CHAPTER 3

RESEARCH METHODOLOGY

3.1 Studied Populations

In this research thirteen ethnic populations were studied. These populations were divided into three groups, based on linguistic and ethnohistorical data.

1) The Tai groups including the Yuan, Yong, Lue and Khuen. Their language belongs to the Tai-Kadai linguistic family.

2) Indigenous group comprises the Mon, Lawa, H'tin, Plang, Paluang and Mlabri. Their language belongs to the Mon-Khmer linguistic subfamily of Austro Asiatic family.

3) The hill tribes including the Karen, Hmong, and Yao. The Karen uses the language in the Tibeto-Burman linguistic subfamily of Sino Tibetan family, while the Hmong and Yao use the language belongs to Hmong-Mien linguistic family.

The studied populations were 256 individuals (137 males and 119 females) from 13 ethnic groups in the upper northern part of Thailand. The details of various studied populations are shown in Table 3.1. Blood samples collection and DNA extraction were done by the population genetics research group, Department of Biology, Faculty of Science, Chiang Mai University.

Table 3.1 Detail description of studied populations

Ethnic group	ID Code	Linguistic affiliation (Family, Subfamily)	Location (District, Province)	Number of samples
Yuan	TU	Tai–Kadai, Tai	San Sai and Mae Tang, Chiang Mai ; Ban Hong, Lamphun ; Saw Hai, Saraburi	20
Yong	TY	Tai–Kadai, Tai	Pa Sang, Lamphun	20
Lue	TL	Tai–Kadai, Tai	Pua, Nan ; Doi Sa Ket, Chiang Mai	20
Khuen	TK	Tai–Kadai, Tai	Mae Wang and San Pa Tong, Chiang Mai	20
Lawa	LW	Austro-Asiatic, Mon-Khmer	Mae La Noi, Mae Hong Son	19
H'Tin	TN	Austro-Asiatic, Mon-Khmer	Tung Chang and Chiang Klang, Nan	20
Mlabri	MA	Austro-Asiatic, Mon-Khmer	Wiang Sa, Nan	19
Mon	MO	Austro-Asiatic, Mon-Khmer	Pa Sang, Lamphun	19
Plang	PP	Austro-Asiatic, Mon-Khmer	Mae Sai and Mae Chan, Chiang Rai	20
Paluang	PL	Austro-Asiatic, Mon-Khmer	Fang and Chiang Dao, Chiang Mai	20
Karen	KA	Sino-Tibetan, Tibeto-Burman	Mae Sariang, Mae Hong Son	20
Hmong	HM	Hmong–Mien, Hmong	Mae Rim , Chiang Mai	20
Yao	YA	Hmong–Mien, Yao	Mae Yao and Muang, Chiang Rai	19

3.2 SNP data

The SNPs genotyping, as a part of Pan-Asian SNP Initiative, was performed by Dr. Metawee Srikummool, using the Affymetrix GeneChip Human Mapping 50K Xba array. The SNP genotyping data of 256 unrelated individuals, from 13 populations, were obtained from the Pan-Asian SNP database (PanSNPdb). A set of 58,960 SNPs, the biallelic markers, which were generated, was used in this study. These SNPs are fairly evenly spaced across all of the autosomes and X chromosome. Table 3.2 illustrates the format of the Affymetrix export being a tab delimited text file consisting of rows of SNPs and their attributes. The first two lines are headers describing the file title and attribute names.

Table 3.2 Affymetrix SNP array export file example of the Hmong ethnic group

Dynamic Model Mapping Analysis

No. of SNP	SNP ID	Chromo-some	Physical Position	dbSNP RS ID	TSC ID	TH-HM-000106-1-01_Call	TH-HM-000106-1-01_Confidence	…
1	SNP_A-1650338	2	168550528	rs836702		BB	0.008301	…
2	SNP_A-1716667	19	40749462	rs725986	TSC58722	AA	0.000488	…
3	SNP_A-1712945	19	53411226	rs2009373	TSC47071	BB	0.000977	…
4	SNP_A-1653742	6	65265069	rs10494882		NoCall	0.480469	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	…
58957	SNP_A-1714915	13	71972490	rs9318082		BB	0.000488	…
58958	SNP_A-1655697	X	86473638	rs10521379		AA	0.007813	…
58959	SNP_A-1724002	8	85324682	rs977858	TSC291521	AA	0.001465	…
58960	SNP_A-1674163	12	66603891	rs1905444	TSC949662	BB	0.016113	…

The SNP summary result information from genotyping analysis is shown in Table 3.3. From the analysis, the SNP call (AA, AB, BB or Nocall) was obtained at any locus for each individual. In this thesis, the individual genotype pattern at any SNP locus is of interested.

Table 3.3 Example of SNP information from Affymetrix SNP array genotyping analysis

Attribute	Sample of attribute	Description
SNP ID	SNP_A-1716667	Affymetrix SNP ID.
Chromosome	19	The chromosome on which the SNP is located on the current genome version.
Physical Position	40749462	The nucleotide base position where the SNP is found. The genomic coordinates given are in relation to the current genome version and may shift as subsequent genome builds are released.
dbSNP RS ID	rs725986	The dbSNP at the National Center for Biotechnology Information (NCBI).
TSC ID	TSC58722	The SNP Consortium (TSC) ID that corresponds to this probe set.
Sample 1	AA	Genotype in an individual.
…	…	…
Sample 20	BB	Genotype in an individual.
Allele A	A	The alternative nucleotides at the SNP position that occur in the population and can be identified by the probe set. All the SNPs measured by the Affymetrix mapping arrays are biallelic.
Allele B	G	

3.3 Computational Method

The computational method comprises four main sections: mutual information, decision tree, correspondence analysis, and genetic distance and phylogenetic tree analysis. The framework of thesis illustrates in Figure 3.1.

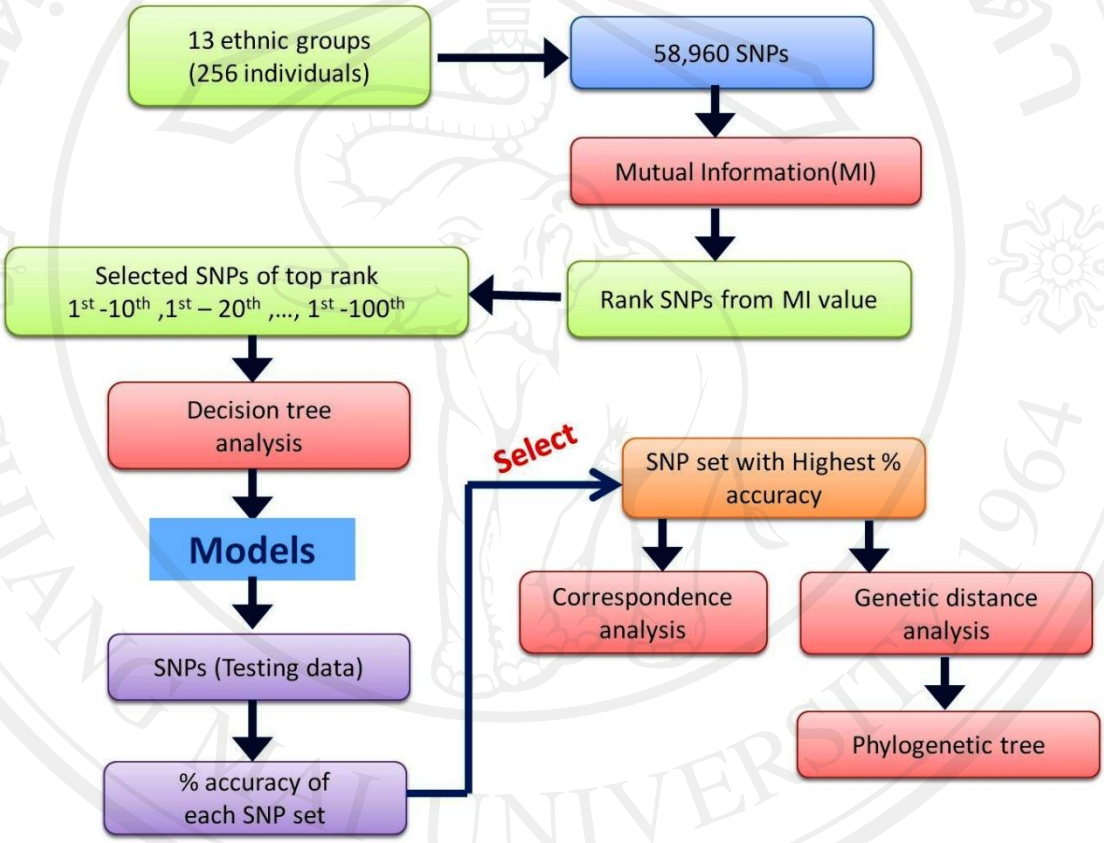


Figure 3.1 Diagram of research design

3.3.1 Mutual information with SNP

Due to the large amount of available SNPs loci, the SNPs which are specific in each population group were extracted. Input features are importance to improve the efficiency of feature selection. Therefore, the feature ranking is used in our

experiment to determine each feature's population classification power. In this work, mutual information estimation is used as the ranking measure.

According to ranking criteria of information theory, this theory was applied to SNPs ranking with the equation (2.2) in CHAPTER 2. The feature selection method is to compute $I(S;E)$ as the expected mutual information of any SNP locus (S) and class (E). Mutual inoformation measures how much information present or absent of a term contributes to make the correct classification decision on E. Formally (Christopher *et al.*, 2008):

$$I(S;E) = -\sum_{e_a \in \{1,0\}\ e_c \in \{1,0\}} P(S= e_a , E= e_c)\ \log_2 \frac{P(S= e_a , E= e_c)}{P(S= e_a)\, P(E= e_c)} \qquad (3.1)$$

When S is a random variable that takes values $e_a = 1$ (the ethnic group contain allele AA) and $e_a = 0$ (the ethnic group does not contains allele AA), when every allelic pattern, AA, AB, BB and Nocall (or NC in the equations) is included, and E is a random variable that takes values $e_c = 1$ (the SNP is in the ethnic group c) and $e_c = 0$ (the SNP is not in the ethnic group c). The counts of the number of individuals with the eight possible combinations of indicator values are in Table 3.4.

Table 3.4 Genotype frequency at any locus

	$e_c = e_{\text{ethnic group}} = 1$	$e_c = e_{\text{ethnic group}} = 0$	Total
$e_a = e_{\text{allele}}$=AA	N_{AA1}	N_{AA0}	$N_{AA.}$
$e_a = e_{\text{allele}}$=AB	N_{AB1}	N_{AB0}	$N_{AB.}$
$e_a = e_{\text{allele}}$=BB	N_{BB1}	N_{BB0}	$N_{BB.}$
$e_a = e_{\text{allele}}$= NC	N_{NC1}	N_{NC0}	$N_{NC.}$
Total	$N_{.1}$	$N_{.0}$	$N_{..}$

The mutual information in terms of maximum likelihood estimations of the probabilities is written as equation (3.2).

$$
\begin{aligned}
I(S;E) = {} & \frac{N_{AA1}}{N_{..}} \log \frac{N_{AA1}N}{N_{.1}N_{AA.}} + \frac{N_{AB1}}{N_{..}} \log \frac{N_{AB1}N}{N_{.1}N_{AB.}} \\
& + \frac{N_{BB1}}{N_{..}} \log \frac{N_{BB1}N}{N_{.1}N_{BB.}} + \frac{N_{NC1}}{N_{..}} \log \frac{N_{NC1}N}{N_{.1}N_{NC.}} \\
& + \frac{N_{AA0}}{N_{..}} \log \frac{N_{AA0}N}{N_{.0}N_{AA.}} + \frac{N_{AB0}}{N_{..}} \log \frac{N_{AB0}N}{N_{.0}N_{AB.}} \\
& + \frac{N_{BB0}}{N_{..}} \log \frac{N_{BB0}N}{N_{.0}N_{BB.}} + \frac{N_{NC0}}{N_{..}} \log \frac{N_{NC0}N}{N_{.0}N_{NC.}}
\end{aligned}
\qquad (3.2)
$$

Given $I(S;\ E)$ is the value of any SNP locus in considered ethnic group.

$N_{.1}$ is the number of individuals in considered ethnic group.

$N_{.0}$ is the number of individuals does not in considered ethnic group.

$N_{AA.}$ is the total number of individuals who have allele pattern AA.

$N_{AB.}$ is the total number of individuals who have allele pattern AB.

$N_{BB.}$ is the total number of individuals who have allele pattern BB.

$N_{NC.}$ is the total number of individuals who have allele pattern NC.

Then, the value is calculated for each SNP locus for the entire ethnic groups and called mutual information value. The value measures how much information - in the information theoretic sense - a term contains about the class. Any SNP which has maximum value; means the SNP is closely related with the ethnic group (class) (Kwak and Choi, 2002). If a term's distribution is the same in the class as it is in the collection as a whole, then $I(S;E) = 0$ (Christopher *et al*., 2008).

3.3.2 Decision tree

3.3.2.1 Problem description

In decision tree problem description, there is a set of n SNPs, $S = \{s_1, \ldots, s_n\}$, of m sample individuals and their ethnic group. There is also decision tree which is able to predict ethnic group for any given subset, $\acute{S} \subseteq S$. The predict accuracy is defined by decision tree for subset $\acute{S}$ as $p(\acute{S}, n)$. The goal is finding a subset SNP which gives a maximal prediction accuracy, P_{max} (Kim *et al*, 2007).

$$\text{Select } \acute{S} \subseteq \{s_1, \ldots, s_n\} \text{ to maximize } p(\acute{S}, n)$$

Thus, the different SNPs number from the top MI value ranking list is chosen, as the subsets, to solve this problem. The subsets are 10, 20, 30, …, 100 SNPs of each ethnic group that use for the training set, then take them as the testing set and fine the one giving the best prediction accuracy.

3.3.2.2 SNP data set in decision tree

According to the Affymetrix export file, the SNP genotype of individual was extracted to use as the data set. The input data for classification is a collection of records. For each record, also known as an instance or example, is characterized by a tuple (x,y), where x is the attribute set as a SNP locus and y is a target attribute as ethnic group, designated as the class label (also known as category attribute). Table 3.5 shows a sample data set which is used for classifying population into one of the following categories:

- Yuan (denoted as “TU”), classifying the Yuan ethnic group.
- Lue (denoted as “TL”), classifying the Lue ethnic group.
- Yong (denoted as “TY”), classifying the Yong ethnic group.

- Khuen (denoted as “TK”), classifying the Khuen ethnic group.
- Lawa (denoted as “LW”), classifying the Lawa ethnic group.
- Mon (denoted as “MO”), classifying the Mon ethnic group.
- Mlabri (denoted as “MA”), classifying the Mlabri ethnic group.
- H’tin (denoted as “TN”), classifying the H’tin ethnic group.
- Paluang (denoted as “PL”), classifying the Paluang ethnic group.
- Plang (denoted as “PP”), classifying the Plang ethnic group.
- Karen (denoted as “KA”), classifying the ethnic group.
- Hmong (denoted as “HM”), classifying the Hmong ethnic group.
- Yao (denoted as “YA”), classifying the Yao ethnic group.

Table 3.5 SNP data set for decision tree analysis

Sample code (Individual)	SNP_A-1650338	SNP_A-1714915	…	SNP_A-1655697	Class label (Ethnic group)
TH-HM-001	AA	AB	…	AA	KA
TH-HM-002	AB	BB	…	AB	KA
…	…	…	…	…	…
…	…	…	…	…	…
TH-KA-019	AB	AB	…	…	TK
TH-KA-020	AA	AA	…	BB	TK

3.3.2.3 Population classification via decision tree using R program

In Rweka qewpackage of R program**,** C4.5 algorithm is used to build decision trees. C4.5 algorithm is an improvement of IDE3 algorithm, developed by Ross Quinlan (1993). It is based on Hunt’s algorithm (Hunts *et al*., 1966 cited in Anyanwu and Shiva, 2009) and also like IDE3. By using the best single feature test, the tree is

first constructed by finding the root node of the tree that is most discriminative for classifying. The criterion of the best single feature test is the normalized information gain, which results from choosing a feature to split the data into subsets. The test selects the feature with the highest normalized information gain as the root node.

3.3.2.4 Decision tree definition

Considering any SNP subset ($\acute{S} \subseteq S$), the training data (S) composing of SNPs which have N loci, is defined to be indexed by $j, 1 \leq j \leq N$. The N depends on the number of SNPs of any subset, which is chosen for data training (e.g., $N = 10,20,\ldots,100$). The possible attribute value is SNP genotype, which has values of AA, AB, BB, and NC. Let there be n_e distinct ethnic group, indexed by $k, 1 \leq k \leq n_e$, and let $Y_i \in \{1,\ldots,n_e\}$ be the ethnic group of individual $i, 0 \leq i \leq m$. Let $n_k \equiv \sum_{i=1}^{n} 1(Y_i = k)$ be number of individual in the training set of ethnic group k. Let G_{ij} be the genotype for individual i at SNPj.

Let Y be the ethnic group (class) from the 13 groups. The class attribute Y is discrete and has value of TU, TY, TL, TK, LW, MO, MA, TN, PL, PP, KA, HM and YA.

The goal is to learn from the training cases a function,

$$DOM(SNP_1) \times DOM(SNP_2) \times \ldots \times DOM(SNP_j) \rightarrow DOM(Y)$$

that maps from the SNP values to a predicted ethnic group.

3.3.2.5 Decision tree construction

The process is started by defining a measure called entropy, which measures the homogeneity of examples.

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus} \quad (3.3)$$

S is a sample of training examples

$p_{\oplus}$ is the proportion of positive examples in S

$p_{\ominus}$ is the proportion of negative examples in S

Information gain is simply the expected reduction in entropy caused by partitioning the examples according to this SNP.

$$Gain\ (S, A) \equiv Entropy(S) - \sum_{v \in Value\ (A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (3.4)$$

Values (A) is the set of all possible genotype for SNP A

S_v is the subset of S for which SNP A has genotype v.

First in the tree, the information gain for SNPs (SNP_A-1650338, SNP_A-1714915, …) are determined. Applying the computation of information gain with the SNP data, which used in this thesis, is shown below (Figure 3.2).

S: $[n_{KA}, n_{HM}, n_{YA}, n_{LW}, n_{MO}, n_{TN}, n_{MA}, n_{PL}, n_{PP}, n_{TU}, n_{TY}, n_{TL}, n_{TK}\ ]$

Thus, S: [20, 20, 19, 19, 20, 19, 19, 20, 20, 20, 20, 20, 20].

According to the equation (3.3)

$$Entropy(S) \equiv -p_{KA} \log_2 p_{KA} - p_{HM} \log_2 p_{HM} - \ldots - p_{TK} \log_2 p_{TK}$$

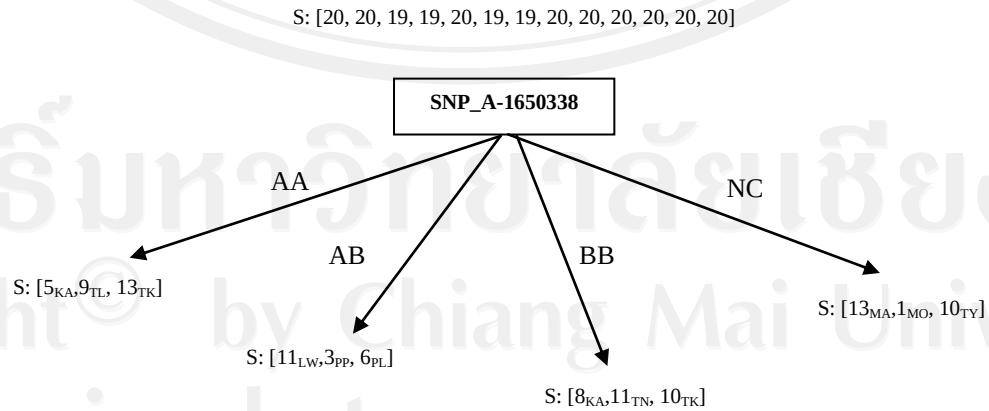


Figure 3.2 Example of information gain calculation

The information gain for SNP is:

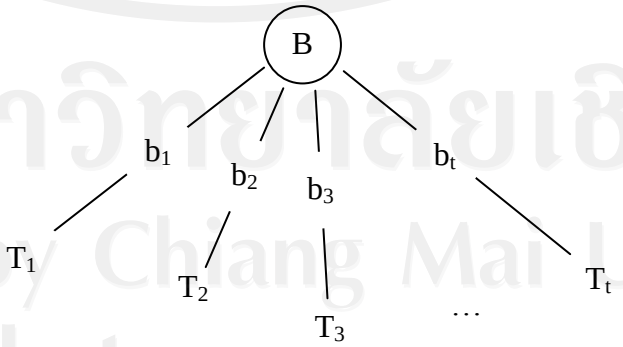
$$Gain\,(S, SNP_A - 1650338\,) \equiv Entropy(S) - \left(\frac{27}{256}\ 0.765\right) - \left(\frac{30}{256}\ 0.149\right) - \left(\frac{24}{256}\ 0.486\right) \equiv 0.151$$

The calculation is applied with all SNP, then the SNP which has highest gain value is selected; it is the decision attribute for the root node.

3.3.2.6 Divide and conquer algorithm

C4.5 first grows decision tree learners, using a method known as divide and conquer to construct a suitable tree from a training set S of cases:

- If all the cases in S belong to the same class (C_j, say), the decision tree is a leaf labeled with C_j.
- Otherwise, let B be a test with outcomes $b_1, b_2, \ldots, b_t$ that produces a non-trivial partition of S, and denote by S_i the set of cases in S that has outcome b_i of B. The decision tree is as shown below, where T_i is the result of growing a decision tree for the cases in S_i (Kohavi and Quinlan, 1999).



3.3.2.7 Candidate tests

C4.5 uses the tests of three types, each involving only a single attribute A_a. Decision regions in the instance space are thus bounded by hyperplanes, each orthogonal to one of the attribute axes.

In SNP case, A_a is a discrete attribute with z values, possible tests are:

-"A_a = ?" with z outcomes, one for each value of A_a, (this is the default.)

-"$A_a \in G$" with $2 \leq g \leq z$ outcomes, where $G = \{G_1, G_2, \ldots, G_g\}$ is a partition of the values of attribute A_a. Tests of this kind are found by a greedy search for a partition G that maximizes the value of the splitting criterion.

3.3.2.8 Selecting tests

In the divide and conquer algorithm, any test B that partitions S non-trivially will lead to a decision tree, but B_s give trees. Most learning systems attempt to keep the tree as small as possible, because small trees are easily understood and, by Occam's Razor arguments, are likely to have high predictive accuracy (Quinlan and Rivest, 1989). Since it is infeasible to guarantee the minimality of the tree (Hya and Rivest, 1976 cited in Kohavi and Quinlan, 1999), C4.5 relies on greedy search, selecting the candidate test that maximizes a heuristic splitting criterion. In C4.5, there are two criterions, information gain and gain ratio.

Let $RF(C_j, S)$ denote the relative frequency of cases in S that belong to class C_j . The information content of a message that identifies the class of a case in S is then

$$I(S) = -\sum_{j=1}^{r} RF(C_j, S) \log\left(RF(C_j, S)\right). \quad (3.5)$$

After S is partitioned into subsets $S_1, S_2, \ldots, S_t$ by a test B, the information gained is then

$$G(S,B) = I(S) - \sum_{i=1}^{t} \frac{|S_i|}{|S|} log\left(\frac{|S_i|}{|S|}\right) \quad (3.6)$$

The gain criterion chooses the test B that maximizes $G(S;\ B)$.

A problem with this criterion is that it favors tests with numerous outcomes - for example, $G(S,B)$ is maximized by a test in which each S_i contains a single case. The gain ratio criterion sidesteps this problem by also taking into account the potential information from the partition itself:

$$P(S,B) = -\sum_{i=1}^{t} \frac{|S_i|}{|S|} log \frac{|S_i|}{|S|} \quad (3.7)$$

Gain ratio then chooses, from among the tests with at least average gain, the test B that maximizes $\frac{G(S,B)}{P\ (S,B)}$.

3.3.2.9 Performance of evaluation measure

The training set is used to build a classification model, which is subsequently applied to the test set, consisting of records with unknown class label.

Evaluation of the performance of a classification model is based on the counts of test records correct and incorrect predicted by the model. These counts are contained in the table, known as a confusion matrix. Consider in 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. 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, the actual ethnic group is compared to the ethnic group that

Table 3.6 Testing data set for decision tree analysis

Sample code (Individual)	SNP1	SNP2	…	SNPn	class label (ethnic group)
TH-HM-001	AA	AB	…	AA	?
TH-HM-002	AB	BB	…	AB	?
…	…	…	…	…	?
…	…	…	…	…	?
TH-KA-019	AB	AB	…	…	?
TH-KA-020	AA	AA	…	BB	?

was assigned by the trained classifier. In Table 3.7, the number of positive (negative) example that is correctly classified by the classifier, is called a true positive (true negative); a number of positive (negative) example that is incorrectly classified, is called a false negative (false positive) (Tan *et al*., 2006).

Table 3.7 Calculation of accuracy

		Number of model prediction	
		+	-
Actual	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

The accuracy obtains from equation (3.8) and error rate from equation (3.9).

$$\text{Accuracy} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3.8)$$

$$\text{Error rate} = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions} = \frac{(FP+FN)}{(TP+TN+FP+FN)} \quad (3.9)$$

or $$\text{Error Rate} = 1 - \text{Accuracy}$$

TP is the number of correct predictions when an example is from positive class;

TN is the number of correct predictions when an example is from negative class;

FN is the number of incorrect predictions when an example is from negative class;

FP is the number of incorrect predictions when an example is from positive class.

3.3.3 Correspondence analysis

Correspondence analysis is an exploratory data analytic technique designed to analyze simple two-way and multi-way tables, containing some measure of correspondence between the rows and columns. The aim, common to all of these, is the representation of a data set by a number of points in multidimensional space, enabling a visual interpretation of the patterns existing in the data (Greenacre and Degos, 1977).

3.3.3.1 Contingency table in research

The following contingency table shows the SNP genotype (AA, AB, BB and NC) frequencies of thirteen ethnic groups in the SNP subset, which have the high accuracy from decision tree, among 256 individuals.

Table 3.8 Example of contingency table of SNP genotype

Ethnic group	$SNP1_{AA}$	$SNP1_{AB}$	$SNP1_{BB}$	$SNP1_{NC}$	…	SNP N_{NC}
KA	1	5	13	1		9
HM	7	10	3	0		13
⋮						
TY	18	0	0	2		2
TL	15	4	0	1		3
TK	1	15	2	2		6

This data matrix in Table 3.8 contains the counts of the S SNP loci multiply with number genotype format (S x 4) for 13 different ethnic groups (rows of matrix). Each row contains genotype counts in a SNP of individual from each ethnic group. Note that S obtains from the SNP set which has the highest decision tree classification accuracy.

3.3.3.2 Correspondence analysis using R program

The package ‘ca’ version 0.33 in R program is used. The basic concepts of profile, mass and chi-squared distance are introduced in an initial simple example using data on the relationship between population ethnic group and SNP loci. The

main result of the correspondence analysis is a geometric map of this relationship, showing the relative frequencies of population ethnic group with the genotype frequency in each SNP loci.

As in principal component analysis, the idea is to reduce the dimensionality of a data matrix and visualize it in a subspace of low-dimensionality, commonly two- or three dimensional.

There are certain fundamental concepts and definition in correspondence analysis which is described below.

1) Correspondence table

The original data matrix, $N\ (I\ ,J)$, or contingency table, is called the primitive matrix or primitive table. The elements of this matrix are n_{ij} . The data of interest in simple correspondence analysis are usually a two-way contingency table, or any other table of nonnegative ratio-scale data, for which relative values are of primary interest. In this application, the matrix consists of SNP genotype frequencies, such that n_{ij} is the frequency of SNP_j in the ethnic group i. Ethnic groups figure as rows and SNP genotypes at each locus as columns of this matrix.

2) Profile (set of proportion)

While interpreting a cross-tabulation, it makes little sense to compare the actual frequencies in each cell. Each row and each column has a different number of respondents, called the base of respondents. For comparison, it is essential to reduce either the rows or columns to the same base.

When consider a contingency table, $N\ (I,J)$ with $I\ rows\ (i = 1,2,I)$ and J columns ($j\ = 1, 2, \dots, J$) having frequencies n_{ij}, marginal frequencies are denoted by n_{i+} and n_{+j}.

$$n_{i.} = \sum_j n_{ij} \quad (3.10)$$

$$n_{.j} = \sum_i n_{ij} \quad (3.11)$$

Total frequency is

$$n = \sum_j \sum_i n_{ij} \quad (3.12)$$

- **Row profile**

The profile of each row i is a vector of conditional densities:

$$R = \frac{n_{ij}}{n_{i.}} \quad (i = 1, 2, \dots, I \text{ and } j = 1, 2, \dots, J) \quad (3.13)$$

Table 3.9 Matrix of rows profile

Rows	**Columns**				**Total**
	1	2	…	J	
1	$n_{11}/n_{1.}$	$n_{12}/n_{1.}$	…	$n_{1j}/n_{1.}$	1
2	$n_{21}/n_{2.}$	$n_{22}/n_{2.}$	…	$n_{2j}/n_{2.}$	1
3	$n_{31}/n_{3.}$	$n_{33}/n_{3.}$	…	$n_{3j}/n_{3.}$	1
⋮	⋮	⋮		⋮	⋮
I	$n_{i1}/n_{i.}$	$n_{i2}/n_{i.}$	…	$n_{ij}/n_{i.}$	1
Column mass	$n_{.1}/n_{..}$	$n_{.2}/n_{..}$	…	$n_{.j}/n_{..}$	1

Define the set of row profile as $i \; x \; j$ matrix R

The average of row profile as follow ;

$$\bar{r} = \frac{n_{.j}}{N} \quad (j = 1, 2, \dots, J) \quad (3.14)$$

- **Column profile**

The profile of each column j is a vector of conditional densities. The complete set of the column profiles may be denoted by (i ´ j) matrix C.

$$C = \frac{n_{ij}}{n_{.j}} \quad (\text{i= 1, 2,… ,I and j= 1, 2,… ,J}) \quad (3.15)$$

Table 3.10 Matrix of columns profile

Rows	**Columns**				**Row mass**
	1	2	…	J	
1	$n_{11}/n_{.1}$	$n_{12}/n_{1.}$	…	$n_{1j}/n_{.j}$	$n_{.1}/n_{..}$
2	$n_{21}/n_{.1}$	$n_{22}/n_{2.}$	…	$n_{2j}/n_{.j}$	$n_{.2}/n_{..}$
3	$n_{31}/n_{.1}$	$n_{33}/n_{3.}$	…	$n_{3j}/n_{.j}$	$n_{.3}/n_{..}$
⋮	⋮	⋮		⋮	
I	$n_{i1}/n_{.1}$	$n_{i2}/n_{i.}$	…	$n_{ij}/n_{.j}$	$n_{.i}/n_{..}$
Total	1	1		1	1

Define the set of row profile as i x j matrix C

The average of column profile ;

$$\bar{c} = \frac{n_{i.}}{N} \quad (i= 1, 2, ..., I) \tag{3.16}$$

3) Mass (Marginal profile)

Another fundamental concept in correspondence analysis is the concept of mass which obtain from the following equations:

$$mass\ value\ of\ \ row\ i^{th} = \frac{Marginal\ frequency\ of\ i^{th}\ row}{\text{Grand total}}$$

$$= \frac{n_{.i}}{n} \tag{3.17}$$

$$mass\ value\ of\ column\ \ j^{th} = \frac{Marginal\ frequency\ of\ j^{th}\ row}{\text{Grand total}}$$

$$= \frac{n_{j.}}{n} \tag{3.18}$$

4) Distance

Distance measure in correspondence analysis is Chi-square distance method. Where $d^2(i,i)$ is the distance between two rows, which row i[th] and i from formula:

$$d^2(i,i) = \sum_{j=1}^{J} \frac{i}{n_{.j}} \left[\frac{n_{ij}}{n_{i.}} - \frac{n_{ij}}{n_{i.}}\right] \quad (3.19)$$

In the same way, the distance between two column j and j calculates from

$$d^2(j,j) = \sum_{i=1}^{I} \frac{1}{n_{i.}} \left[\frac{n_{ij}}{n_{.j}} - \frac{n_{ij}}{n_{.j}}\right] \quad (3.20)$$

The Chi-square distance differs from the usual Euclidean distance in that each square is weighted by the inverse of the frequency corresponding to each term.

5) Inertia

Inertia is a term borrowed from the "moment of inertia" in mechanics. A physical object has a center of gravity (or centroid). Every particle of the object has a certain mass *m* and a certain distance *d* from the centroid. The moment of inertia of the object is the quantity md^2 summed over all the particles that constitute the object.

$$Moment\ of\ inertia\ = md2 \quad (3.21)$$

This concept has an analogy in correspondence analysis. There is a cloud of profile points with masses adding up to 1. These points have a centroid (*i.e.*, the average profile) and a distance (Chi-square distance) between profile points. Each profile point contributes to the inertia of the whole cloud. The inertia of a profile point can be computed by the following formula.

Where r_{ij} is the ratio $\frac{nw}{n_{\ i+}}$ and is $\frac{n.j}{n}$

The inertia of the j^{th} column profile is computed similarly.

The total inertia of the contingency table is given by:

$$Total\ inertia = \sum m_i d_i^2 \quad (3.22)$$

Where m_i is mass of the point i^{th}

d_i is the distance from i^{th} to controid measure by chi-square distance method.

Also, the proportion of inertia is proportion of inertia in each dimension as follow:

$$Proportion\ of\ inertia = \frac{inertia_i}{total\ inertia} \quad (3.23)$$

6) Score in dimension

Score in dimension is the co-ordinate of each variable in dimension 1 and 2 demonstrate in correspondence mapping.

3.3.3.3 Visualization of correspondence analysis result

The correspondence analysis results are presented on graphs that represent the configurations of points in projection planes, formed by the first principal axes taken two at a time. It is customary to summarize the row and column coordinates in a single plot. The graph is commonly done with so-called symmetric maps. In that case, the row and column coordinates on each axis are scaled to have inertias (weighted variances) equal to the principal inertia (eigenvalue) along that axis: these are the principal row and column coordinates. Depending on the situation, other types of display are appropriate (Nenadić and Greenacre, 2007).

3.3.4 Population genetic distance and relationship visualization

Genetic distance analysis, which focuses on average genetic distance between populations, is quite efficient while constructing an evolutionary tree from allele frequency data. In this research, the genetic distance calculates in pairwise difference, using PEAS program (Xu and Jin, 2010). Then, phylogenetic tree analysis was performed as an implement in the MEGA5 software (Tamura *et al*., 2011). There are several genetic distances perform well for reconstruction of phylogenetic when the populations are of the same species and are very closely related (Dínç H., 2003). Thus, genetic distance matrix of Nei's standard and Cavalli–Sforza are provided, applicable to SNP genotype data that are widely used in human genetic studies. The SNP loci are obtained from the subset which has the highest decision tree classification accuracy.

3.3.4.1 Genetic distance using PEAS program

This format is the same style as HapMap genotype data, with SNPs in rows and genotypes of sample in columns. But the genotypes are coded by single character, with 'A' and 'B' coding for two homozygotes, 'H' coding for heterozygote and 'U' coding for missing genotype. Because of the large SNP surveys which have much larger number of SNPs than that of individuals, thus this format is more readable than the others. The genotype data file is supplied by the user to specify how many individuals there are to be analyzed, how many sites each individual has been typed at, and the genotypes for each individual. The information that the user has to provide includes also ID of SNPs (the first column), which chromosome that each SNP is of (the second column), the physical position of each SNP (the third column),

the two possible allele state of each SNP (the fourth column), which DNA strand each SNP was genotyped (the fifth column), followed by genotype data (the rest columns). One example of standard format of genotype data can be seen in Appendix D:

The population distances estimation use PEAS program, which can provide including Wright's F_{ST}, F_{ST} distance, Nei's standard distance, Nei's D_A distance and Cavalli-Sforza's distance. The program generates also output files which can be recognized by MEGA and PHYLIP programs for further processing. In this research, the only two distances, including Nei's standard and Cavalli-Sforza's distance can be calculated. The notation of the distance measures are shown below:

1) Nei's standard distance

Nei (1972), developed a genetic distance measure (called standard genetic distance) whose expected value is proportional to evolutionary time, when both effects of mutation and genetic drift are taken in to account. The Nei's standard genetic distance, consider two populations X and Y, is defined as follows.

$$D = -ln(I) \qquad (3.24)$$

where I is the normalized identity of SNPs between X and Y with respect to the average in all loci, is defined as

$$I = \frac{\hat{J}_{XY}}{\sqrt{\hat{J}_X \hat{J}_Y}} \qquad (3.25)$$

$\hat{J}_X$, $\hat{J}_Y$ and $\hat{J}_{XY}$ are the unbiased estimates of average of $\sum x_i^2$, $\sum y_i^2$ and $\sum x_i y_i$ for all loci respectively. Let x_i and y_i be the frequencies of the k-th alleles ($i = 1, \dots, N$). For a single locus, the unbiased estimates of $\sum x_i^2$, $\sum y_i^2$ and $\sum x_i y_i$ are:

$$\hat{J}_X = \frac{2m_x \sum \hat{X}_i^2 - 1}{2m_x - 1} \qquad (3.26)$$

$$\hat{J}_Y = \frac{2m_y \sum \hat{Y}_i^2 - 1}{2m_Y - 1} \quad (3.27)$$

$$\hat{J}_{XY} = \sum \hat{X}_i \hat{Y}_i \quad (3.28)$$

where m_X and m_Y are the number of diploids sampled from population X and Y respectively, $\hat{x}_i$and $\hat{y}_i$ are allele frequencies in samples of allele A_i in population X and Y (Nei, 1987). Therefore, $\hat{J}_X$, $\hat{J}_Y$ and $\hat{J}_{XY}$ are the averages of $\hat{J}_X$, $\hat{J}_Y$ and $\hat{J}_{XY}$ in all loci, respectively

The variances of I and D can be computed by the formulas given by Nei (1978, 1987)

2) Cavalli-Sforza's distance

Distance measures based on geometric consideration. According to Cavalli-Sforza and Edwards (1967) also used an angular transformation. They proposed that the genetic distance between two populations be measured by the chord length between points X and Y on the q-dimensional hypersphere. This chord length is given by $[2(1 - cos\theta)]^{1/2}$. Since $\theta = \pi/2$ corresponds to a complete gene substitution, it is convenient to work in terms of $2\pi/\theta$, where θ is in radians, for the unit distance is then one gene substitution.

$$d_c = \frac{\sqrt{2(1 - \sum_{i=1}^{q} \sqrt{x_i y_i})}}{\pi} \quad (3.29)$$

where q is the number of alleles in k^{th} locus.

3.3.4.2 Phylogenetic tree and multidimensional scaling

The distance value is used to construct the phylogenetic tree by Neighbour Joining (Saitou and Nei, 1987). MEGA5 software is used to graphically display the results. The input data file for MEGA5 is matrix format generating from PEAS program. The distance matrix ($i\ x\ j$) shows the distance between the ethnic groups. The example is shown in Table 3.11. Note that the matrix is symmetric, therefore only values of one side of the diagonal need to be computed.

Table 3.11 Distance matrix example

		i						
		KA	HM	YA	LW	…	TL	TK
	KA	-						
	HM	0.195031	-					
j	YA	0.166200	0.167043	-				
	…							
	TK	0.169001	0.211146	0.377981	0.54663	…	0.195137	0.193316

Since a tree presentation of the distance matrix might be misread as a succession of population splits (Kampuansai, 2007), multidimensional scaling was also performed by R program. The purpose of using multidimensional scaling is to provide a visual representation of the complex pattern of genetic distance among a set of populations. These distance values were projected onto two-dimensional space applying classical multidimensional scaling to the distance matrix (Table 3.11) using the R function cmdscale. Multidimensional scaling is the methods for reconstructing

a map from a distance matrix. The map is not restricted to two dimensions-it can be one dimensional, three dimensional, or higher dimensional. Multidimensional scaling techniques attempt to find a set of coordinates for the objects, and representation of the units in a given number of dimensions, so that the most similar objects are plotted close together and the most dissimilar objects are plotted furthest apart (Everitt and Hothorn, 2009). Thus, populations that are perceived to be very similar to each other are placed near each other on the map, and those that are perceived to be very different are placed far away.