CHAPTER 1 INTRODUCTION

1.1 Rationale

Two humans differ at approximately 3 million nucleotides, or about 1 in every 1,000 nucleotides. This degree of genetic diversity is low comparing to that of other species—for example, the estimated 50% higher levels of nucleotide diversity in chimpanzees, or about 1 in every 500 nucleotides. It is, however, through this small fraction of differences that we could gain insights on human phenotypic variation as well as disease susceptibility. The studies of human genetic variation have been recently accelerated owing to the technological advancement in sequencing (Ansorge, 2009; Shendure and Ji, 2008), genotyping technology with microarray chips (Kirov *et al.*, 2006) and availability of data in the last decade, including the completion of human genome project (Collins, 2003) and many other relevant projects in gathering and analyzing human genetic polymorphism databases (Frazer *et al.*, 2007; Sherry *et al.*, 2001). Nonetheless, this low level of differentiation poses a challenge in the use of genetic data in researches on human genetic differentiation. This chapter explains precisely the scope of this thesis and the reason why the study of genetic variation among the Thais is important.

1.1.1 Why Thai population?

This is the first attempt to study Thai population genetics at large with the use of high-density genome-wide autosomal SNPs. There have been very few population genetic studies on Thai people (Besaggio *et al.*, 2007; Listman *et al.*, 2007; Xu *et al.*, 2010). The existing studies were mostly conducted on ethnics in Northern Thailand. To illustrate, Besaggio *et al.* (2007) carried out a study on 10 populations of Northern hill tribal people, including Karen, Lahu, Hmong and Lisu, for instance, using mitochondrial DNAs and Y-chromosomal Short Tandem Repeats (STRs). Listman *et al.* (2007), focusing on the differentiation between Thai, Hmong and Chinese, still used multi-locus STRs. Another more recent and highly cited work on Mlabri (Phi Tong Luang) and other 12 neighboring populations by Xu *et al.* (2010) had begun to use autosomal SNPs as genetic markers.

Naturally, despite their contribution on the demographic history of the populations in the North, the specific focuses of these past studies can provide only limited demographic history and genetic characteristic information of the Thai people at large, *i.e.* the Thais who live in the other regions of Thailand. Therefore, the present study includes Thai populations in wider area ranging from the Central area, where long history of inhabitance has been recorded, the Southern, the Northeastern as well as the Northern regions. Moreover, the current study is the first attempt to study Thai population genetics with the use of highly dense autosomal SNP markers. The use of genome-wide SNP markers ensures that a large proportion of genome region, including the protein-coding and non-coding parts, is to be explored. The high density (over 200,000 loci) of SNP markers used in this study also has a more powerful differentiation power and is therefore suitable for detailed investigation into the discreet substructure within Thai population.

To conclude, the major contribution of the present study is two-fold. First, it provides a non-biased view of Thai people genetic makeup, through the use of dense genome-wide autosomal SNPs. Second, it provides a large picture of Thai populations in general, using the samples from four major regions of the country, i.e. the North, the Northeast, the Central and the South.

Thai people are a distinctive population that worth investigation. Furthermore, preliminary results in this study showed that the genetic makeup of Thai population was distinctive (Figure 4.5 in Results) and indicated a unique pattern of genetic admixture, composing of mixed ancestry. Investigation into Thai genetic characteristics and population differentiation would surely leverage applications in the fields of the modern molecular medical advancement (e.g. genome-wide association studies, identification of disease susceptibility and pharmacogenetics) and forensics (individual assignment to population) (see section 1.1.2 on *Practical uses*).

Thai people could represent a large group of genetic makeup in the mainland Southeast Asia. The geography of mainland South East Asia composes of mostly plains with no major natural boundaries, such as extreme habitats e.g. deserts or exceedingly high mountainous areas. Consequently, interbreeding between the members in previously separated populations are likely. Previous studies on global mitochondrial DNAs and Y chromosomes (Hill *et al.*, 2006; Tajima *et al.*, 2002; Feng Zhang *et al.*, 2007) indicated multiple migrations into this area. Also, as geographical distance plays important roles in patterns of genetic variation (Bamshad *et al.*, 2004; Handley *et al.*, 2007), the greater distance between the Central Thai region and other countries in East Asia means that it is likely that our genetic makeup is more distant to the Chinese but more closely related to populations that are scattered in the mainland Southeast Asia. It would be interesting to see who we are close to and whether we can share genetic study results and implications.

1.1.2 Why study human genetic variation?

Practical uses. Investigation of the genetic differences between populations has intensified, largely due to the successes in genome-wide association studies which deliver numerous reports on putative associations between polymorphisms and morphological characteristics and disease susceptibility. For example, single nucleotide polymorphisms have been used in identification of genetic variants that contribute to complex human traits (McCarthy *et al.*, 2008) such as obesity (van Ooij, 2009) and hepatitis C viral infection (Walley, Asher, and Froguel, 2009) . However, in order to make an extensive use of such results, we have to take into consideration the existence of population stratification. For example, population differentiation is a major issue that affects the design of case-control study (McCarthy *et al.*, 2008). In addition, the classification of human through genetic variation data has applications in pharmacogenetics and forensic science.

Evolutionary insights. The study of human genetic variation can provide insights on both the investigation of human origin and the more recent demographic history of populations. The study on population substructure, like the investigations on the Thai population genetics to be carried out in this thesis, can reveal the degree of different natural processes that shape the variations in the population. For example, the demographic parameter such as heterozygosity indicates the levels of nucleotide variation and provides insights on the age of the population and the size of founder population (Jobling, 2004). Degrees of admixture can indicate the past migration into the geographical areas (Jobling, 2004). In addition, the pattern of allele frequency or site frequency spectrum can provide insights on the forms of natural selection having taken place in the population (Dermitzakis, 2010). Inferences regarding selection could provide powerful tool in functional studies, such as the prediction of possible disease-related genomic regions. Therefore, the knowledge of a population's past can even allow us to make predictions about the future or, putting more precisely, about unknown sets of data. It could provide direction upon which we should focus in order to answer some present biological questions.

In conclusion, evolutionary perspective on population genetic variation not only contributes to the scientific community in terms of its philosophical aspects, but also produces concrete yields. Understanding in a population history and its shared ancestry can be linked to the understanding in the population's pathological heritage. Furthermore, in combination with geographical and ecological perspectives, knowledge on evolutionary history of a modern population can elucidate significant phenotypic differences and the prevalence of disease-causing mutation. After all, the past is not something that merely happened, but it has accumulated and acted as the source of the present. Such view plays important roles in the analyses in the field of population genetics and is to be revisited throughout this thesis, particularly in the rationale of some methodologies and result interpretation.

1.1.3 Why SNPs and haplotypes

There are many forms of polymorphisms at the DNA level. They serve as genetic markers and have become fundamental tools for modern studies in molecular biology, among which includes population genetics, disease association studies and pharmacogenomics. Each type has both advantages and disadvantages. There are a number of reasons this study focuses on single nucleotide polymorphisms (SNPs).

Due to technological limitation, early researches on human genetic differentiation relied mainly on mitochondrial DNA (mtDNA), non-recombinant Y chromosome (NRY) and, a bit later on, microsatellites (short tandem repeats or STRs). These types of genetic markers are highly polymorphic (*i.e.* having more discrimination power) and hence are necessary when only a few number of individuals could be interrogated (Dermitzakis, 2010).

However, there have been recent concerns over the use of these molecular markers. Mitochondrial DNA, the most popular marker of molecular diversity, was recently reviewed as not suitable in studying population history, mainly due to its unpredictable mutation process and evolutionary rate (Galtier *et al.*, 2009). The problem of Y chromosome is that its effective population size is a quarter that of autosome. Y chromosomes thus are more susceptible to genetic drift.

The completion of human genome project (Collins, 2003) and the availability of human genetic polymorphism databases (Frazer *et al.*, 2007; Sherry *et al.*, 2001) have made the use of neutral and autosomal polymorphisms possible. Currently, due to the robustness in SNP genotyping data (Vignal *et al.*, 2002), the use of SNPs as molecular markers has become popular, replacing the more polymorphic form, yet less robust, microsatellites. In other words, SNPs, by their nature of variation among the four possible genetic bases, do not contain as much polymorphic information as STRs, which could exist in many forms ranging up to many hundreds possible copies in one genome. However, SNPs contain fewer artifacts and can be obtained in high density. This promotes their suitability in a large-scaled study of multifactorial diseases and population genetics which rely on delicate molecular signals in human genome.

1.1.4 Why study in fine scale?

This study encompasses more than 220,000 SNP loci and thus is considered a fine-scale genetic differentiation study. Molecular markers in such density have discrimination power advantages and are necessary to distinguish discreet subpopulation (Rosenberg, 2002; Bamshad *et al.*, 2004). Insufficient number of markers can lead to biases in the differentiation results (Bamshad *et al.*, 2004). Due to the reasons discussed in section 1.1.1 pertaining to the discreet sub-structure of Thai population, a fine-scale genetic variation study would not only be useful but necessary in obtaining clear results in this study.

1.2 Scope

This study investigates the genetic characteristics and differentiation of Thai populations from four regions of Thailand: the North, the Northeast, the Central and the South. Also, published data that include sample populations from Thailand, as well as other Asian and world populations would be used for comparison. The genetic data to be used in this study are multi-locus autosomal SNPs and haplotypes. SNPs are genotyping data. Haplotypes would be inferred from SNPs.

Methods covered mainly are those used in population genetics, namely statistics, principal component analysis, and probability models for population stratification analysis. This work would also interpret and discuss the results using concepts in population genetics, evolution theories and ethnolinguistics.

1.3 Goals

The goals of this study are to gain a better understanding of the genetic diversity and recent demographic history of Thai populations and to provide reference information on such aspects for potential future uses in the studies of population genetics, case-control studies, forensics or pharmacogenetics in Thailand.

1.4 Objectives

1. To analyze the genetic data (SNPs and haplotypes) of Thai population, both in large scale and in detail, to construct a descriptive information of Thai genetics and genetic variation.
2. To compare the large-scale result with other Asian and world populations to find genetic relatedness of Thai people to other populations.