CHAPTER 3 METHODOLOGIES

3.1 Samples

Thai individual samples from Ayutthaya and surrounding cities in the central region of Bangkok are from Thai SNP2 project conducted by BIOTEC, Thailand. Ayutthaya situated in the central area of Thailand and was one of the biggest Thai capitals which flourished in a relatively more recent historical period in A.D. 1350-1569. Thus, the samples from Ayutthaya and surrounding cities in central region were primarily chosen to represent recent Thai people. Genotyping for this set of data was based on the 5th generation Affymetrix SNP genotyping 500K array, which comprises common SNPs reported by Affymetrix.

Recent Thai individuals from four regions of Thailand were obtained through case-control projects conducted by the Thalasemia Research Center, Mahidol University and DNA profiling project on depression disorder by Rajanukul Institute's Centre for Medical Genetic Research. The two projects provide highly dense sets of 570,856 and 566,200 sets of SNPs using Illumina SNP genotyping platforms. Individual samples used in this study include those diagnosed as having the diseases and the control (healthy) individuals. The individual genetic samples are classified into four geographical regions. Thalassemia set of data also have individual samples from Bangkok, in addition to those from Central region of Thailand.

This study includes global samples derived from *Xing et al.* (2010) to provide global comparison. This set of SNP data was genotyped with identical platform with Thai Ayutthaya data. Table 3.1 shows the original sets of genotype data obtained from various sources for this study.

Finally, genetic data from the HUGO Pan-Asian SNP Consortium (2009) were used for comparison. It provides a set of 54,794 autosomal SNPs in 1,928 individuals representing 73 Asian populations. Only a number of populations which are relatively more closely related to Thai populations are used for comparison in this study. The genotyping for Pan-Asian data is the Affymetrix Gene Chip Human Mapping 50K Xba Array. However, this SNP genotyping platform not only contain fewer number of SNPs, but also cause a large number of SNP markers to be excluded when cross-compared with the 500K platform used in the Thai Ayutthaya and surrounding cities set of data. This poses a limitation in genetic analysis of recent Thai people in comparison with Asian indigenous people. The results concerning Thai people and indigenous Asian people are from minimal number of SNP markers as shown in Table 3.2.

Table 3.1 Original sets of samples used in this study

Sets of samples	Number of SNP markers	Genotyping SNP platform	Number of individuals
Thai samples from Ayutthaya and surrounding cities	224,477	Affymetrix 500K	32
Recent Thai populations from Thalassemia project	566,200	Illumina 500K	421
Recent Thai populations from Depression project	570,856	Illumina 500K	618
Pan-Asian individuals	54,794	Affymetrix 50K	1,928
Global populations	246,554	Affymetrix 500K	850

3.2 Data preparation

Only overlapping SNP data from different sources were used. Table 3.2 shows the actual overlapping SNPs used in this study. Then they were checked for matching strand (deriving from the same DNA strands, not the complimentary ones). Minor and major alleles in each polymorphism were calculated. Genotypes were encoded into numeric format according to requirements in different analyses. SNP identification number matching, strand checks, major and minor allele count were performed by python coding (see Appendices A and B).

Table 3.2 Sets of combined samples and overlapping markers used in all analyses

Sets of samples	Populations	Number of overlapping SNP markers	Number of individuals
Global populations	All populations in Xing *et al.*, 2010 and samples from Thai Ayutthaya (and surrounding cities)	226,412	882
Eurasian and Polynesian	Eurasian and Polynesian populations in Xing *et al.*, 2010 and samples from Thai Ayutthaya (and surrounding cities)	226,412	564
Populations with expected shared ancestry with Thai population	Xing *et al.*, 2010 and samples from Thai Ayutthaya (and surrounding cities)	226,412	173 and 234 (include more Indian individuals and Nepalese)

Table 3.3 Sets of combined samples and overlapping markers used in all analyses (continue)

Sets of samples	Populations	Number of overlapping SNP markers	Number of individuals
Asian indigenous populations	HUGO, 2010 and Thai Ayutthaya (this study)	12,737	1, 960
Recent Thai and Central indigenous Thai populations	Thai individuals from Thalassemia project, depression project and samples from Thai Ayutthaya (and surrounding cities) (this study)	39,049	1,024 and 455
Recent Thai populations	Thai individuals from Thalassemia project, depression project	553,892	992

3.3 Principle Component Analysis (PCA) and iterative pruning PCA (ipPCA)

As PCA brings out the significant components in genetic data, clusters become more observable. PCAs were performed at individual levels using ipPCA (Intarapanich *et al.*, 2009) implementation on MATLAB (version r, 2008a). This tool performs normal PCA on its first iteration. Then it separates and clusters the more distantly related sets of samples together in the next iterations, providing a zoom-in analysis of the genetic data without the process of manual removing and re-grouping sets of samples for PCA. It is therefore easier to monitor the relationship between sets of individual genetic data through clustering in each iteration of ipPCA. In other words, at lower iterations, ipPCA clusters major groups of individuals together and it distinguishes the difference among the individuals in the initial clusters at higher iterations. This tool is suitable for detecting discreet sub-populations, particularly those in proximal geographical area like mainland Southeast Asia.

3.4 ADMIXTURE analysis

A model-based algorithm implemented in ADMIXTURE (Alexander *et al.*, 2009) was used to determine the genetic ancestries of each individual in a given number of populations without using information about population designation. Similar to ipPCA iteration, we can understand the detailed relationship between groups of individuals when higher numbers of genetic founders (K) are given. For example, ADMIXTURE analysis at 3 K clusters individuals into 3 main continental groups; the model views sources of the genetic data as coming from 3 founders or ancestries. ADMIXTURE tests are sensitive to population samples included in the analysis; more populations could mean higher number of founders present. Figure 3.1 provides a simplified illustration of how individuals from different populations could be clustered in ADMIXTURE and how numbers of populations included in each ADMIXTURE test is important in making sense of ADMIXTURE analysis.

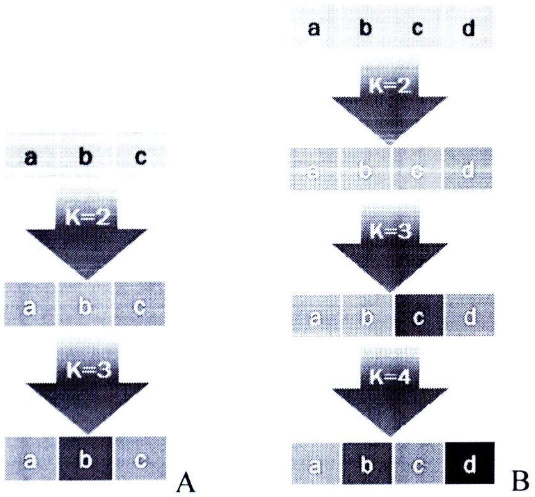

Figure 3.1 Illustration of how ADMIXTURE identify and cluster individuals into groups of populations.
The grey boxes represent unclassified populations (composing of individuals--this is a simplified representation). The colored boxes represent ADMIXTURE results after its clustering analysis. Same colors mean the populations are grouped together, having shared ancestry. A) When fewer populations are included in the analysis, lower K's could suffice in clustering genetic similarity among the samples. B) When more distantly related populations are included, higher K's could reflect actual genetic similarity among the included samples better.

3.5 Distance analysis

Individual allele sharing genetic distances (ASD) were used as a measure of genetic distance between individuals. ASDs were calculated from SNP genotype data using AWClust (Gao and Starmer, 2008) which is a non-parametric population structure analysis written for R. ASD matrices are visualized as phylogenetic trees. While a model-based approach like ADMIXTURE provides specific analysis, ASD analysis together with PCA provide non-model reference which is imperative in the case that model assumptions, e.g. Hardy-Weinberg, do not hold. Mismatch results from the two approaches could also provide us insights into the characteristics of the genetic data in hand.

In addition, in order to quantify and visualize degree of difference from ADMIXTURE genetic component ratio results, we used pairwise similarity distances. The values were calculated from population's averaged ADMIXTURE ancestry estimates, using norm of genetic vectors. Average ancestry estimates are calculated for each genetic component in each population. The sum of averaged ancestry estimate vector still equals to one. The norm of vector was determined using this formula:

$$\theta = arccos\left(\frac{\mathrm{a}\,.\,\mathrm{b}}{\|\mathrm{a}\|\,\|b\|}\right)$$

When *a* and *b* are averaged ADMIXTURE ancestry estimates. For example:
Averaged ancestry estimates for Northern Thai population and Northeast Thai population at K = 3 is

N	0.605785	0.042879	0.351336
NE	0.385007	0.109164	0.50583

Therefore, sizes of Northern ancestry estimations can be calculated by:

$$||a|| = \sqrt{0.605785^2 + 0.42879^2 + 0.351336^2}$$

$$a.b = \sqrt{(0.605785)(0.385007) + (0.042879)}(0.109164) + (0.351336)(0.50585)$$

Then arccosine values could be calculated. The greater the angle between ancestry estimations is, the higher their degree of difference will be.

3.6 F_{ST}

Pairwise F_{ST} distances (Wright, 1949) were calculated for four recent Thai populations from four regions of Thailand at 553,892 SNPs. The model included three interrelated parameters for diploid populations: F_{IT} (correlation between alleles within an individual relative to the entire population), F_{IS} (correlation between alleles within an individual relative to the subpopulation to which the individual belongs) and F_{ST} (correlation between alleles chosen randomly from within the same subpopulation relative to the entire population). An example of formula used for pairwise genotype frequency calculation is as follow:

$$x_{11.1} = p_1^2 + f_1 p_1(1-p_1)$$
$$x_{12.1} = 2p_1(1-p_1)(1-f_1)$$

...

Where p_1 and p_2 are the frequencies of allele A_1 in the first and second population, respectively. $x_{11.1}$ is the frequency of genotype A_1A_1 in the first population and $x_{12.1}$ is the frequency of genotype A_1A_2 in the first population. f_1 is an coefficient for measure of the frequency of observed heterozygotes compared with that of expected one when genotypes are in Hardy-Weinberg proportions.

3.7 Individual tree reconstruction

The dendrograms of individuals were reconstructed based on ASDs using the neighbor-joining algorithm in Molecular Evolutionary Genetics Analysis software package (MEGA version 4.0) (Tamura *et al.*, 2007) .

3.8 Haplotype inference

Haplotype of 2 chromosomes were inferred for each individual from its genotypes with fastPHASE version 1.2 (Scheet and Matthew Stephens 2006). The number of haplotype clusters was set to 20, the number of random starts of the EM algorithm (-T) was set to 20. The number of iterations of EM algorithm (-C) was set to 50. This analysis was used to generate a best guess estimate of the true underlying patterns of haplotype structure. As the number of individual samples could affect the accuracy of the estimation, only populations with more than 20 individual samples were included.

3.9 Haplotype analysis

The analysis focuses on the number of haplotypes each population has and the number of haplotypes two populations share, aka. haplotype sharing. The shared number of haplotype could indicate the degree of their close relationship. This was calculated by counting the number of haplotypes that are shared between two populations. Counting was performed by python codes.