CHAPTER 2

LITERATURE REVIEW

2.1 General information of Northern Thailand

Northern Thailand is the region full of history, customs, traditions and culture since long back, which is different from other parts of the country. In addition, the major topography is mountain ranges, which create a beautiful nature and the proper places for many hill tribes habitation. Furthermore, the neighbouring countries of the northern land: Myanmar, China, and Laos, making this area even richer of traditional, cultural and custom collections. However, the tradition and culture of northern Thai people, as found through art and culture, and especially language, had still remained.

In the past, the north of Thailand was called Lan Na, but today, Lan Na is understood to cover the area comprising eight provinces: Chiang Mai, Mae Hong Son, Chiang Rai, Lamphun, Lampang, Phayao, Phrae, and Nan (Ongsakul, 2005). However, geographically designated (Figure 2.1), the upper northern Thailand composes of nine provinces; Chiang Mai, Mae Hong Son, Chiang Rai, Lamphun, Lampang, Phayao, Phrae, Nan, and Uttaradit (Boontang; the Royal institute, 2003). The number of population in the upper northern Thailand is about 6.1 millions (Department of provincial administration; Ministry of Interior, 2011).

.

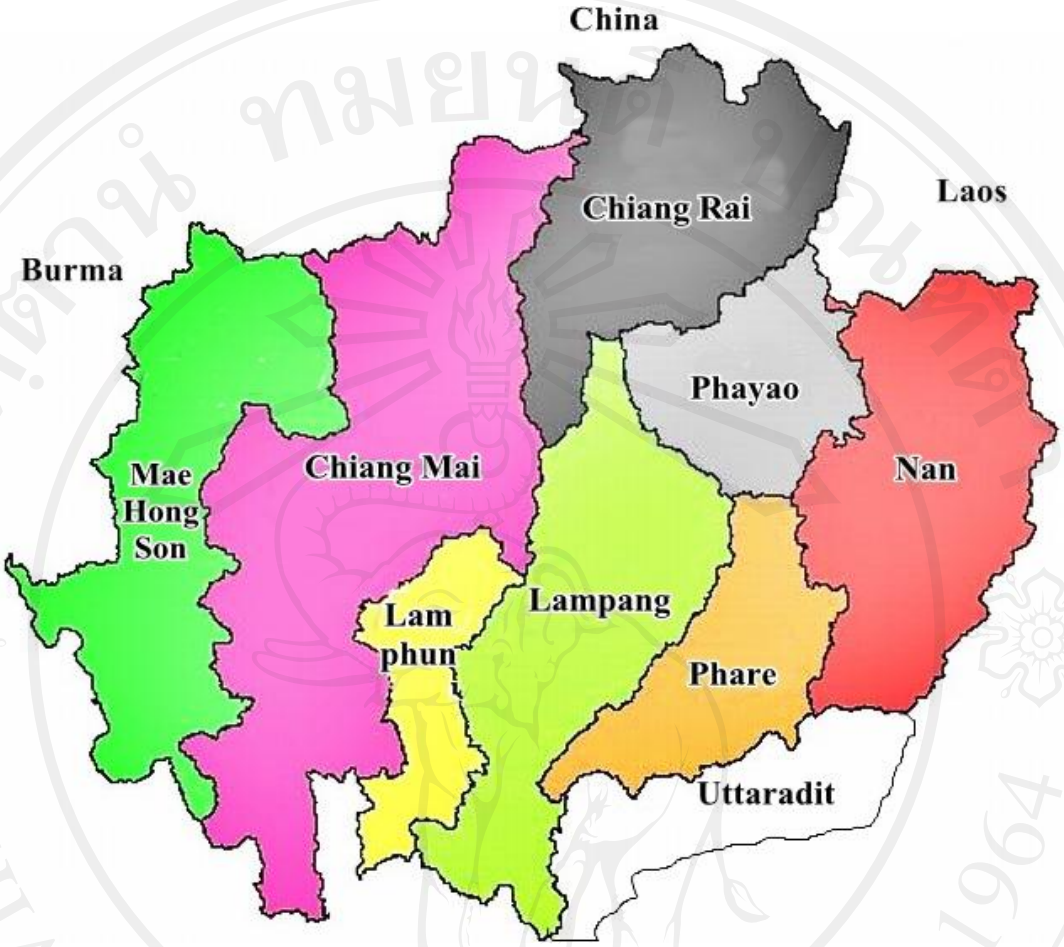

Figure 2.1 Map of northern Thailand

2.2 Ethnic groups in northern Thailand

Thailand is one of the major countries of Southeast Asia and can be considered as the source of strategic importance in the history, since Thailand locates in the centre of Southeast Asia and surrounds by Myanmar, Laos, Vietnam, Cambodia, and Malaysia. Therefore, Thailand was among the distress of racial war and these neighbor countries falling under the colonial hunting of western countries. These suffering events had resulted in the immigration of small minorities from the neighbouring countries in order to survive, hoped to have a better life and could live peacefully in Thailand. The linguists use the language as one of the ethnic group

classification whiles the cultural anthropologist using the common culture characteristic, civil society, and history for the classification (Eriksen, 1993). Although the ethnic groups in northern Thailand have always assimilated in the society, but some of them are still disregarded as Thai citizens and have no citizen right.

Yuan origin: according to an immigrant hypothesis, Yuan moved southward from the area where now three countries, Burma, Laos and Thailand, meet. Then they conquered the native Mon-Khmer peoples along their southern migration route until they settled in Chiang Mai-Lamphun basin, where it is the present day northern Thailand (Penth, 2000). Nowadays, 87.5% of Tai people in the north of Thailand use northern Thai dialect or Khum Mueang. The Lan Na art and culture, which are the trait of the Yuan, can be seen all around Chiang Mai-Lamphun basin, such as architecture and painting in many temples and the earth walls around Chiang Mai city. These are the reflection of civilization of the Lan Na Kingdom.

In this study, Yuan or Khon Mueang in northern Thailand, were from Chiang Mai and Lamphun. Another Yuan group was from Saraburi province in central Thailand. The Yuan of Saraburi is believed to be the direct descendent of the Yuan who used to occupy Chiang Saen city. They migrated from the north to central Thailand about 200 years ago, and still call themselves as Yuan as well as speak the Yuan language (Kampuansai, 2008).

Lue ethnic group speak Tai Lue language. They live in Xishuangbanna of southern Yunnan, the first Tai Lue Kingdom. Some of the Lue live in Yong city and the eastern part of the Shan State of Myanmar. The Lue migrated to cities in north Thailand because of the King Kawila's states policy "put vegetables into the basket

and put the people into the towns", and incorporating them into the new Lan Na kingdom. This was especially true during the early Rattanakosin period, in the reign of King Kawila, where people had migrated to Lan Na Kingdom to cut down the power of the enemy, the Burmese army. Now the Tai Lue communities have scattered in different provinces in the north, as follow: Chiang Mai, Lamphun, Chiang Rai, Phayao, Nan, Phrae and Lampang (Wichaenkaew, 1986).

Lue under this study were from two provinces, Nan and Chiang Mai, and they have difference histories. Lue people of Nan had migrated from Chiang Rung (Jinghong) city, Yunnan, passed through Laos in 1772 A.D. The Chiang Mai group was first founded in 1389 A.D., during King Saen Mueang Ma of Chiang Mai period, but more Lue migrated into this village later in 1804 A.D.(Kampuansai, 2008).

Yong used to live in Yong city in Shan state of Myanmar. The period between 2339-2356 Buddhist era was important to the history of movement and settlement of the Yong people. They moved from the Yong city and were forcibly settled in Pasang district, Lamphun, and became the majority of citizen in the area (Malasam, 1997). The Yong language is classified into Tai-Kadai subgroup of Autro-Asiatic language family. Their culture and traditions have been carried over to the present time. Now there are about 240,000 to 320,000 Yong people living in Chiang Rai, Chiang Mai, and Lamphun provinces.

Yong in this study was from Pa Sang district, Lamphun province. This group had been founded since the first settlement of Yong people in Lamphun in 1805 A.D.

Khuen is one of the original Tai-Kadai speaking groups, which have a strong and stable foundation of civilization. Chiang Tung city is the origin settlement and the capital city of this ethnic group.

The migration of Khuen into Thailand occurred in the same period as Lue and Yong (1804-5 A.D.). The speaking and writing languages of Khuen are very similar to Lue and Yong (Kampuansai, 2008). They reside in Mae Wang and San Pa Tong district, Chiang Mai. These groups were the samples in this study.

Lawa, or Lua, is a native group of northern Thai people. They are also recognized as the first group who settled in this area before other immigrant groups. The Lawa are still holding their traditional way of life on the high areas. Their language is defined as the Mon-Khmer subfamily of Austro-Asiatic (Boonyasaranai, 1988; Condominas, 1990; Schliesinger, 2000). Lawa now resides in Chom Thong, Mae Cham and Hod districts in Chiang Mai, and in Mae Hong Son province. Lawa in this study were from Mae La Noi district, Mae Hongson province.

H'tin tribal history is still not clear. It is estimated that H'tin people settled in Thailand approximately 40-80 years ago. These tribal people migrated from Laos and have settled only in the Nan province, near the Thai and Laos boarder (Jaroenwong, 1991). The Thai people in the Nan province usually call them "Lua". Their language belongs to the Mon-Khmer branch of the Austro-Asiatic linguistic family, which is the same as for Lawa, Khmu and Mlabri. The H'tin samples from Tung Chang and Chiang Klang districts in the Nan province were used in this study.

Mlabri, or Phi Tong Luang, who used to live deep in the jungles and on the high mountain of Thailand, were rarely seen. Mlabri have ancient Mongoloid features. They are groups of nomadic "hunter-gatherers" that do not have their own culture, tradition, religion, or tribal history (Munwalee, 1984). Mlabri in Thailand migrated from Sayaburi province in Lao during the last century (Pookajorn, 1992). They used to live dispersedly in Laos and Thailand, but they are found now only in

Rong Kwang district, Phrae province and Wiang Sa district, Nan province. Mlabri samples in this study were from Wiang Sa district of the Nan province.

Mon founded the great city of Hariphunchai at the site of today's Lamphun in 769 A.D. Later, the Mon of Hariphunchai expanded their kingdom to the Lampang area. The Mon dominated Northern Thailand until the arrival of the Tai in the 13th century A.D. The Mon of Hariphunchai had advanced writing and administration culture which were partly influenced from Burma and India (Penth, 2000). The Mon language is defined as the Mon–Khmer subgroup of Austro-Asiatic family language (Thianpanya et *al.*, 2004).

Ongsakul (2000) mentioned that around Chiang Mai-Lumphun basin there were four original Mon communities, comprising Nongdoo and Buacow villages, Pa Sang district, Lamphun province; and Korchoke and Nongkrob villages, Maeka, San Pa Tong district, Chiang Mai province. These Mon groups still preserve their language, culture and traditions. The Mon in this study were from Pa Sang district, Lamphun province.

Plang, or Blang, live mainly in the southwestern part of Yunnan province of China, and in scattered communities in Myanmar and Thailand. The Chinese call them Blang, Hkawa, or Pulang. In Thailand, the Plang people call themselves Tai Doi, Khon Doi, or Tai Loi. The Plangs speak a language belonging to the Mon–Khmer subfamily of Austro-Asiatic language. The Plang samples in this study were from Mae Sai and Mae Chan district, Chiang Rai province. They migrated from Xishuangbanna, Yunnan, China, through Myanmar and into Thailand in 1974 A.D. (Kampuansai, 2008).

Palaung normally call themselves Ta-Ang (Di-Ang). Small number of them immigrated to Thailand in 1983 A.D., from Shan and Kachin states in order to escape persecution and oppression of Burmar's military rulers, and have settled in Doi Ang Khang area, Chiang Mai. Nowadays there are about 10 Paluang villages in Fang, Chiang Dao, and Mae Ai district, Chiang Mai province (Laaongpliw, 2003). The Palaung in this study were from Fang and Chiang Dao districts, Chiang Mai province.

Karen is the largest hill tribe group in Thailand. The Karen originated from the East of Tibet and then moved to China. Later, they migrated along the Khong and Salween rivers, and finally reached the northern Thailand about 200 years ago. Most of them live in the dense jungle along the Thailand-Myanmar border. Karen language belongs to the Tibeto-Burman group of the Sino-Tibetan language family. The major groups in the North of Thailand are Sgaw or white Karen, and Pwo. Each group is distinguished not only by language but also by their women clothing design. Nowadays, the Karen people mostly inhabit the provinces of Chiang Mai, Mae Hong Son and Tak (Ministry of Social Development and Human Security, 2003). Karen in this study were from Mae Sariang district, Mae Hong Son province. The people from this location settled there about 137 years ago (Srikummool, 2005).

Hmong migrated from Huang He River in China, where their ancestor lived, due to the Chinese political unrest. A portion of Hmong moved to the mountainous regions of southwestern China and northern Myanmar. Most of Hmong people in the northern part of Thailand migrated from Burma and Laos around 60 years ago.

The Hmong language is classified to the Hmong-Mien linguistic family. The subgroups in Thailand are the black and white Hmong. Today, they inhabit the upper northern provinces: Chiang Mai, Chiang Rai, Nan, Phayao, Lampang, Phrae, Mae

Hong Son, and the lower northern provinces: Phetchabun, Tak, and Phitsanulok. They can also be found in Loei province in northeastern Thailand (Leepreecha, 2005). The Hmong in this study were from Mae Rim district, Chiang Mai province. People of this group were there about 42-47 years ago (Srikummool, 2005).

Yao call themselves Mien. They are a minority group which have a native land in the Valley of Huang He River of northern China since one thousand years ago. They were almost wiped out by invasions and many political disasters. The majority of the Hmong immigrated to Guizhou and Yunnan cities while a small group immigrated to Vietnam, Lao and Thailand (Leepreecha *et al*, 2004). Most of the Yao villages in Thailand are located on the highland of Chiang Rai, Chiang Mai, Phayao, Lampang, and Nan provinces. The Yao had immigrated into Lan Na Kingdom approximate 200 years ago (The association of hilltribe cultural studies of Thailand, 2002). The Yao samples from Chiang Rai province were used in this study. This group settled in the mentioned location about 57 years ago (Srikummmol, 2005). Their language is in the Hmong-Mien family.

2.3 Single nucleotide polymorphisms

The human genome has approximately $3x10^9$ base pairs in DNA of two people, except for the identical twins, which are genetically identical. Any two people have about 6×10^6 base pairs, only 0.1% of the entire human genome, which are different. And yet around the world, all populations of humans are essentially the same; while the differences lie among individuals, not among populations. This leads some geneticists to question the validity of defining race; the biological differences between races are much fewer than the differences among individuals in one race

variation is difference. Genetic variation is the difference in DNA. The “letters” of DNA are molecules called nucleotides: adenine, cytosine, guanine and thymine (A, C, G and T) strung together in long chains called sequences.

A genetic variant that occurs when there is a difference or change in a single base or nucleotide (A, T, G or C) within different members of a species is termed a single nucleotide polymorphism (SNP). For example a SNP might change the DNA sequence GTGT**G**GCT to GTGT**A**GCT (Figure 2.2 and 2.3). A single base change occurs in a population at a frequency of >1% is termed a SNP, while a base change that occurs at a frequency of <1% is considered to be a mutation. SNPs are inherited from one generation to the next without much change (i.e., they are evolutionarily stable) and about 90% of all human genetic variations are SNPs, which occur every 100 to 300 bases along the 3-billion-base human genome (Kwok, 2003: 3-5).

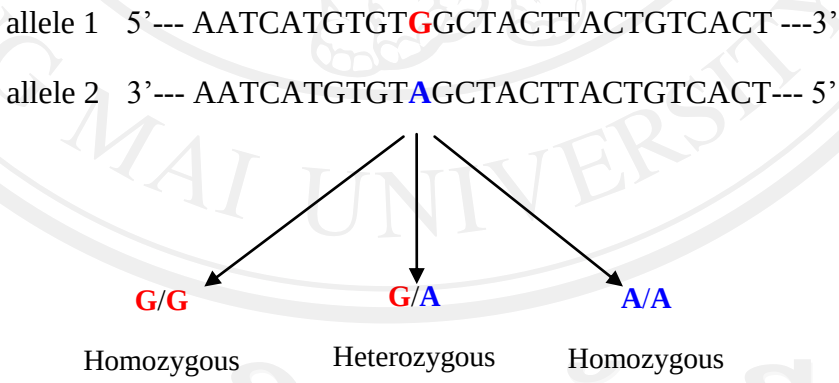

Figure 2.2 Single nucleotide polymorphisms in an individual

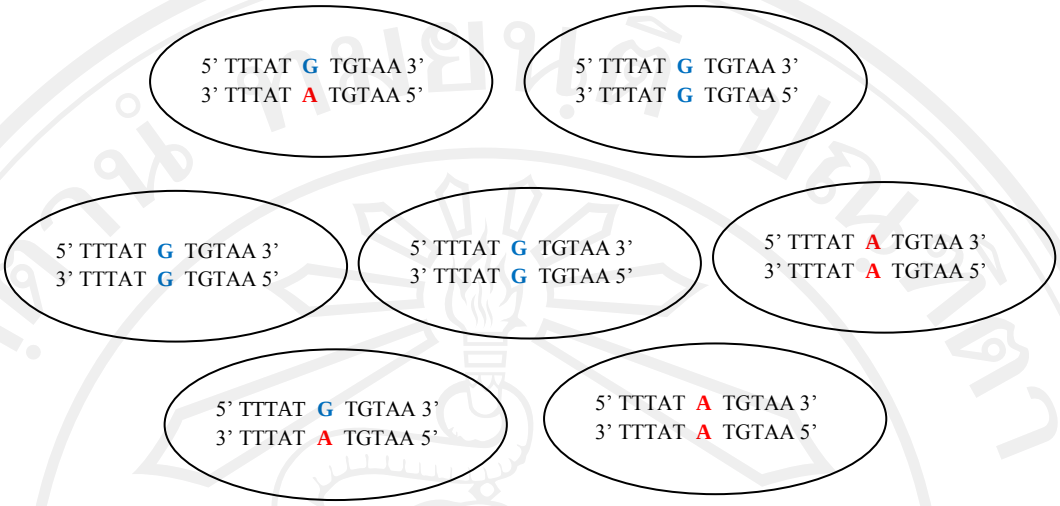

Figure 2.3 Single nucleotide polymorphisms in population

SNP is the simplest type of genetic variation which occurred in a single DNA base. There are about 10 millions known SNPs in the human genome. SNPs result from error copying that cells make while copying the DNA sequence during the process of cell division. Most SNPs is arisen many centuries ago and can be found to be shared between large groups of people with a common origin. If a SNP arises in a germ cell (egg cell or sperm cell) then it may be passed on to that individual's offspring and all future generations. In many cases, the impact of a SNP is not significant, but in other cases, it can alter the regulation of a gene or result in an altered gene product. SNPs can be synonymous or non-synonymous. A SNP is synonymous when the base change does not result in a change in the amino acid (and therefore has no functional effect). Conversely, a non-synonymous SNP is a base change that results in a change in the amino acid. An amino acid change may affect the protein product of that gene. There are SNPs that are associated with variation in physical traits between individuals, such as the shape of the nose or the color of hair.

Of greater interest to the biomedical community are SNPs that are associated with disease risk or predisposition to certain medical conditions.

2.3.1 SNP discovery

Although numerous approaches for SNP discovery have been described, including some also currently used for genotyping, the main ones are based on the comparison of locus-specific sequences, generated from different chromosomes. The simplest, when targeting a defined region, for instance containing candidate genes, is to perform direct sequencing of genomic PCR products obtained in different individuals. However, on a large scale, this approach tends to be costly due to the need for locus-specific primers. This approach is limited to regions for which sequence data is available and produces a diploid sequence in which it is not always easy to distinguish between sequencing artifacts and polymorphism when double peaks, as expected in heterozygotes. Therefore, different approaches based on the comparison of sequences obtained from cloned fragments can be considered for developing an SNP map of a genome. For SNP genotyping, there are many techniques available. One key feature of most SNP genotyping techniques, apart from those based on direct hybridization, is the two step separation: 1) generation of allele-specific molecular reaction products; 2) separation and detection of the allele specific products for their identification (Vignal *et al*., 2002). SNP-microarrays provide a relatively fast and inexpensive genotyping method, which requires only tiny amounts of DNA. Different kinds of technologies are used in SNP-microarray production and platforms varied in their hybridization and staining techniques. Commercial arrays are marketed with brands such as: Asper, TaqMan, Sequenom, Illumina, and

Affymetrix (Laakso, 2007). Considerable progress has been made in the technological ability to assay humans for genetic variation. Commercial probe-based SNP array platforms can now genotype, with >99% accuracy, about one million SNPs in an individual in one assay (Affymetrix, 2006 ; Illumina, 2009).

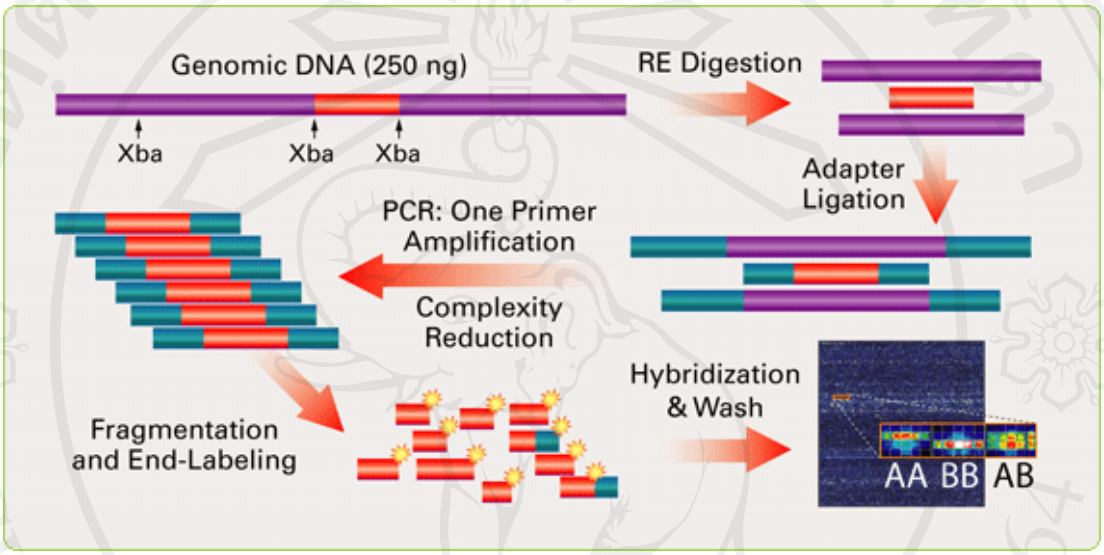

Figure 2.4 Sample preparation and array processing of genomic DNA.

Figure 2.4 shows the sample that is digested with a restriction enzyme and ligated to adaptors that recognized as the cohesive four basepair (bp) overhangs. All fragments resulting from restriction enzyme digestion, regardless of size, are substrates for adaptor ligation. A generic primer that recognizes the adaptor sequence is used to amplify adaptor ligated DNA fragments. PCR conditions have been optimized to preferentially amplify fragments in the 250 to 2,000 bp size range. The amplified DNA is then fragmented, labeled, and hybridized.

In this thesis, the Affymetrix SNP array was used to create a data sets. The Affymetrix SNP arrays work using chemistries that rely on the biochemical principle (Figure 2.4). The nucleotide bases bind to their complementary partners specifically,

A binds to T and C binds to G, in Watson–Crick base pairs. The array protocols call for the hybridization of fragmented single-stranded DNA to arrays containing hundreds of thousands of unique nucleotide probe sequences. Each probe is designed to bind to a target DNA subsequence. A specific hypothetical example for one SNP is shown in Figure 2.5. In the cases, specialized equipment can produce a measure of the signal intensity associated with each probe and its target after hybridization. The underlying principle is that the signal intensity depends upon the amount of target DNA in the sample, as well as the affinity between target and probe. Extensive processing and analysis of these raw intensity measures yield SNP genotype inferences. The manufacturers report genotyping accuracy well over 99.5%. This section details some of the computational algorithms that have been developed to convert the set of probe intensities into genotypes (LaFramboise, 2009).

In Figure 2.5 at the top, is the fragment of DNA harboring an A/C SNP to be interrogated by the probes. In the Affymetrix assay, there are 25-mer probes for both alleles, and the location of the SNP locus varies from probe to probe. The DNA binds to both probes regardless of the allele it carries, but it does so more efficiently when it is complementary to all 25 bases (yellow) rather than mismatching the SNP site (red). This impeded binding manifests itself in a dimmer signal. For the platforms, the computational algorithms convert the raw signals into inferences regarding the presence or absence of each of the two alleles

When the analysis is completed; a report of summarizing data from the samples will be displayed. An important indicator that must be checked for each array is the call rate (SNPCall). Values > 95% represent an acceptable data accuracy.

Another critical factor, the signal detection values, should be above 99%. The Affymetrix Raw Data Files; Flat Files and CEL Files were obtained (Table 2.1).

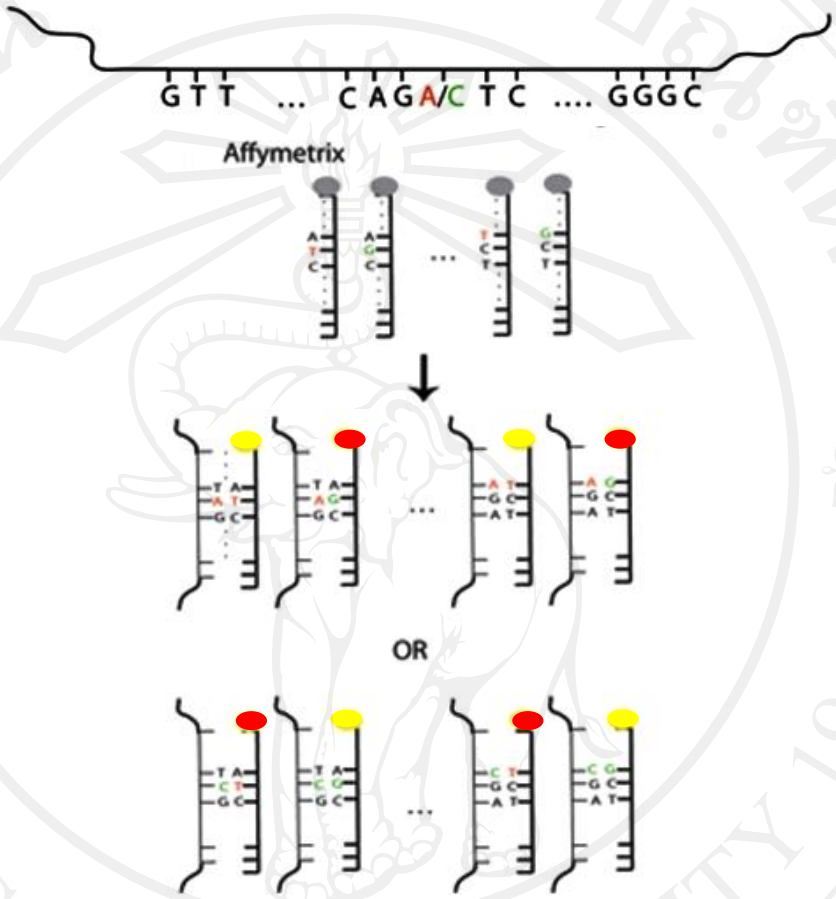

Figure 2.5 Overview of Affymetrix SNP array technology.

2.3.2 Pattern of human SNP variation

Humans arose about 100,000–200,000 years ago in Africa, and spread from there to the rest of the world (Tishkoff *et al*., 1996). The original population was polymorphic, and so populations around the world share most polymorphisms from our common ancestors. For example, all populations are variable at the gene for the ABO blood group. About 85–90% of human variation is within all populations

(Barbujani *et al*., 1997 cited in Kwok, 2003). Thus any two random people from one population are almost as different from each other as being any two random people from the world. Mutations have arisen in populations since humans spread around the world, so some variation is mostly within particular populations. Rare variants are likely to have arisen recently, and are more likely than common variants to be found in some populations but not others (Rieder *et al*., 1999 and Nickerson *et al*., 1998). Common variants are usually common in all populations. Only a small proportion of variants are common in one population and rare in another.

Table 2.1 Example of summary data from Affymetrix raw file

		Mapping Array Report					
		Report File Name - E:\Program Files\Affymetrix\GeneChip\Affy_Data\Data\TH-HM-000106-1-01.RPT					
		Date:	03/21/06 19:35:31				
		Total number of SNPs:	58960				
		Total number of QC Probes	4				
		Probe array type:	Mapping50K_Xba240				
		SNP Performance					
		CEL Data	Called Ger	SNP Call	AA Call	AB Call	BB Call
1	1	TH-HM-000106-1-01	F	98.47%	38.51%	23.29%	38.20%
2	2	TH-HM-000109-1-01	M	98.61%	37.90%	24.64%	37.46%
3	3	TH-HM-000110-1-01	F	97.80%	37.58%	24.81%	37.61%
4	4	TH-HM-000111-1-01	M	98.27%	37.36%	25.41%	37.23%
5	5	TH-HM-000120-1-01	F	99.06%	37.34%	25.91%	36.75%
6	6	TH-HM-000122-1-01	F	98.04%	37.50%	25.24%	37.26%
7	7	TH-HM-000136-1-01	F	94.76%	38.87%	23.18%	37.95%
8	8	TH-HM-000138-1-01	F	97.22%	37.99%	24.68%	37.33%
9	9	TH-HM-000144-1-01	F	97.68%	37.49%	25.32%	37.19%
10	10	TH-HM-000147-1-01	M	99.17%	36.95%	25.91%	37.14%

Figure 2.6 shows the pattern of human variation. The outer circle is the entire amount of human variation, and each of the other circles shows the variation within one population**.** The large overlap among the circles shows that all populations contain mostly the same variation. The small nonoverlap regions are still important for population differences in susceptibility to disease, but even then not all people in a

population get any particular disease. Most differences in disease risk are among individuals regardless of population, rather than among populations. (Kwok, 2003)

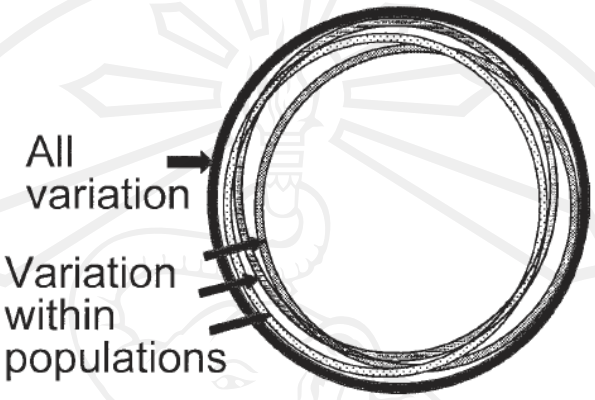

Figure 2.6 Distribution of human variation within and between populations

2.4 Variable ranking and feature selection

2.4.1 Introduction to variable and feature Selection

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. The selections are the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Feature selection serves two main purposes. First, it makes the training and applying a classifier more efficient by reducing the measurement and storage requirements, and reducing training and utilization times. Second, feature selection often increases classification accuracy by eliminating of noise features. A noise feature is one that, when added to the document representation, increases the classification error on new data (Christopher *et al*., 2008).

2.4.2 Variable ranking

Many variable selection algorithms include variable ranking as a principal or auxiliary selection mechanism because of its simplicity, scalability, and good empirical success. Several papers in this issue use variable ranking as a baseline method (Bekkerman *et al*., 2003; Caruana and de Sa, 2003; Forman, 2003; Weston *et al*., 2003). Variable ranking is not necessarily used to build predictors. In this section consider ranking criteria that defined for individual variables being independently of the context of others. Correlation methods belong to that category.

For the principle of the method, let consider a set of m examples $\{x_k, y_k\}$ $(k = 1, \dots m)$ consisting of n input variables x_k, i $(i = 1, \dots n)$ and one output variable y_k. Variable ranking makes use of a scoring function $S(i)$ computed from the values x_k, i and y_k, k, $k = 1, \dots m$. By convention, it is assumed that a high score indicating a valuable of the variable and sort variables in decreasing order of $S(i)$. To use variable ranking to build predictors, nested subsets progressively incorporating more and more variables of decreasing relevance are defined (Guyon and Elisseeeff, 2003).

2.4.3 Information theoretic ranking criteria

Several approaches to the variable selection problem using information theoretic criteria have been proposed (Bekkerman *et al*., 2003, Dhillon *et al*., 2003, Forman, 2003, Torkkola, 2003). Many rely on empirical estimates of the mutual information between each variable and the target:

$$I\,(i) = \int_{x_i} \int_{y}\ p(\,x_i\,, y)\, log\, \frac{p(x_i\,, y)}{p(x_i)p(y)}\, dxdy \qquad (2.1)$$

Where $p(x_i)$ and $p(y)$ are the probability densities of x_i and y ,and $p(x_i), y$ is the joint density. The criterion $I(i)$ is a measure of dependency between the density of variable x_i and the density of the target y_i. The difficulty is that the densities $p(x_i)$, $p(y)$ and $p(x_i, y)$ are all unknown and hard to estimate from data. The case of discrete or nominal variables is probably easiest because the integral becomes a sum:

$$I(i) = \sum_{x_i} \sum_{y} P(X = x_i, Y = y) \, log \frac{P(X = x_i, Y = y)}{P(X = x_i) P(Y = y)} \qquad (2.2)$$

The probabilities are then estimated from frequency counts. For example, in a three-class problem, if a variable takes 4 values, $P(Y = y)$ represents the class prior probabilities (3 frequency counts), $P(X = x_i)$ represents the distribution of the input variable (4 frequency counts), and $P(X = x_i, Y = y)$ is the probability of the joint observations (12 frequency counts). The estimation obviously becomes harder with larger numbers of classes and variable values. The case of continuous variables (and possibly continuous targets) is the hardest. One can consider discretizing the variables or approximating their densities with a non-parametric method such as Parzen windows (see, e.g., Torkkola, 2003). Using the normal distribution to estimate densities would bring us back to estimating of the covariance between X_i and Y, thus giving us a criterion similar to a correlation coefficient.

2.5 Data classification and decision tree

2.5.1 Basic knowledge in data classification

Classification is the process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to specific variables or class labels (for predicting) is known as classification (Thearling, 2010). For instance, determining whether a protein binds to DNA or not based on sequence and structural motifs, cancer type classification based on micro array expression are good examples of classification problems.

Data classification is a scientific discipline researching for the ways of assigning class membership values (labels) to unknown (in the sense that they have not previously be seen to an observer) observations or samples, based on a set of observations or samples provided with class membership values (labels). Each observation is represented by a feature vector associated with it. Unknown observations form a test set while their labeled counterparts together with class labels compose a training set. Labeling unknown observations is done by means of a classifier, which is a data classification algorithm implementing a mapping of feature vectors to class labels. An algorithm is a sequence of steps necessary for the solution of a data classification task at hand. Let us restrict ourselves to two-class problems. Thus, the process of data classification involves either one or two following steps: 1) Training a classifier (optional) and 2) Testing the trained classifier (Figure 2.7).

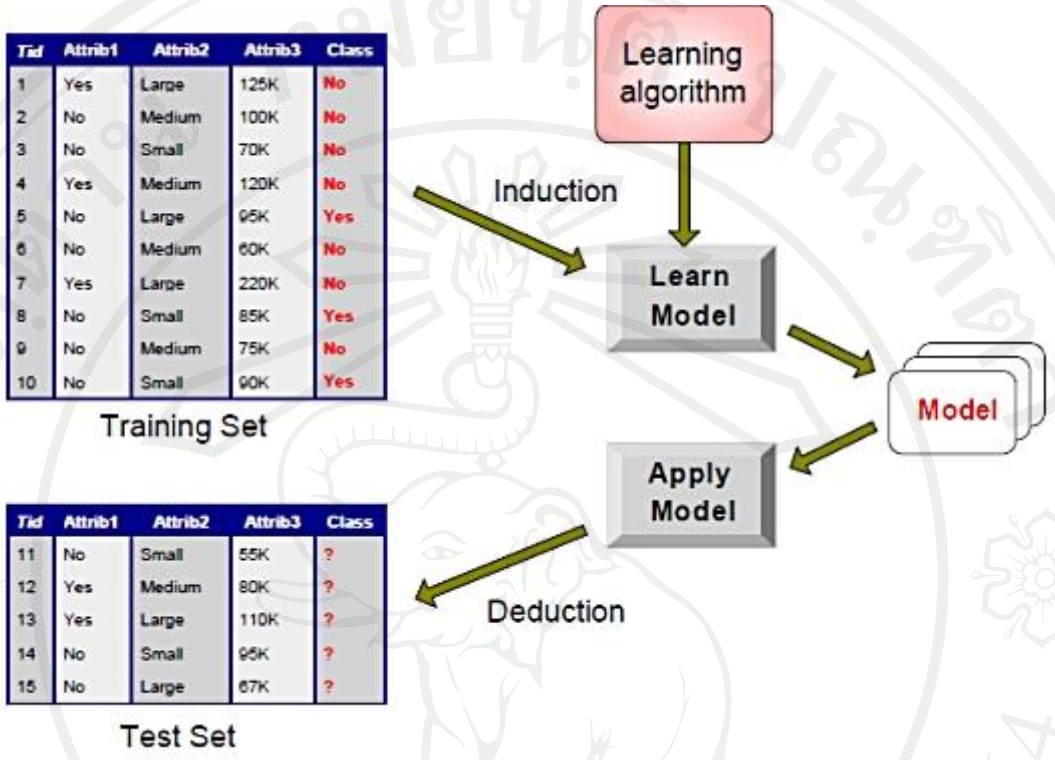

Figure 2.7 Illustrate classification task

Many classifiers have one or several parameters to be pre-defined before classification will start. Without knowing the optimal values of these parameters, data classification would be akin to random walk in search for the right solution. The words 'optimal values' mean such parameter values that allow a classifier to learn the correct mapping from features, describing each observation, to class labels. This learning done from the training data is possible, because the learner can always check the answer: for this, it needs to compare its output and the correct result as specified by class labels assigned to the observations from the training set. If there is a mismatch (classification error) between the two, the learner knows there should be some work to do (Okun, 2011).

Typically, classification error on the training data is not precisely 0% and sometimes it simply cannot be so due to the finite size of the training set. On the contrary, the zero error rates may indicate that you over-trained the classifier so that it learned every minute detail, which is often nothing but noise. Such a classifier will be unable to generalize properly. In other words, when presented with previously unseen data, its classification performance will be very bad. The smaller the training set is, the higher your chances to over-train a classifier, because the different classes are likely to be under-represented. The more sophisticated classifier is, the higher chances are for its over-training, since sophisticated classifiers are capable of partitioning data classes in more complex ways than simpler classifiers. So, the training could be both evil and blessing. Sometimes, a classifier does not need training at all, which, however, does not automatically imply that this classifier will do its job well in all cases. Independently of the fact whether classifier training is required or not, the testing phase has still to be carried out to complete the data classification task. This is done with the trained classifier applied to the test data. If no training was needed, then the word 'trained' is omitted before 'classifier' in the last sentence. As a result of testing, test error and other performance characteristics such Area Under the Receiver Operating Characteristic (ROC) Curve are computed and which can be further compared (by means of statistical tests) with errors/characteristics of other classifiers attained on the same test set. Given that microarray gene expression that data are high dimensional, it is advisable and even required to reduce the number of features prior to data classification in order to alleviate the effect of classifier over-training. That is, dimensionality reduction should always precede classification when dealing with gene expression data sets

(Okun, 2011). There are many different classifiers that can be widely applied in bioinformatics such as decision trees, support-vector machine based classifier, Bayesian classifiers, and neural network classifiers, etc. These classifiers were also recently named among the top 10 algorithms in data mining. In addition, they form a bulk of base classifiers used in building classifier ensembles (Wu *et al*., 2008).

2.5.2 Principle of decision tree

A decision tree is a predictive model which can be used to represent both classifier and regression models. On the other hand, decision trees refer to a hierarchical model of decision and their result. When a decision tree is used for classification tasks, it is more suitably referred to as a classification tree (Rokach and Maimon, 2008).

A decision tree is a simple tree-shaped structure where each internal node represents a test on one attribute, the arcs show the results of a test and the leaf nodes reflect classes. They are easy to understand and can be easily converted to a set of rules. Moreover, they can classify both categorical and numerical data and require no priori assumptions of the data. Because of the advantages listed above, the decision tree approach is extensively utilized for both classification and prediction purposes (BarKir *et al*., 2006).

The main objectives of decision tree classifiers are (Safavian and Landgrebe, 1991):

1) To classify correctly as much of the training sample as possible.

2) Generalize beyond the training sample so that unseen samples could be classified with as high of accuracy as possible

3) Easy to update as more training sample becomes available and

4) Have as simple a structure as possible.

The method objective is to select the attribute which most useful for classifying the examples. In order to measure the worth of an attribute, a statistical property is defined and information gain, which measures how well a given attribute, separates the training examples according to their target classification.

Decision tree classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified b starting at the root node of the tree, testing the attribute specified b this node, then moving down to the tree branch corresponding to the value of the attribute in the given example. This process is then repeated for the subtree rooted at the new node Figure 2.8 illustrates a typical learned decision tree.

2.5.3 Evaluate measurement using confusion matrix

The confusion matrix is used as an indication of the properties of a classification rule. It contains the number of elements that have been correctly or incorrectly classified for each class. Seeing on its main diagonal the number of observations that have been correctly classified for each class; the off-diagonal elements indicate the number of observations that have been incorrectly classified.

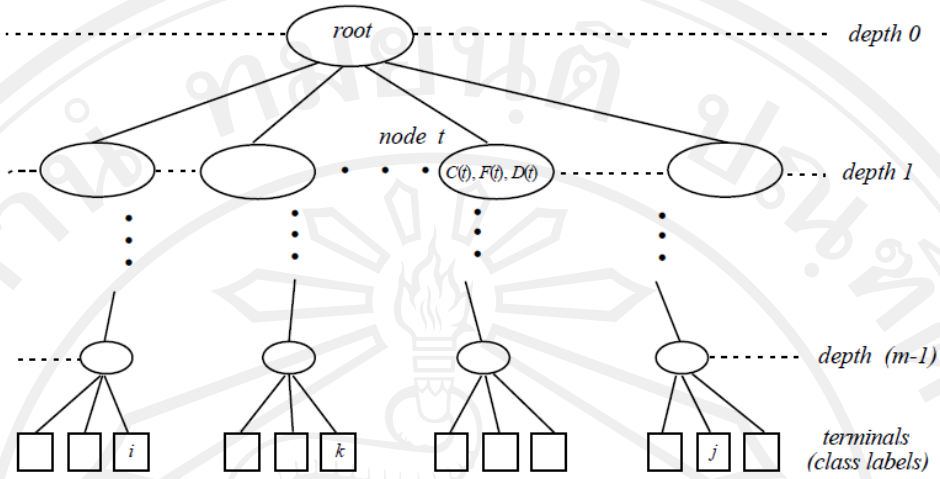

Figure 2.8 Simple decision tree *C(t)* – subset of classes accessible from node *t*, *F(t)* – feature subset used at node *t*, *D(t)*- decision tree rule base at node *t*

One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). For every instance in the test set, we compare the actual class to the class that was assigned by the trained classifier. A positive (negative) example that is correctly classified by the classifier is called a true positive (true negative); a positive (negative) example that is incorrectly classified is called a false negative (false positive) (Rokach and Maimon, 2008). These numbers can be organized in a confusion matrix shown in Table 2.2.

Table 2.2 Confusion matrix example

	Predicted negative	Predicted positive
Negative Examples	A	B
Positive Examples	C	D

Based on the values in Table 2.2, one can calculate all the measures defined above:

- Accuracy is: *(a+d)/(a+b+c+d)*
- Misclassification rate is: *(b+c)/(a+b+c+d)*
- Precision is: $d/(b + d)$
- True positive rate (Recall) is: $d/(c + d)$
- False positive rate is: $b/(a + b)$
- True negative rate (Specificity) is: $a/(a + b)$
- False negative rate is: $c/(c + d)$

2.6 Categorical data and correspondence analysis

2.6.1 Basic concept in categorical data analysis

Let's first define categorical data. Categorical variables represent the types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level. While the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of groups.

Analysis of categorical data generally involves the use of data tables. A two-way table (Table 2.3) presents categorical data by counting the number of observations that fall into each group of two variables: one divided into rows and the other divided into columns. For example, suppose a survey was conducted of a group of 20 individuals who were asked to identify their hair and eye color, a two-way table presenting the results might appear as follows (Ender, 2005):

Often the outcomes of interest are frequency counts of observations occurring in specified response categories rather than continuous variables. But the outcomes are expressed as frequency counts of observations occurring in the response categories.

Table 2.3 Example of a two-way table

Hair color	Eye Color				
	Blue	Green	Brown	Black	Total
Blonde	2	1	2	1	6
Red	1	1	2	0	4
Brown	1	0	4	2	7
Black	1	0	2	0	3
Total	5	2	10	3	20

2.6.2 Principle of correspondence analysis

Correspondence analysis represents yet one more method for analyzing data in contingency tables. Correspondence analysis is a statistical visualization method for picturing the associations between the levels of a two-way contingency table. Correspondence analysis is a descriptive/exploratory technique designed to analyze two-way and multi-way tables containing measures of correspondence between the row and column variables. The results produced by correspondence analysis provide information which is similar to that produced by principal components or factor analysis. They allow one to explore the structure of the categorical variables included in the table. Correspondence analysis seeks to represent the relationships among the categories of row and column variables with a smaller number of latent dimensions.

It produces a graphical representation of the relationships between the row and column categories in the same space (Lee, 2007).

In a two-way contingency table, the observed association of two traits is summarized by the cell frequencies, and a typical inferential aspect is the study of whether certain levels of one characteristic are associated with some levels of another. Correspondence analysis is a geometric technique for displaying the rows and columns of a two-way contingency table as points in a low-dimensional space, such that the positions of the row and column points are consistent with their associations in the table. The goal is to have a global view of the data that is useful for interpretation. To illustrate the correspondence analysis is to consider the multidimensional time series on the number of science doctorates conferred in the USA from 1960 to 1975 that is shown in Table 2.4 (Greenacre, 1984). Correspondence analysis of these data yields the graphical display shown in Figure 2.9. It has two sets of points, as indicated by the two types of point symbols. The points are row points for the rows of the data and column points for the columns. In Figure 2.9, there are row points for the disciplines, and column points for the years.

The distance between the row points is a measure of similarity between the row-frequency profiles. The anthropology degree and the engineering degree are far from each other because their profiles are different, whereas the mathematics degree is near the engineering degree because their profiles are similar. The distances between the points representing years are interpreted in the same way of each year point representing the profile of that year across the various disciplines.

Table 2.4 Table of Science Doctorate in USA, 1960-1975 (Greenacre, 1984)

Discipline/Year	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1263	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	134	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Articulture	414	576	830	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

Note that the positions of two sets of points with respect to each other are not directly comparable and should be interpreted with caution. The interpretation given by Greenacre (1984) for this example is that each discipline point will lie in the neighborhood of the year of which the discipline's profile is prominent. Thus, there are relatively more agriculture, earth science and chemistry degrees in 1960, while the trend from 1965 to 1975 appears to be away from the physical sciences towards the social sciences. Furthermore, noticin points such as earth sciences and economics lie within the parabolic configuration of the years points; this implies that the profiles of these disciplines are higher than average in the early and later years. This example

illustrates how a low-dimensional graphical representation of what is basically a deterministic trend supports a rich description of the data.

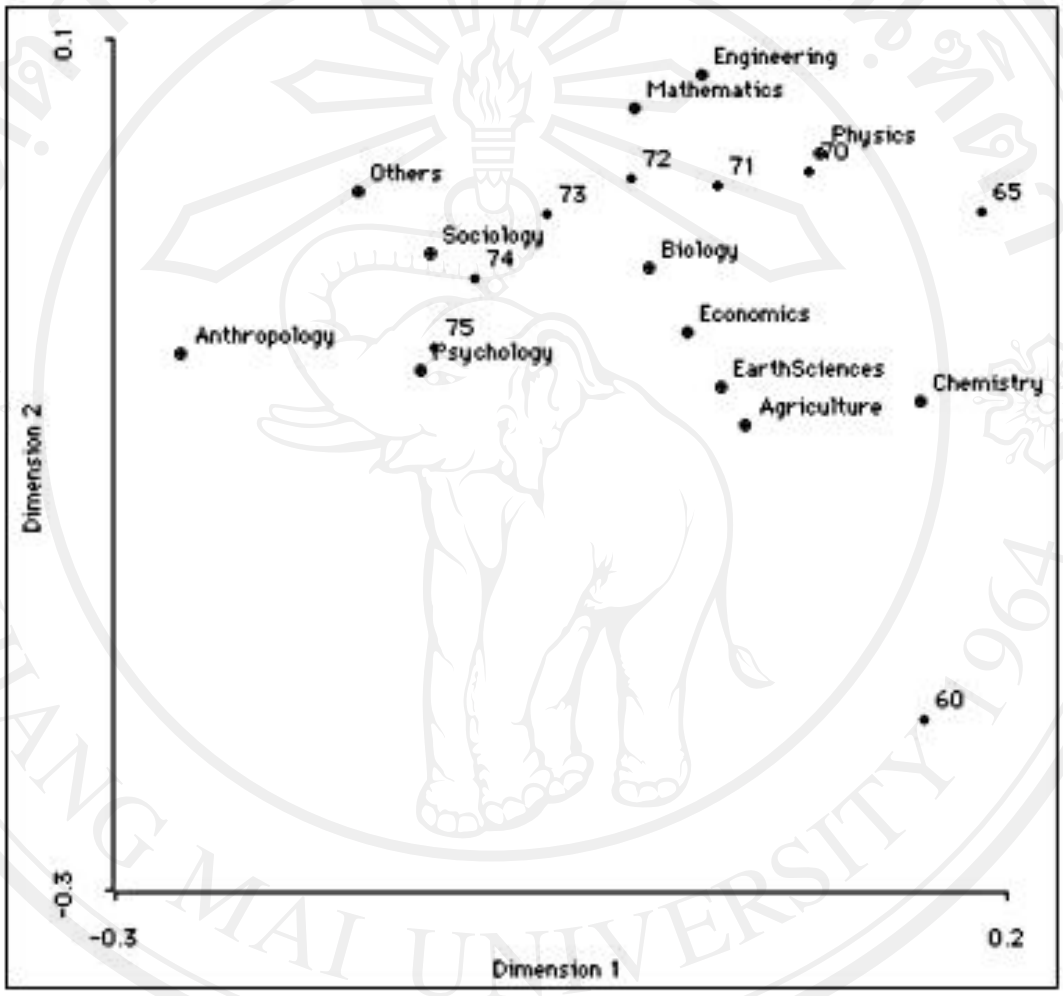

Figure 2.9 Correspondence analysis of doctorate data

2.7 Population genetic distance

Population that is reproductively isolated, normally they are separated geographically, gradually become different in genetic. The principal reason for the differences is that the selection force are different. Also, if a portion of a population moves to a different territory, or becomes isolated from the rest of the population due

to waters rising, rivers shifting, glaciers and deserts forming, or other reasons and, if some of those isolated people just happen to be a little genetically different from the remainder of the population, which is probable, the entire isolated population is likely to become even more genetically different, which is called the "Founder Effect." Chance mutations may also arise in one population that do not arise in another population, or only one of the populations may interbreed with a third population (Fuerle, 2008).

2.7.1 Principle of genetic distance

Genetic distance is a value tell that the number of gene differentiate inside each locus after the two population began to separate, or that, the measure of the allelic substitutions per locus that have occurred during the separate evolution of two populations or species. In the situation of linked gene, the distance between them usually consider in terms of recombination frequency or map units: as explain that, everyone has the same genes, e.g., we all have a gene for eye color, but each gene comes in an average of 14 different A-C-G-T sequences, called "alleles".

To determine the genetic distance between two individuals, the number of alleles that differ between them have to be counted (Sailer, 2006), for populations, the number of people in each population who have a particular allele is counted (preferably using a large number of alleles to increase precision), and the results are expressed mathematically.

There are many software which suitable to calculate population to population genetic distance from allele frequency such as Arlequin v3.01, PHYLIP, GDA, PopGene, Populations, and SGS. All of them estimate genetic distance using one of

three methods: Nei's, Cavlli-Sforza's or Reynold's (Cavalli-Sforza and Edwards, 1967; Nei, 1983; Nei, 1996). After those methods, genetic distance matrix data were obtained then choose the program estimates phylogenies. A tree can be obtain using one of the following components of the program also draws a phylogenetic tree using the genetic distance matrix data. It uses either Nei and Saitou's (1987) Neighbor Joining (NJ) Method, or the UPGMA (unweighted pair group method with arithmetic mean; average linkage clustering) method (Sneath and Sokal, 1973).

Because of different assumptions they are based on the NJ and UPGMA methods may construct dendrograms with totally different topologies. Both methods use distance matrices. The principle difference between NJ and UPGMA is that NJ does not assume an equal evolutionary rate for each lineage. Since the constant rate of evolution does not hold for human populations, NJ seems to be the better method. For the genetic loci subject to natural selection, the evolutionary rate is not the same for each population and therefore UPGMA should be avoided for the analysis of such loci.

To find the root of the tree, one can add an outgroup. The point in the tree where the edge to the outgroup joins is the best possible estimate for the root position. One persistent problem with tree construction is the lack of statistical assessment of the phylogenetic tree presented. This is best done with widely available bootstrap analysis originally described. For a discussion of statistical tests of molecular phylogenies, topology to be statistically significant the bootstrap value for each cluster should reach at least 70% whereas 50% overestimates accuracy of the tree. Bootstrap tests should be done with at least 1000 (preferably more) replications. (Felsenstein, 1985; Efron *et al*, 1996)

Nei noted that some genes are more suitable than others in phylogenetic inference and that most tree-building methods tend to produce the same topology whether the topology is correct or not (Nei, 1996). He also added that sometimes adding one more species/ population would change the whole tree for unknown reasons. An example of this has been provided in a study of human populations with genetic distances (Nei and Roychoudhury, 1993). The properties of most popular genetic distance measures have been reviewed (Kalinowski, 2002). Whichever is used, large sample sizes are required when populations are relatively genetically similar, and loci with more alleles produce better estimates of genetic distance. However, in a simulation study, Nei et al concluded that more than 30 loci should be used for making phylogenetic trees (Nei, 1983). There seems to be a consensus that estimated tress are nearly always erroneous (i.e., the topological arrangement will be wrong) if the number of loci is less than 30 (Nei, 1996). If populations are closely related even 100 loci may be necessary for an accurate estimation of the relationships by genetic distance methods. Cavalli-Sforza has noted important correlations between the genetic trees and linguistics evolutionary trees with the exceptions for New Guinea, Australia and South America (Cavalli-Sforza, 1994 cited in Dorak, 2011).

2.7.2 Estimating genetic distance

The calculation of a genetic distance between two populations gives a relative estimate of the time that has passed since the populations have existed as single cohesive units. However, estimations of distance may also be present because the populations are completely isolated but have only been separated for a short period of time. When two populations are genetically isolated, the two processes of mutation

and genetic drift lead to differentiation in the allele frequencies at selectively neutral loci. As the amount of time that two populations are separated increases, the difference in allele frequencies should also increase until each population is completely fixed for separate alleles. A number of methods have been developed which estimate genetic distance from these allele frequency differences. Many researchers are interested in the genetic differentiation of populations, there are some genetic distances are used to describe the genetic difference between populations, such as F_{ST}, Nei's standard distance, Nei's *DA* distance, F^*_{ST} distance and Cavalli-Sforza's *dC* distance (Xu and Jin, 2011).

2.8 Related work

Previous studies which related to this thesis are divided into two sections. Firstly, the studies were population genetic of the northern Thailand population, and then used SNPs as the genetic marker.

2.8.1 Genetic study of the northern Thailand population

First, the study of genetic structure and genetic affinity of Tai ethnic group in the Northern Thailand aimed to use the mitochondrial DNA to trace back biological ancestors, as well as to reconstruct the past history of four northern Thai populations-the Yuan, Lue, Yong, and Khuen. Their genetic structures and relationships with the nearby populations in northern Thailand and southern China were also investigated. The results suggested that the genetic structure of each village generated by the founder effect. Genetic clustering of Tai, Mon-Khmer, and Tibeto-Burman groups, as revealed by mtDNA HVRI sequences, did not correlate well with linguistic

classification. The Yuan or Khon Muang in northern Thailand exhibited a close relationship to the Tai group in South and Southeast China, which supported the immigrant hypothesis of their ancestors. However, more genetic data of the endogenous Mon-Khmer ethnic groups is needed to reconstruct a clear history of the Tai, as well as their biological ancestor (Kampuansai *et al*., 2007).

Next, the study in the genetic origin of the Hill Tribes used a clustering analysis, and two specific databases for mtDNA and Y-chromosome assembled from several published that studies on Asian populations. Overall the sex specific genetic signature of different post marital habits of residence in the Hill Tribes was robust. However, specific perturbations related to linguistic differences, population specific traits, and the complex migratory history of these groups, could be identified (Besaggio *et al*., 2007).

Finally, the male and female gene pool of geographically diverse Khon Mueang along the historical Yuan migratory route were analyzed to investigate the genetic diversity and genetic structure , evaluate if any geographic or demographic factors have shaped pattern of variation, and examine the degree of concordance between mtDNA and Y-chromosome. The result of analysis of this study provided insights into the genetic structure of present day Khon Mueang. Their data substantiate that the genetic structure of the Khon Mueang paternal lineage differs from that of the maternal lineage. These could have been caused by sex-bias demographic processes such as migration and admixture. Factors of geography and extensive gene flow among populations play an important role in shaping genetic differentiation of geographically diverse Khon Mueang groups, especially in the Chiang Mai-Lamphun basin (Kutanan *et. al*., 2010).

2.8.2 Population genetic using SNPs marker

The first, Park *et al*. (2006) construct SNP@Ethnos, which is a catalog of human SNPs and genes that contain human ethnic variation. The database contains following results for detecting natural selection and population difference. They compared the genotype profiles of three ethnic groups: Yoruba in Ibadan, Nigeria (YRI) and combination of Japanese in Tokyo (JPT) and Han Chinese (CHB) in Beijing (CHB+JIP), and Utah residents with ancestry form northern and western Europe (CEU). The study identified 100736 SNPs that could classify the ethnic groups based on the nearest shrunken centroid method (NSCM). Of those SNPs, 5515 were in well known loci of natural selection and disease-associated genes. Using the Sorting Intolerant From Tolerant system, 85 coding nonsynonymous ethnically variant SNPs (ESNPs) were predicted as damaging, showing that these SNPs may be highly relevant in disease research. The result of these SNPs, 1009 were within disease-associated genes, and 85 were predicted as damaging, using the SNPs. This study resulted in the creation of the SNP@Ethnos database that contains genetic-variation information for use in human differentiation studies.

Population classification uses SNPs data in the Hapmap Project (Zhou and Wang, 2008). They try to find as few SNPs as possible from the original nearly 4 million SNPs to classify the 3 populations in the Hapmap genotype data. A modified t-test was proposed to use in SNPs ranking, where the higher ranking value, the stronger corresponding classification power. Then, from the ranking value, they randomly choose different numbers of top ranked SNPs, e.g. 2, 5, 7, 10 and so on, combine the ranking result with the support vector machine to find the best SNP subset. The SNP subset, which has the smaller size and highest classification

accuracy, was determined as best subset. In the result, this obtained that perfect classification using only 11 SNPs.

The HUGO Pan-Asian SNP consortium was the study aimed to study a genetic patterns of more than 70 Asian populations. The Affymetrix GeneChip Human Mapping 50K Xba Array was used to genotype 1,928 individuals from 73 Asian populations at 54,794 autosomal SNPs. They found that most genetic cluster corresponded to language groups, though geography was also a factor in these patterns. In addition, the study suggests an influx of individuals from Southeast Asia contributed genetically to many populations found in East Asia today. Their ancestry analyses suggest Asian populations harbor genetic contributions from five language groups, three ethnic groups, and two small groups representing specific populations in Borneo and Thailand. Most of the genetic patterns corresponded with language groups, the researchers reported. For instance, they found eight populations in which genetic and language patterns did not match. Moreover, individuals in these populations tended to cluster more closely with nearby geographic populations. In general, haplotype diversity was highest in southern Asia and dwindled in samples taken further north. Most East Asian haplotypes (some 90 percent) turned up in Southeast or Central-South Asia. Nevertheless more of these haplotypes were unique to Southeast Asia: about half of East Asian haplotypes were present only in Southeast Asia, the researchers reported, compared with the five percent of East Asian haplotypes that were found in Central-South Asia alone. Such patterns indicate that migration from Southeast Asia into East and North Asia, the team explained (The HUGO Pan-Asian SNP Consortium *et al*., 2009).

The selection of SNP indentifying ancestry discusses how to select an optimal set of SNPs, from the millions of known SNPs, which should be genotyped to best predict ancestry. Their SNP selection procedure is designed to use the database collected by the Human Genome Diversity Project, which includes genotypes for 500,000+ SNPs from 100's of individuals, spanning 54 populations. They note two unique features of our procedure that can greatly improve prediction accuracy. First, they incorporate the phylogenetic tree among the populations when estimating the allele frequencies in each population. Second, they develop a better estimate for the error rate. They demonstrate the increased accuracy gained by these improvements using both simulated and HGDP data. They also provide a list of the 100 optimal SNPs for identifying ancestry (Sampson *et al*., 2011).