Thai Ayutthaya	819	1485
Indian Brahmin	Thai Ayutthaya	779	1744
Thai Ayutthaya	Tongan	753	1566
Samoan	Thai Ayutthaya	728	1328

Chinese has higher number of haplotypes (1,241 in chromosome 21 and 1,458 in chromosome 20) than Japanese (1,214 in chromosome 21 and 1,431 in chromosome 20). Iban has lower number of haplotypes (1,226 in chromosome 21 and 1,485 in chromosome 20) than Thai Ayutthaya and Chinese but still higher than Japanese. The population with lowest number of haplotypes is the Samoan of Polynesia.

Haplotype sharing or the numbers of haplotypes that are shared between two populations are consistent in the two chromosomes. The populations in Table 4.5 and 4.6 are ranked according to their number of haplotypes shared. Thai Ayutthaya has the highest haplotype sharing with Thai Phuket and Moken, Kyrgyzstani, Japanese, Chinese and Iban, respectively. The numbers of sharing decrease with relatively bigger difference between Thai-Indian and Thai-Polynesian populations.