CHAPTER 4

RESULT AND DISCUSSION

4.1 Specific SNP selection

The first 100 SNP with highest mutual information values, in each ethnic group, were selected. The 10 SNP subsets comprising the 1st -10th, 1st – 20th,….....1st - 100th, were grouped. Using the classifier, all 10 subsets were calculated for their prediction accuracy, and the only set with the highest accuracy and the least of SNP was selected.

4.1.1 Mutual information technique

When estimating the mutual information values for 58,960 SNP loci separated from the ethnic groups, the values were ranked descending. The high value of any SNP locus means that the SNP is related to the considerable ethnic group, and that SNP should be selected as the specific SNP for their ethnic group.

Results of the highest and average mutual information value of 13 ethnic groups are shown in Table 4.1, as for the lowest values of all are zero. The multhist() function, which is the function for plotting a multiple histogram in plotrix package of R program, was used to compute the frequency of number of SNP loci. Table 4.2 shows that in each ethnic group, the higher MI value range, the lesser number of SNP there are. The high value indicated that the SNP is specific with the considered population. The histogram graphs are used to display the number of SNP loci distributions, which the horizontal axis is the log base 2 of the number of SNP loci

Table 4.1 Mutual information value of 13 ethnic groups

Ethnic group	Code	Linguistic affiliation	MI values	
		(Family, Subfamily)	Max	Average
Karen	KA	Sino-Tibetan, Tibeto-Burmese	0.1070	0.0038
Hmong	HM	Hmong–Mien, Hmong	0.1813	0.0113
Yao	YA	Hmong–Mien, Yao	0.1524	0.0039
Lawa	LW	Austro-Asiatic, Mon-Khmer	0.1047	0.0101
H'tin	TN	Austro-Asiatic, Mon-Khmer	0.2061	0.0122
Mlabri	MA	Austro-Asiatic, Mon-Khmer	0.3321	0.0092
Mon	MO	Austro-Asiatic, Mon-Khmer	0.1154	0.0017
Paluang	PL	Austro-Asiatic, Mon-Khmer	0.1614	0.0007
Plang	PP	Austro-Asiatic, Mon-Khmer	0.0833	0.0002
Yuan	TU	Tai–Kadai, Tai	0.0746	0.0032
Yong	TY	Tai–Kadai, Tai	0.0868	0.0087
Lue	TL	Tai–Kadai, Tai	0.0918	0.0084
Khuen	TK	Tai–Kadai, Tai	0.0991	0.0080

and the vertical axis is the mutual information value range (Figure 4.1). The results obviously indicate that the groups of SNPs with small number and high MI value are suitable to be used as the feature in the classification method.

4.1.2 Population classification using decision tree

The SNP locus of each ethnic group was selected base on the MI values ranking list (Appendix A, Table A.1). Due to the limitation of classification program that

Table 4.2 Number of SNP at mutual information value ranges

Ethnic group	MI value range																	Total SNP
	0-0.019	0.02-0.039	0.04-0.059	0.06-0.079	0.08-0.099	0.10-0.119	0.12-0.139	0.14-0.159	0.16-0.179	0.18-0.199	0.20-0.219	0.22-0.239	0.24-0.259	0.26-0.279	0.28-0.299	0.30-0.319	0.32-0.339	
KA	52844	5417	593	90	14	2	0	0	0	0	0	0	0	0	0	0	0	58960
HM	48986	8081	1497	313	59	17	4	2	0	1	0	0	0	0	0	0	0	58960
YA	51766	6370	692	98	24	5	2	3	0	0	0	0	0	0	0	0	0	58960
LW	50908	6885	975	162	28	2	0	0	0	0	0	0	0	0	0	0	0	58960
TN	47675	8807	1887	438	102	37	7	5	1	0	1	0	0	0	0	0	0	58960
MA	32622	12299	6136	3133	1889	1144	660	425	285	159	92	46	37	20	4	3	6	58960
MO	53102	5326	458	68	5	1	0	0	0	0	0	0	0	0	0	0	0	58960
PL	48117	8775	1625	350	75	12	3	2	1	0	0	0	0	0	0	0	0	58960
PP	53477	5081	380	21	1	0	0	0	0	0	0	0	0	0	0	0	0	58960
TU	55562	3244	151	3	0	0	0	0	0	0	0	0	0	0	0	0	0	58960
TL	53787	4756	374	41	2	0	0	0	0	0	0	0	0	0	0	0	0	58960
TY	53162	5343	413	40	2	0	0	0	0	0	0	0	0	0	0	0	0	58960
TK	54312	4372	259	16	1	0	0	0	0	0	0	0	0	0	0	0	0	58960

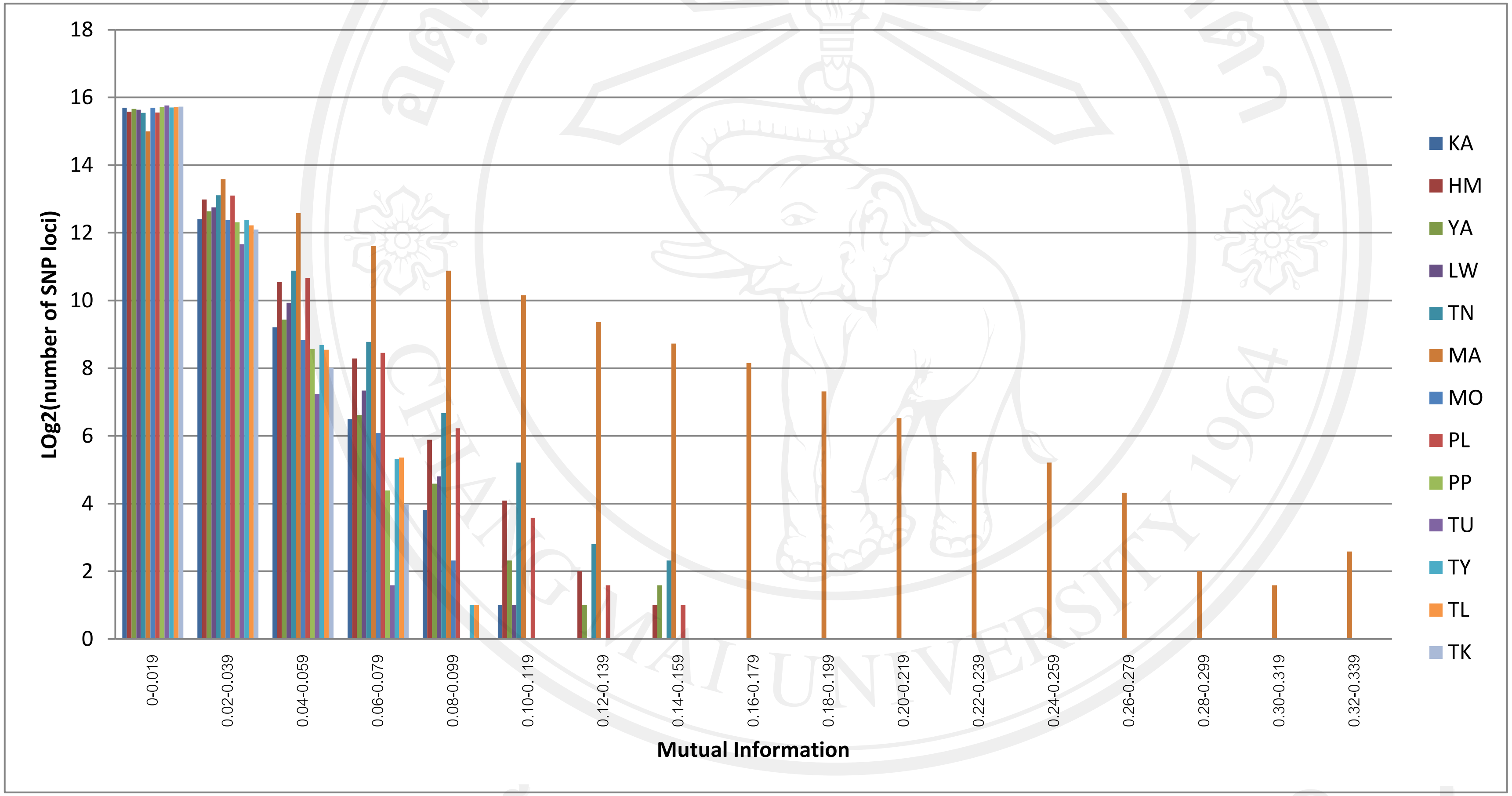

Figure 4.1 Mutual information value distribution of all SNP among 13 ethnic groups

cannot manage the large number of SNP, thus the limited number of SNP would be used. The top 100 SNP loci, with highest MI values to the lower one, were then chosen consecutively, to form 10 new feature sets which are different in the number of SNP loci. The ten new feature sets compose of SNP 1-10, 1-20, 1-30 ……1-100 from the already chosen ranking list (Table 4.3). Then input them as the training and testing data to the decision tree classifier. The criterions for SNP specific set selection are the highest classification accuracy, the smaller number of SNP loci, the lesser number of leaves and the smaller size of tree.

Table 4.3 Classification accuracy of different SNP numbers from ranking list

SNP Numbers	Accuracy (%)	Number of leaves	Size of the tree
10	82.03 (210/256)	117	162
20	87.11 (223/256)	137	188
30	87.50 (224/256)	133	183
40	89.06 (228/256)	143	198
50	89.45 (229/256)	139	194
60	**89.84 (230/256)**	**131**	**181**
70	87.50 (224/256)	135	184
80	87.89 (225/256)	126	174
90	88.28 (226/256)	129	178
100	87.11 (223/256)	121	167

The experimental results show that, when 60 SNPs set of each ethnic group is input, the highest classification accuracy of 89.84% is observed (Table 4.3). The percentage of classification accuracy means that when 100 unknown populations are used for prediction, approximately 90 populations can be correctly predicted.

Moreover, the data in confusion Matrix (Table 4.4) also indicate the correct ethnic group prediction of 90 percent by 230 individuals. The decision tree generated by Graphviz software is shown in Appendix B, Figure B.2 (a), (b), (c) and (d)). The tree identifies all of the 761 SNP loci. The nodes of tree are SNP loci, and branches are SNP genotypes, while the leaves are the ethnic groups.

Table 4.4 Confusion matrix for genotype data of 60 SNP loci

		Model predict												
		KA	HM	YA	LW	TN	MA	MO	PP	PL	TU	TY	TL	TK
Actual	KA	**19**	0	0	0	0	0	1	0	0	0	0	0	0
	HM	0	**20**	0	0	0	0	0	0	0	0	0	0	0
	YA	1	1	**13**	0	0	0	0	0	0	1	1	1	1
	LW	3	0	0	**16**	0	0	1	0	0	0	0	0	0
	TN	0	0	0	0	**20**	0	0	0	0	0	0	0	0
	MA	0	0	0	0	1	**18**	0	0	0	0	0	0	0
	MO	0	0	0	1	0	0	**18**	0	0	0	0	0	0
	PP	0	1	0	0	0	0	1	**20**	0	0	0	0	0
	PP	0	0	0	1	0	0	0	0	**19**	0	0	0	0
	TU	0	0	0	0	0	1	0	0	1	**17**	0	0	1
	TY	0	0	0	1	0	0	0	0	1	0	**17**	0	0
	TL	0	0	0	1	0	0	1	0	1	0	1	**17**	0
	TK	0	0	0	1	0	0	0	1	1	0	0	0	**18**

4.1.3 Correspondence analysis of SNP set

To further examine the 60 SNP set discriminating efficiency, the correspondence analysis was performed to find the relationships among the 13 ethnic groups. The outputs of correspondence analysis contain eigenvalues, relative percentages, cumulated percentage and screen plot of explained inertia in all available similarity between the SNP genotype-frequency profiles of group - the Mlabri (MA) is far from the others because the profile is different, whereas the Karen (KA), Yong (TY) and Lawa (LW) are close together because their profiles are similar. However, the result is not correspondent with the linguistic identification

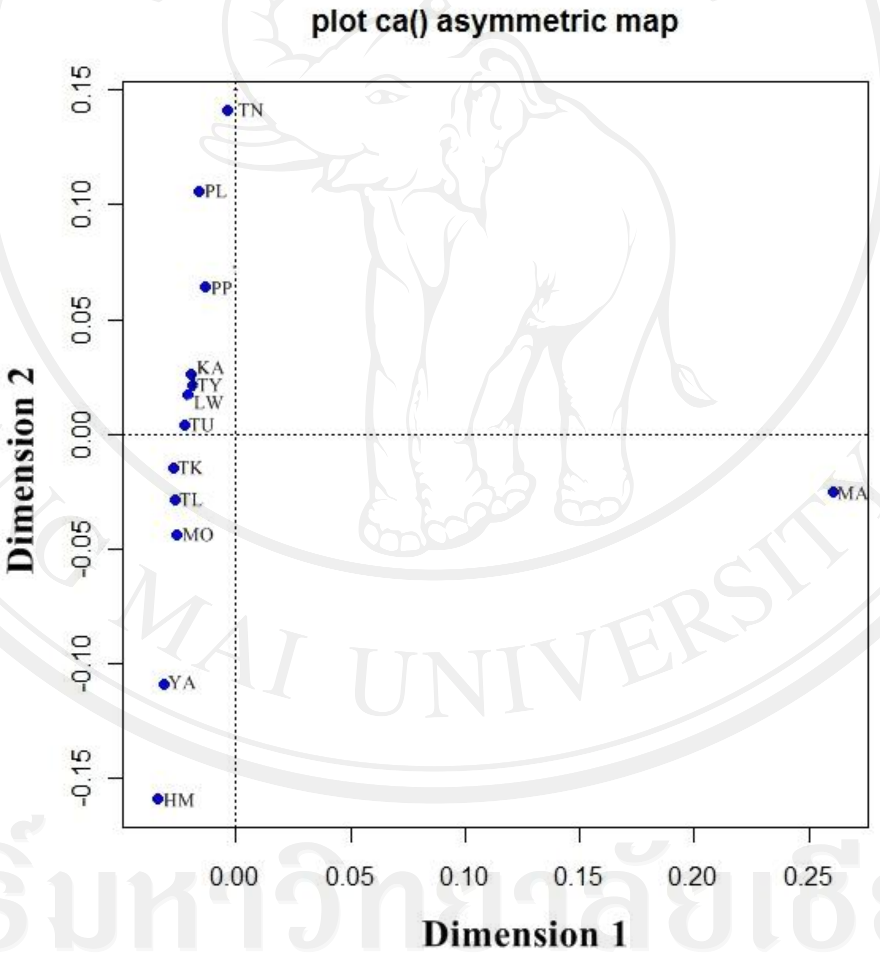

Figure 4.2 Two dimensional visualizing correspondence analysis based on genotype frequency of the top 60 SNP loci

4.2 Population genetic distances using SNP data

To examine SNP efficiency in term of the genetic relationship among the ethnic groups, the selected SNP set (top 1st - 60th loci) was analyzed. Their genotype frequencies were used to compute the genetic distance between populations. The genetic distances provide a relative estimation of the time that must have elapsed since the populations existed as a unified and a cohesive population (Tamura *et al*., 2007).

The Nei's standard genetic distance and the Cavalli-Sforza distance were estimated among the ethnic groups (Table 4.5). Maximum Nei's standard distance was observed between Mlabri and Paluang (PL) (0.6288) where minimum genetic distance was between the Khuen (TK) and Yuan (TU) (0.0527). Cavalli-Sforza distance revealed minimum genetic distance between the Khuen and Yong (0.0828). The maximum genetic distance was between the Mlabri and Paluang (0.4062).

The unrooted neighbor joining trees were constructed based on Nei's standard distance (Figure 4.3), and Cavalli-Sforza's distance (Figure 4.4). Both the Nei's standard and Cavalli-Sforza distance phylogenetic tree reveals Yong, Khuen and Lue have closed genetic relationship and the same origin. The Paluang is far apart from other populations, while the Yong and Khuen are the closest ones. This can be explained that the Khuen and Yong ancestors might have separated recently from each other. In the tree which was constructed from Nei's standard distance value, the Karen and Lawa joined with one another although they speak different languages. Even though the geographic distance between them is quite close, since the Karen live in Mae Sariang and the Lawa live in Mae La Noi districts of Mae Hongson province, it cannot be assumed that they have genetically mixed.

The two-dimensional solution from classical multidimensional scaling of Nei's standard distance matrix is shown in Figure 4.5 and Cavalli-Sforza's distance is shown in Figure 4.6. These are visualized the genetic relationship among the 13 ethnic group populations. The results show that, the Tai speaking populations (TU, TY, TL and TK) clustered together indicate a close genetic relationship among them. The Mlabri and Palaung are separately plotted from the Mon-Khmer populations and segregated away from the cluster, which indicates high genetic differentiation. This result is correspondent with Kutanan *et al*. (2011). From the results, the visualization of multidimensional scaling can clustered better than correspondence analysis when consider on the population linguistic affinity.

Table 4.5 Cavalli-Sforza (above the diagonal) and Nei's standard genetic distance (below the diagonal) among studied populations

	KA	HM	YA	LW	TN	MA	MO	PP	PL	TU	TY	TL	TK
KA	-	0.1281	0.1028	0.1122	0.1304	0.2058	0.1080	0.1000	0.3625	0.1004	0.1097	0.1199	0.1109
HM	0.0857	-	0.0978	0.1350	0.1462	0.2149	0.1189	0.1202	0.3723	0.1113	0.1094	0.1160	0.1053
YA	0.0666	0.0657	-	0.1153	0.1247	0.2030	0.0957	0.1012	0.3731	0.0921	0.0915	0.0961	0.0972
LW	0.0720	0.0940	0.0797	-	0.1341	0.2144	0.1125	0.1023	0.3646	0.1064	0.1147	0.1235	0.1155
TN	0.0900	0.1058	0.0882	0.0964	-	0.2050	0.1281	0.1168	0.3741	0.1209	0.1256	0.1233	0.1264
MA	0.1650	0.1855	0.1703	0.1805	0.1773	-	0.2115	0.2004	**0.4062**	0.2054	0.2032	0.2118	0.2091
MO	0.0660	0.0739	0.0594	0.0694	0.0834	0.1659	-	0.0996	0.3601	0.0903	0.1018	0.1082	0.1001
PP	0.0657	0.0787	0.0677	0.0648	0.0760	0.1622	0.0612	-	0.3602	0.0894	0.0979	0.1063	0.0956
PL	0.5721	0.5872	0.5904	0.5698	0.5816	**0.6288**	0.5614	0.5592	-	0.3635	0.3697	0.3714	0.3740
TU	0.0654	0.0747	0.0622	0.0694	0.0792	0.1685	0.0536	0.0565	0.5677	-	0.0835	0.0911	0.0862
TY	0.0754	0.0744	0.0645	0.0758	0.0838	0.1694	0.0607	0.0608	0.5784	0.0542	-	0.0907	**0.0828**
TL	0.0798	0.0791	0.0638	0.0811	0.0851	0.1752	0.0637	0.0653	0.5810	0.0574	0.0594	-	0.0941
TK	0.0721	0.0684	0.0624	0.0742	0.0845	0.1794	0.0579	0.0602	0.5846	**0.0527**	0.0533	0.0600	-

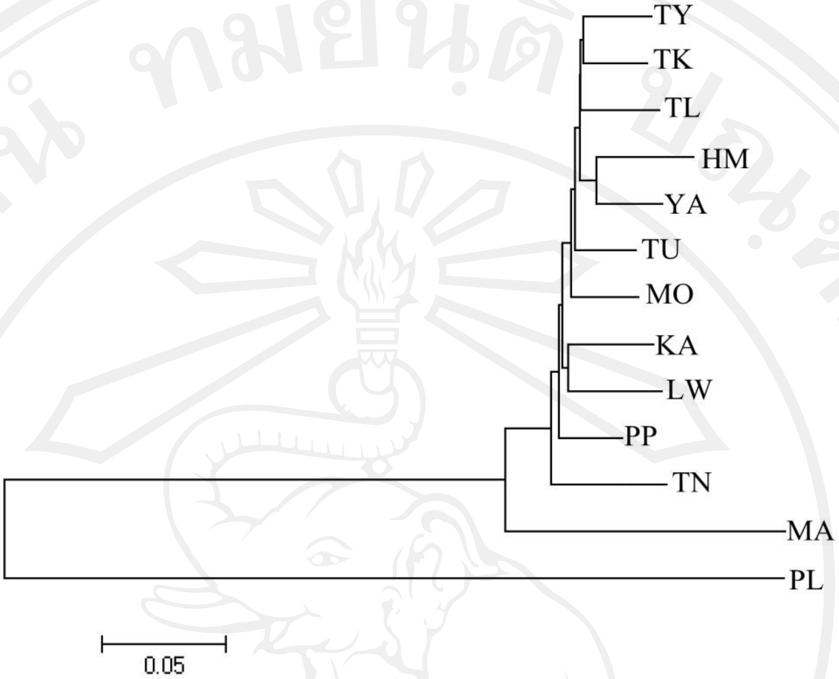

Figure 4.3 Unrootd neighbor-joining tree based on Nei's standard distance from 60 SNPs of each population

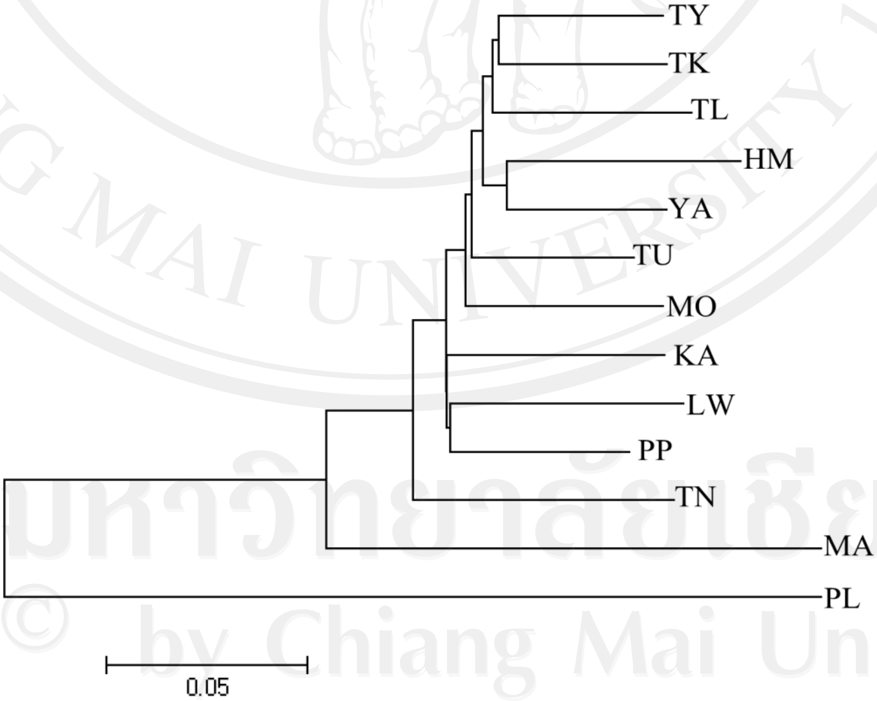

Figure 4.4 Unrooted neighbor-joining tree of 13 populations based on Cavalli-Sforza's distance from 60 SNPs of each population

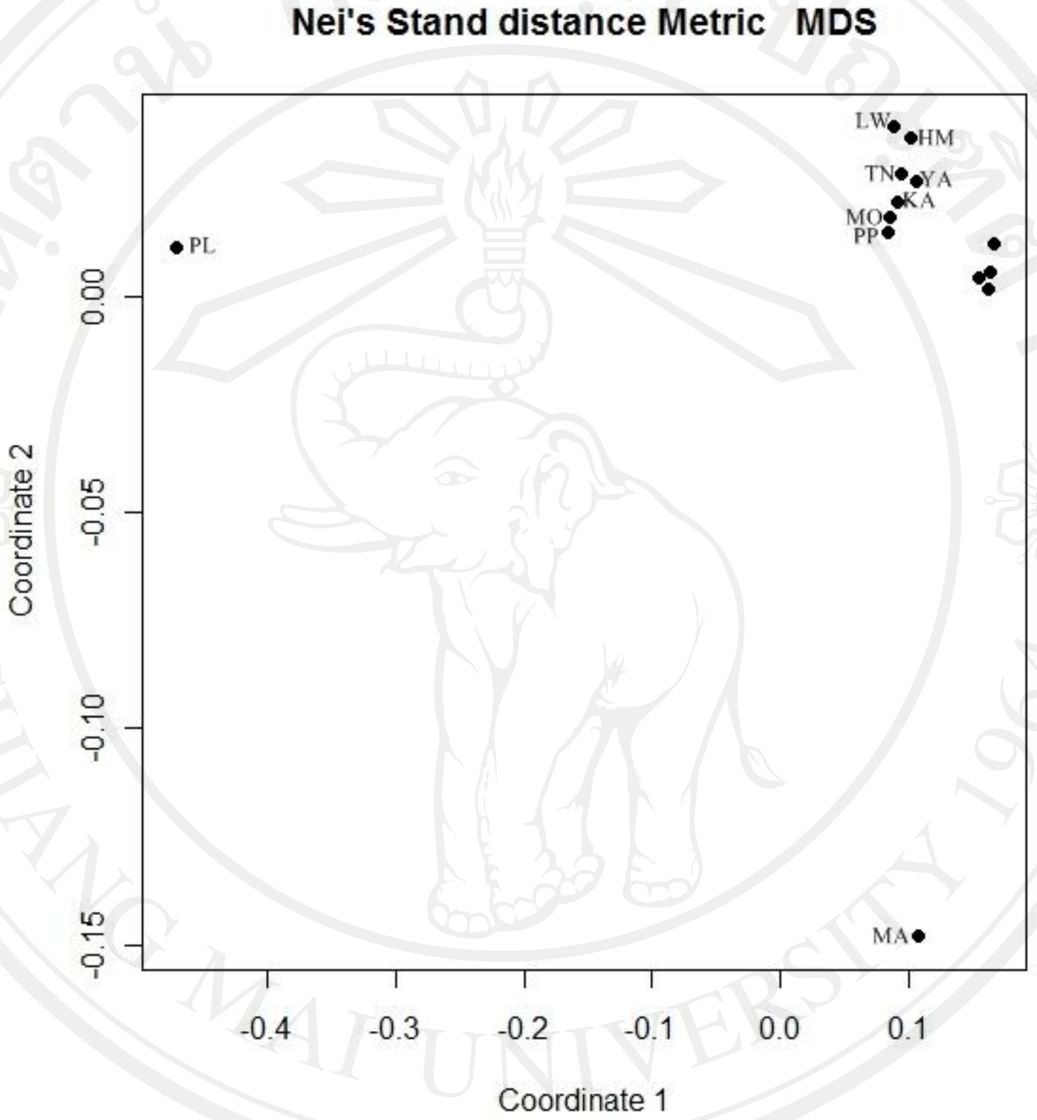

Figure 4.5 Two-dimensional graph from classical multidimensional scaling of 13 populations based on Nei's standard distance

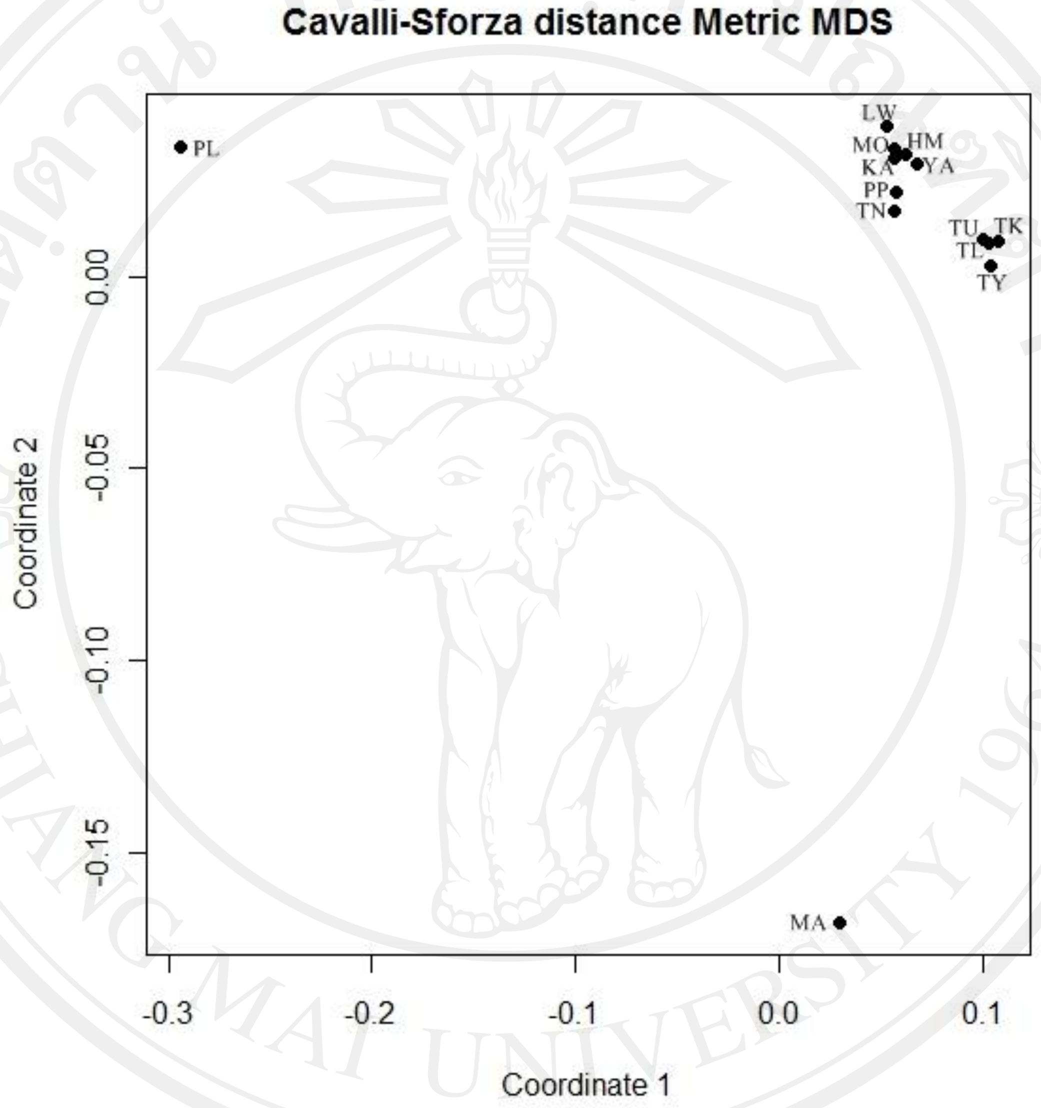

Figure 4.6 Two-dimensional graph from classical multidimensional scaling of 13 populations based on Cavalli-Sforza's distance