CHAPTER 2 BACKGROUND

2.1 Human genetic variation

2.1.1 Single nucleotide polymorphisms (SNPs)

A SNP is a single base change in DNA sequence, mostly derived from a point mutation. SNPs are usually *biallelic*, having two possible alternatives of nucleotide at a certain position, even though a nucleotide alteration at a single locus theoretically produces four possible alternatives. There are two explanations for this observation (Vignal *et al.*, 2002). First, mutation generally occurs very slowly. Studies in non-functional regions (*i.e.* neutral mutation which is not subjected to natural selection) like pseudogenes (Martinez-Arias *et al.*, 2001) reported substitution rates in human to be at only about 1 x 10^{-9} per nucleotide per year. This means that the probability of alteration in two independent bases occurring at the same position is very low. Secondly, transition (changes between the same types of bases) occurs more often than transversion (changes between different types of bases). The observed ratios for transition over transversion vary, ranging from about 1.5 (Collins, 2003) to 2.2 (non-CpG regions) and 3.6 (CpG region) (Nachman and Crowell, 2000). Therefore, polymorphisms data tend to exhibit bias over transition change; that is between two purines, A/G, or two pyrimidines, C/T.

As they are biallelic co-dominant markers, SNPs have to be studied in a higher number in the field of population genetics in order to obtain a sufficient resolution power. In the present time, this can be achieved through the advancement in sequencing and genotyping techniques, particularly the gene chip technology (Kennedy *et al.*, 2003; Syvänen, 2005; Gunderson *et al.*, 2005) which allows a large number of standardized sets of SNP (as well as Copy Number Variance or CNV and STR) genotypes to be produced for new population samples with a reference to previously studied results. A number of studies on population genetics (Jennifer B Listman et al. 2007; Holsinger and Weir 2009; Price et al. 2010; Luikart et al. 2003; Reich et al. 2009) made use of this type of standardized microarray platforms. The chips are produced from published information on whole-genome sequencing and SNP discovery. For example, this study used Affymetrix 500K microarray in the process of genotyping SNPs in Thai population. This means we studied only some selected regions of human genome, around which previous whole-genome researches have declared an abundant presence of SNPs. There are some concerns over this protocol that it produces bias toward common alleles (as oppose to rare alleles) and results from European population (Rosenberg *et al.*, 2010). However, it provides a cost-efficient method for preliminary research on new populations. With the molecular technological development going on at this rate, the cost of sequencing and genotyping is deemed to be dramatically reduced (Hawkins *et al.*, 2010), facilitating future genome sequencing and producing a non-biased genome variation results for Thai population.

2.1.2 LD and Haplotype

For detecting population expansion, a single SNP or even a few would be less useful because the accumulation of new mutations require longer time period for slowly evolving loci. Knowledge of haplotypes and linkage disequilibrium is therefore

essential for the construction of phylogenies of individual loci which are important in many areas of evolutionary genetics.

A haplotype refers to the combination of allelic states of polymorphic markers along the same DNA molecule, *i.e.* on the same chromosome. If we type (as in the process of finding SNP genotypes) Y-chromosomal, a male's X-chromosomal or mtDNA markers, we immediately derive a haplotype because these molecules are haploid. Typing autosomal markers (and X-chromosomal markers in females) do not directly produce a haplotype, except for the case that both chromosomes have the same haplotype (homozygosity). Therefore, the SNP genotypes obtained from genotyping process in the study of diploid organisms, like in this study, are diploid genotypes, composing of two haplotypes combined. In this case, haplotype inference has to be performed to deduce the most probable haplotype from the genotype information. This process is sometimes called *haplotype phasing*.

Processes shaping haplotype patterns include mutation and recombination. In the case of mtDNA and NRY chromosome, both of which are nonrecombining, the haplotype diversity is due only to mutation. However, in autosomes, meiotic recombination plays the major part in creating haplotype diversity, breaking and shuffling existing haplotypes, yielding new ones in the next generation. When recombination is at play, the level of diversity could increase dramatically than with mutation alone (Jobling, 2004).

Reduced recombination between certain alleles at separate loci results in the higher tendency of the alleles to be co-inherited. Statistics could reveal association between alleles that are tightly linked by their loci. This kind of property is known as linkage disequilibrium (LD) (Figure 2.1) (Ardlie *et al.*, 2002). In other words, LD occurs when two alleles are found together on the same chromosome more often than expected if they were segregated at random. Biologically, LD arises in the similar way to how polymorphisms are generated by mutation and then shuffled by recombination. As shown in Figure 2.2, LD is created when a new mutation occurs on a chromosome and is gradually eroded by recombination.

In the past decade, LD has become the focus of intense study because it facilitates the mapping of complex disease loci through whole-genome association studies. Interest in LD has risen further because of the increasing availability of genome sequences and high marker density genetic maps. This opened up possibilities to detect genes and mutations underlying quantitative genetic variation by association mapping.

The oldest and simplest measure of LD is to find the difference between the observed frequency and the expected frequency (theoretical allele frequency for random segregation) of a haplotype which has two loci. The expected frequency is derived by probability. The two types of frequencies (observed and expected) are to be tested with statistics (Fisher exact test) for their difference. If they are significantly different, LD is said to exist. Later models are built on this simple concept, making several adjustments to better suit genetic data.

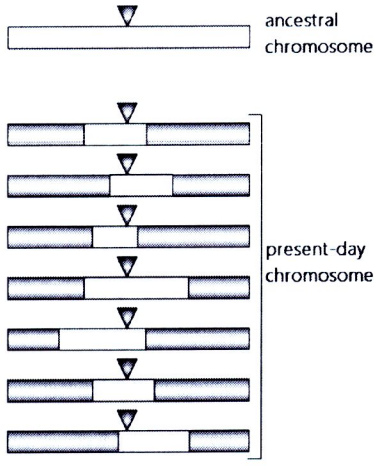

Figure 2.1 Linkage disequilibrium around an ancestral mutation (triangle). The present-day chromosomes have both the chromosomal stretches derived from common ancestor (lighter areas), in whose chromosome mutation occurs, and the new stretches introduced by recombination (darker areas). Markers that are physically close (lighter areas in present-day chromosome) tend to remain associated with the ancestral mutation. Recombination limits the extent of the associated regions over time (different sizes of the lighter regions).
Source: Ardlie and Seielstad, 2002

Furthermore, analyses of LD and haplotype structures in human chromosomes clearly indicated a block-like structure of the genome (Cardon, 2003; Stumpf and Goldstein, 2003; The International HapMap Consortium, 2005). These structures, termed *haplotype blocks*, are segments of consecutive alleles that appear to be co-inherited, *i.e.* having been through little or no recombination. LD is high within the blocks and low at the regions between them (indicating recombination hotspots).

Relatively high levels of LD extended several tens of centimorgans, and LD was frequently detected between unlinked markers. The magnitude of LD varied depending on how the population was sampled. It also varied across different chromosomes, and was shown to be a function of sample size, inter-marker distance and marker heterozygosity. A recent admixture event in the population led to an ephemeral increase in LD.

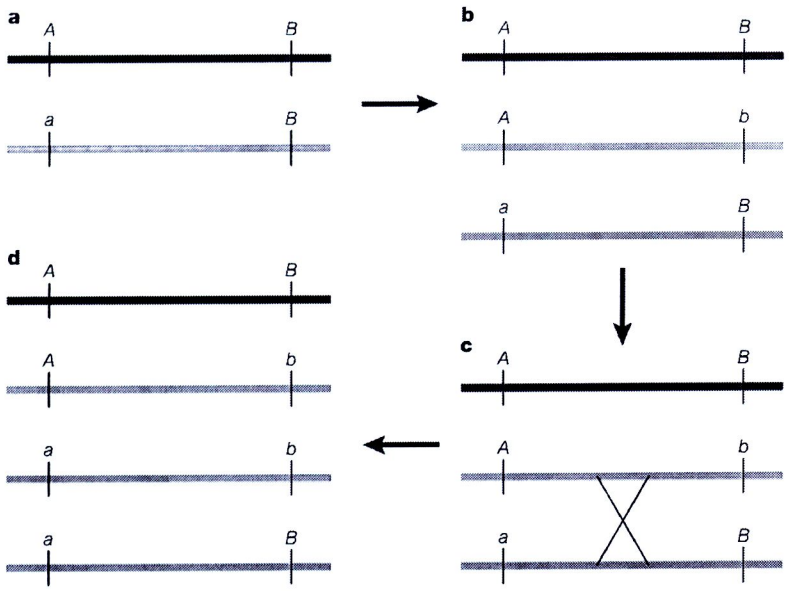

Figure 2.2 The erosion of linkage disequilibrium by recombination.
At the beginning, there is a polymorphic locus with allele *A* and *a* (a). A mutation occurs, changing the allele *B* to *b*, at a locus near allele *A* on one chromosome, resulting in another new haplotype, bearing a new polymorphism (b). Now there are three haplotypes in the population. The *b* allele will always be found on a chromosome with the *A* allele at the adjacent locus. However, the association between alleles at the two loci will gradually be disrupted by recombination between the loci (c). This results in the creation of the forth haplotype. The LD among the markers in the population will eventually decline as recombination continues to take place.
Source: Ardlie and Seielstad, 2002

Characterizing LD can provide insights into the biology of recombination human demographic history and is of central importance for gene-mapping studies. Understanding the extent of LD is relevant to many evolutionary genetic topics including the evolution and rate of genetic recombination, the mechanistic basis of genetic correlations, explanations of why heterozygosity and fitness are often positively correlated, and as a tool to measure effective population size, particularly of bottlenecked populations.

As the correlation of LD between populations will be reduced each generation by an amount that approximately depend of the recombination rate between markers, observed LD in different populations descended independently from a common ancestral population can be used to estimate their divergence. The extent of linkage disequilibrium has been well established to differ across populations, particularly between African and non-African groups (Slatkin 2008; Conrad et al. 2006). A novel strategy has been developed for assessing interpopulation variations in regional patterns of linkage disequilibrium (Gu *et al.*, 2007). This approach has been applied to the genomes of populations in public databases, identifying regions where linkage disequilibrium is considerably different, some of which exist in regions associated with phenotypic variation. It has been shown that such diversity in linkage disequilibrium

can challenge replication studies and meta-analyses while benefiting the pursuit for the functional variants in fine-mapping studies.

2.1.3 Other types of markers: mitochondrial DNA, microsatellite and Y-chromosomes

Mitochondrial DNA has been used extensively in evolutionary study because of their uniparental mode of inheritance and high rate of mutation accumulation, up to 5-10 times higher than nuclear DNA. However, strong within-species homoplasy poses problems when analyzing mtDNA population data, confounding its evolutionary history (Ballard and Whitlock, 2004; Ballard and Rand, 2005; Galtier *et al.*, 2009). Despite its popularity, mtDNA on its own is not the most suitable marker for the study of recent historical events. It should thus be supplemented with analyses of other types of markers such as Y-chromosome or autosomal DNAs (Pakendorf and Stoneking, 2005; Rubinoff and Holland, 2005).

Microsatellite is a DNA motif of tandemly repeat units of one to six base pairs. Microsatellites display relatively higher mutation rates, having substitution rates of around 10^{-9} mutation per nucleotide per generation (Wang 2006). New microsatellite mutations were shown to differ from those of the parental allele by one or more repeats. This makes it suitable for constructing stepwise mutation model in familial studies. However, it is not an optimal choice in population genetic studies.

A typical error with microsatellite allele calling is size determination (Vignal *et al.*, 2002). New alleles could be described where in reality are artifacts or an allele at the locus in consideration could be missed. Although this type of error could be easily corrected in family analyses, it could cause drastic consequences in data interpretation in a large scale study like population genetics. In the case of SNPs, the only errors in allele calling are the non-detection of one allele, resulting in a heterozygote individual being genotyped as homozygote, and *vice versa*.

STRs and mtDNA sequences have been tools of choice in molecular studies of evolution and population since the early 1990s (Morin, 2004). Both kinds of genetic markers represent rapidly evolving DNA sequences that are informative for answering population-level questions. However, the high information content, a result of high mutation rates which pose limitation on subsequent data analyses. Also, the loci can be sparse in the genome. By contrast, mutations observed as SNPs are abundant, covering both coding and non-coding regions (Liu and Cordes, 2004). Another technical problem with STRs is that it is not always possible to compare data produced by different laboratories, due to the inconsistencies in allele size calling. Even though this does not pose much problem in familial studies, it can be a serious issue when genotyping data from isolated individuals. The inconsistencies of STR, particularly in the case of large size differences between alleles, calling are mainly due to the large variety of laboratory techniques and different calling software (Vignal *et al.*, 2002). The allele calling problem is simpler in the case of SNPs, for which the results are mostly binary in nature, leading to either one or two of the possible alleles or, in most cases, simply the presence or absence of the alternative form.

Y-chromosome. 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 databases of Y-chromosome variations are still more limited than those of mtDNA.

2.2 Genetic admixture

The term genetic admixture is reserved for the formation of a hybrid population from the mixing of ancestral populations that have previously been in relative isolation from one another, e.g. expansion of one population into a region inhabited by a previously isolated population. It is the result of cumulative gene flow from when the population first met to the present day.

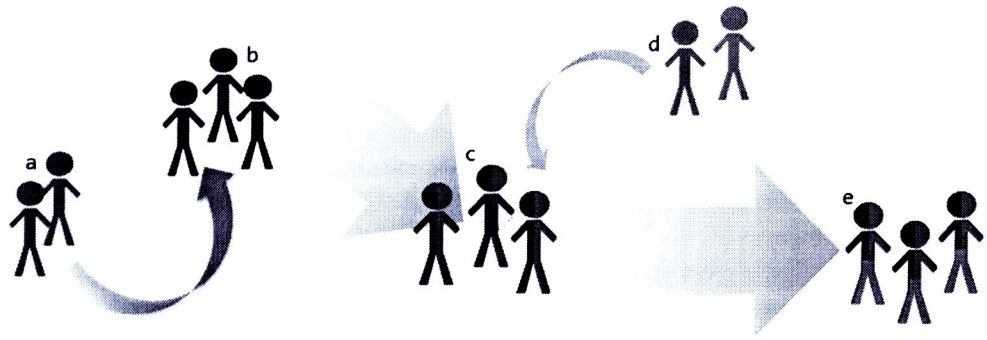

Figure 2.3 Formation of a hybrid population or genetic admixture
A hybrid population (e) arises from genetic mixing of ancestral populations (a, b, c, d) that have previously been in relative isolation from one another.

2.3 Allele frequencies

SNP allele frequency differentiation between and within human populations has been extensively studied both for characterizing population structure and demographic history and for detecting loci that experience the effects of natural selection.

Demographic events affect the distribution of SNP allele frequencies. For example, a large proportion of rare alleles indicates recent expansion, as mutations that have occurred since expansion will not have had time to spread through the population. Understanding neutral allele frequencies is useful for identifying regions of the genome affected by natural selection.

When there are at least two alleles present in the population, the least frequency for a minor allele should be 1% or greater.

2.4 F_{ST}

F_{ST} is a convenient statistical measure of population differentiation (reviewed by Holsinger and Weir, 2009). It is directly related to the variance (amount of variation around a mean value) in allele frequency among populations. Therefore, the values inversely indicate the degree of resemblance among individuals in the populations. For example, small F_{ST} means that there is little variance between the allele frequencies of the tested populations, indicating the populations are genetically similar. On the contrary, large F_{ST} means that the allele frequencies are different. F_{ST} has been extensively used to characterize population structure and demographic history (Jakobsson *et al.*, 2008; Reich *et al.*, 2009; Keinan *et al.*, 2007; The International

HapMap Consortium 2005). When there are at least two alleles present in the population, the least frequency for a minor allele should be 1% or greater.

2.5 Language families in mainland Southeast Asia

There are five distinct language families in Mainland Southeast Asia: 1) Austroasiatic (AA), 2) Sino-Tibetan (ST), 3) Tai-Kadai (TK), 4) Hmong-Mien (HM), and 5) Austronesian (AN) (Ooi, 2004) (Figure 2.4). The distribution of AA language is fragmented from northeastern region of India to the west of Vietnam and from Yunnan to Malay Peninsular. On the contrary, there is a continuous distribution pattern of Tai-Kadai speaking populations from the southern region of China, Shan States of Myanmar, to Thailand.

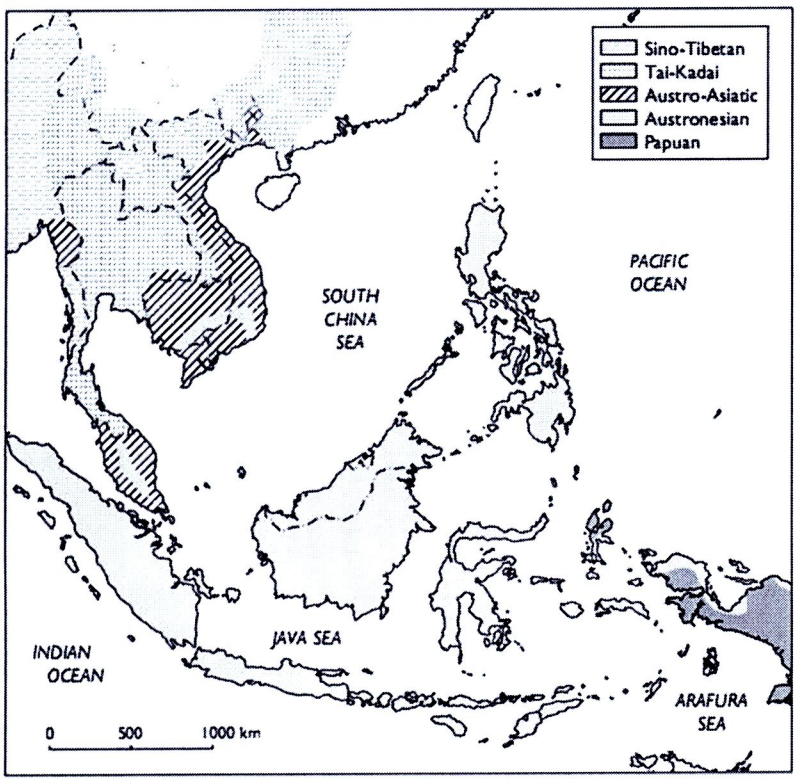

Figure 2.4 Ethnolinguistic pattern of modern Southeast Asia
Adapted from Ooi (2004) with extension of Tai language family in Southern China