CHAPTER 1

INTRODUCTION

1.1 Principle and rationale of research topic

Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide in genome sequence is changed, for example, from AAG**G**TAA to AAG**C**TAA. For a variation to be considered as SNP, it must occur at least 1% in the population. SNPs are the most common type of genetic variations among peoples and appear in every 100 to 300 base pairs along the 3 billions base pairs of the human genome. The SNPs can occur in both coding and noncoding regions of the genes, and many SNPs have no effect on cell function (Kwok, 2003). The properties of SNPs are stable from an evolutionarily standpoint and not changing much from generation to generation. Because the SNPs have these occurrences and properties, they are easier to follow in population studies, and were believed to determine the human difference between any two unrelated individuals (Duerinck, 2001). Moreover, the scientists believe some SNPs can predispose people to disease or influence their response to medicine. Therefore, SNPs can be effective biological markers for scientists to diagnose disease and track population ancestry (Kwok, 2003; Zhou and Wang, 2008).

The North of Thailand has been flourished with the cultural and traditional uniqueness since the prehistoric time. According to the historical evidence of northern Thailand, there were changes and movement of the populations for over 1000 years ago. There are the discovered evidences of the Mon-Khmer speaking group's settlements in Chiang Mai-Lamphun basin. Later on, the area was established as Hariphunchai state, which was prosperous more than any other states in the north before King Mengrai took possession on Hariphunchai state and founded Chiang Mai as the center of Lan Na Kingdom. The Lan Na Kingdom had been thriving, but still had a war between races in the Kingdom and the surrounded settlements for a very longtime. In the King Kawila period about 200 years ago, the Tai: Yong, Lue and Kheun, and the red Karen, were forced to move from the upper city to settle in Lan Na. Consequently, Lan Na had become diverse and mixed society with different people and cultures. In addition, the migration of populations from neighboring states that bordered the upper northern part of Thailand, i.e., Myanmar, Xishuangbanna, and Laos, has been continuing. Therefore, the population of northern Thailand was composed of many ethnic groups (Malasam, 2010).

There are the growing evidences that genetic variation plays an important role in the population studies. For instance, it could be used for studying the genetic structure and relationship between populations, including their ancestor information, of the four Tai speaking groups (Kampuansai *et al*., 2007), or to study the origins and relationships with social structure and linguistic differences in the hill tribes (Besaggio *et al*., 2007). However, these researches considered only the uniparental genetic markers, the mitochondrial DNA and Y chromosome data sets. The SNPs in the genome, therefore, give more information and the researches on SNPs analysis have been increased, such as the developing of systematic approach based on nearest shunken centroid method (NSCM) to identify ethnically variant of SNPs (Park *et al*., 2007). Additionally, the usage of the autosomal SNPs of 73 Asians and two non-Asian HapMap populations (European and Yoruba) to study the general description of Asian population structure and its relation to geography, language, and demographic history (The HUGO Pan-Asian SNP Consortium *et al.*, 2009). Recently, research on the using of an individual's SNP genotypes to predict that individual of ethnicity or ancestry was published (Sampson *et al*., 2011).

1.2 Objective and outline of this thesis

In this thesis, an intuitive approach applying to the Pan-Asian SNP consortium data, which includes genotype data of 58,960 autosomal SNP loci (using Affymetrix SNP array 50K Xba) obtained from 256 individuals of 13 ethnic groups of northern Thailand, was proposed (Table 1). The analytical approaches were as followed: 1) Specific SNP loci for each ethnic group were selected by ranking each SNP from mutual information values. The top SNPs in ranking list were then trained and tested by decision tree technique. 2) The selected SNPs from (1) were examined, for population discriminative efficiency, by correspondence analysis. 3) The same set of SNPs was used for genetic distance analysis and the phylogenetic tree was then reconstructed.

The thesis had five chapters, which were organized as follows:

Chapter 1 presented the principle and rationale.

Chapter 2 provided basic knowledge of the history and background of populations in the North of Thailand. Then, the review of genetic variation, particularly the SNPs was mentioned. In the part of data analysis, the introduction of variable ranking, the theory of ranking criteria, the general information about data classification and their problems were described. Lastly, the analysis of categorical data and population genetic distance was recounted.

Chapter 3 was the material and method part, describing the characteristics of SNP genotype data. The theoretical formulation extensions and calculation for SNP data application were as follow: mutual information (for SNPs ranking), decision tree technique (for population classification), correspondence analysis (for visualization among populations), and genetic distance among populations (for phylogenetic analysis and multidimensional scaling).

Chapter 4 showed the results of applying the computational methods (in chapter 3) to the SNP genotype data from 13 ethnic groups, and the discussion.

Chapter 5 was about the conclusion of the thesis.