CHAPTER 2

IDENTIFICATION OF ANCESTRY INFORMATIVE MARKERS

2.1 Introduction

Human evolution can be traced via various forms of genetic information (Quintana-Murci et al., 1999; Jobling, 2001; Jobling and Tyler-Smith, 2003). Many studies involving mitochondrial DNA (mtDNA) (Quintana-Murci et al., 1999; Salas et al., 2002; Metspalu et al., 2004) and DNA variation on the Y chromosome (Jobling and Tyler-Smith, 2003; Karafet et al., 2008) reveal that the present human species is originated from Africa (Lewin, 1987). In fact, the migration of behaviourally modern humans from Africa to all continents takes place only approximately 70,000–50,000 years ago (Mellars, 2006). Through this course of migration, the population subdivision has occurred and has resulted in the emergence of new populations and ethnic groups.

With the presence of strong evidence that supports the occurrence of population subdivision, the genetic description of a population can be established. Furthermore, the clustering of individuals into many populations with different genetic backgrounds can be done automatically (Pritchard, Stephens and Donnelly, 2000; Falush, Stephens, and Pritchard, 2003, 2007; Gao and Starmer, 2007; Paschou et al., 2007). The task of assigning an unknown individual to the correct population can be carried out by inspecting his or her population-specific genetic patterns once the population boundary is defined via genetics or self-reported ethnicity. These patterns usually consist of ancestry informative markers (AIMs)—genetic markers that exhibit substantially different allele frequencies between populations of descendants derived from mutually inbred ancestors. The identification of AIMs has been proven to be beneficial to many research areas including genetic epidemiology (Enoch et al., 2006; Seldin, 2007; Tian, Gregersen and Seldin, 2008; Baye et al., 2009) and forensic science (Phillips et al., 2007; Budowle

and van Daal, 2008).

The international HapMap project discovers over 3,000,000 single nucleotide polymorphisms (SNPs) in the genome of each human individual (The International HapMap Consortium, 2003, 2005, 2007). As a result, the search for SNP-based AIMs usually involves genome-wide SNP screening. Many measures including informativeness (Rosenberg et al., 2003; Rosenberg, 2005), t statistics (Park et al., 2007; Zhou and Wang, 2007) and F statistics (Zhou and Wang, 2007) have been proposed for SNP prioritisation. The screening is then carried out via a greedy search (Rosenberg, 2005) or a ranking method (Zhou and Wang, 2007). Once SNPs have been selected, their capability as AIMs can be validated via classification model construction. The classification task specifically involves the use of genotypic attributes from selected SNPs as inputs for identifying the ethnicity or population label of an individual. Standard machine learning techniques that have been successfully implemented as classifiers include a support vector machine (Zhou and Wang, 2007) and genetic programming (Nunkesser et al., 2007). The same two-step protocol, which involves SNP screening and classification model construction, has also been successfully applied to genetic association studies (Moore et al., 2006).

Genome-wide SNP screening indicates that AIMs extracted from the HapMap data spread across the whole genome (Park et al., 2007; Paschou et al., 2007; Zhou and Wang, 2007). In fact, only 14 SNPs are required for the complete classification between three populations namely the CEU (Utah residents with northern and western European ancestry) population, the YRI (Yoruba in Ibadan, Nigeria) population and the Asian population obtained by merging the JPT (Japanese in Tokyo) and CHB (Han Chinese in Beijing) populations together (Paschou et al., 2007). However, 64 SNPs are needed for the near complete classification between all four HapMap populations, indicating that additional 50 SNPs are required for the classification between CHB and JPT populations (Paschou et al., 2007). This implies that large AIM panels are necessary when the classification task involves multiple populations which are closely

related to one another. In order to make the AIM identification task tractable for this kind of scenario, it is crucial to develop a protocol that leads to the discovery of the smallest possible AIM panels.

Early works on AIM identification are usually conducted by exploiting little prior knowledge regarding population subdivision. By incorporating the prior knowledge into the AIM search protocol, it is possible that the search can be limited to specific genomic regions. The regions that are strong candidates for this consideration are positive selection regions (Olson, 2002; Sabeti et al., 2002; Bamshad and Wooding, 2003; Akey et al., 2004; Vallender and Lahn, 2004; Voight et al., 2006; Sabeti et al., 2007). This is because one of the main signatures of positive selection is the decrease of heterozygosity over Hardy-Weinberg expectations (Beaumont and Balding, 2004), which also signifies population subdivision. The search for positive selection has been conducted on samples from many populations including European, African and Asian (Hinds et al., 2005; Myles et al., 2008; Oleksyk et al., 2008; Pickrell et al., 2009). The discovered selective regions spread across the whole genome and cover genes that govern growth, pigmentation, immune defence, carbohydrate metabolism, behaviour and other functions.

It has been suggested that SNPs from positive selection regions can be used as AIMs (Lao et al., 2006; Phillips et al., 2007; Seldin, 2007; Tian, Gregersen and Seldin, 2008). This is because an AIM from a positive selection region detected in a multiple population data set has a strong potential for being directly applicable as an AIM for other data sets containing similar populations. Evidence that supports this suggestion includes the selection of a SNP from *EDAR* as a member of an AIM panel for inferring ancestors of many common populations in the US (Kosoy et al., 2009). This gene involves in the development of hair follicles and has undergone positive selection in Asian populations (Sabeti et al., 2007; Bryk et al., 2008; Fujimoto et al., 2008; Mou et al., 2008). Nonetheless, an attempt to extract entire AIM panels from positive selection regions has never been made.

In this thesis, a protocol for identifying AIMs from potential positive selection regions is proposed. It is aimed that by concentrating the AIM search on potential positive selection regions, the resulting AIM panels should be smaller than those identified without the genomic region restriction. The proposed protocol involves three main steps: identification of SNPs in potential positive selection regions, SNP screening via attribute selection and classification model construction. Potential positive selection regions are located by means of F_{ST} extremity measurement (Bamshad and Wooding, 2003; Sabeti et al., 2007; Bryk et al., 2008; Fujimoto et al., 2008; Myles et al., 2008). SNPs with extreme F_{ST} values are subsequently screened by the most appropriate technique selected from a number of filter- and wrapper-based attribute selection techniques (Saeys, Inza, and Larrañaga, 2007) including a correlation-based feature selection technique (Hall and Holmes, 2003), a wrapper embedded with a naïve Bayes classifier (Kohavi and John, 1997), a simple symmetrical uncertainty (Press et al., 1988) ranking technique and a newly proposed round robin symmetrical uncertainty ranking technique. Finally, the classification model is constructed by a naïve Bayes classifier. The functionality of the proposed protocol is demonstrated via an application to the HapMap data.

2.2 Materials and Methods

2.2.1 Data Set

The data set explored in this study is obtained from the public release #23a of HapMap data set (Phase II, release date: March 2008), which is available in NCBI build 36 (dbSNP b126) coordinates. The data set consists of 3,619,209 SNPs in which the genotypic attribute value according to each SNP can be a homozygous wild-type, heterozygous or homozygous mutant genotype. These SNPs are extracted from 270 samples representing four populations: CEU, CHB, JPT and YRI. Both CEU and YRI data sets consist of 90 related samples—30 father-mother-offspring trios. In contrast, both CHB and JPT data sets contain 45 unrelated samples. Since the original HapMap

data set is composed of related and unrelated samples, only 210 unrelated samples are considered. The sample reduction is carried out by removing offspring samples from both CEU and YRI data sets.

2.2.2 F_{ST} Extremity Measurement

The decrease of heterozygosity over Hardy-Weinberg expectations due to population subdivision can be described by an F_{ST} measure (Wright, 1951). F_{ST} is defined by

$$F_{ST} = \frac{H_T - H_S}{H_T} \qquad \text{Eq.2-1}$$

where H_S is the average of expected heterozygosities over all populations and H_T is the expected heterozygosity in the combined population. H_S is given by

$$H_S = \sum_i d_i(2p_i(1-p_i)) \qquad \text{Eq.2-2}$$

where d_i is the proportion of the ith population in the combined population and p_i is the major allele frequency of the ith population. Similarly, H_T is denoted by

$$H_T = 2\bar{p}(1-\bar{p}) \qquad \text{Eq.2-3}$$

where $\bar{p}$ is the average of p_i over all populations and is equal to $\sum_i d_i p_i$. Since population subdivision causes a perceived deficiency of heterozygotes, an F_{ST} value is always between zero and one.

The search for SNPs with extreme F_{ST} values is proven to be useful for preliminary screening for positive selection (Bamshad and Wooding, 2003; Sabeti et al., 2007; Bryk et al., 2008; Fujimoto et al., 2008; Myles et al., 2008). An empirical distribution of F_{ST} is first estimated from either some or all of available SNPs in the recruited samples (Sabeti et al., 2007; Fujimoto et al., 2008; Myles et al., 2008). The F_{ST} extremity of each SNP is subsequently defined in terms of the percentile from the distribution. In this study, the F_{ST} distribution is calculated for every population pair using all SNPs in the HapMap data.

2.2.3 Simple Symmetrical Uncertainty Ranking

Symmetrical uncertainty is an information-theoretic measure discussed by Press et al. (1988). Consider a classification problem that involves a sample set in which each sample is described by n discrete-valued attributes (SNPs) and a class (population) label. Let A be an attribute and C be the class. The entropy H of the class before and after observing the attribute is given by

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c) \qquad \text{Eq.2-4}$$

and

$$H(C|A) = -\sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a), \qquad \text{Eq.2-5}$$

respectively where p denotes the probability value as estimated from the sample set. The difference between the entropy of the class before and after observing the attribute is the information gain (Quinlan, 1993) which is given by

$$\begin{aligned} \textit{Information Gain} &= H(C) - H(C|A) \\ &= H(A) - H(A|C) \\ &= H(A) + H(C) - H(A, C). \end{aligned} \qquad \text{Eq.2-6}$$

The degree of correlation between the attribute and the class can subsequently be estimated via symmetrical uncertainty (SU) which is defined by

$$\begin{aligned} SU &= 2 \times \left[\frac{H(A) + H(C) - H(A, C)}{H(A) + H(C)} \right] \\ &= 2 \times \left[\frac{H(C) - H(C|A)}{H(A) + H(C)} \right]. \end{aligned} \qquad \text{Eq.2-7}$$

It is noticeable that symmetrical uncertainty can be calculated from a quotient between the information gain and the sum of class entropy and attribute entropy. An attribute that has a high SU value is highly correlated with the class and is also an important attribute for classification. A rank can be assigned to each attribute according to its SU value where selected attributes are simply the top n_r attributes with the highest ranks.

2.2.4 Round Robin Symmetrical Uncertainty Ranking

From the previous section, it is noticed that SU can be directly measured in classification problems with the number of classes greater than or equal to two. However, the calculation of SU for a common SNP from two populations at a time is more useful for the identification of AIMs that are located in potential positive selection regions. This is because positive selection is generally confirmed when at least one new population is emerged from the ancestral population. For clarification, the measure is referred to as SU_2 when SU is evaluated to determine the suitability of using an attribute for the classification between two classes. After the SU_2 values have been derived from all attributes, a rank can be assigned to each attribute; high SU_2 values lead to high ranks. The top n_r attributes with the highest ranks are subsequently selected as screened attributes. For a multi-class problem, $\binom{m}{2} = m!/((m-2)!2!)$ sets of top-ranked attributes can be extracted from the data where m is the number of classes. The merging of top-ranked attribute sets is subsequently carried out where the size of the merged attribute set is between n_r and $n_r \times \binom{m}{2}$. The summary of round robin symmetrical uncertainty ranking (SU_2 ranking) is illustrated in Figure 2-1.

2.2.5 Correlation-Based Feature Selection Technique

A correlation-based feature selection technique (Hall and Holmes, 2003) is an attribute subset evaluation heuristic that considers both the usefulness of individual features (attributes) in the classification task and the level of correlation among features. Each attribute subset is assigned a score given by

$$Merit_F = \frac{n_f \overline{r}_{cf}}{\sqrt{n_f + n_f(n_f - 1)\overline{r}_{ff}}} \qquad \text{Eq.2-8}$$

where $Merit_F$ is the heuristic merit of an n_f-attribute subset F, $\overline{r}_{cf}$ is the average feature-class correlation and $\overline{r}_{ff}$ is the average feature-feature correlation. The correlation is obtained from the SU measure. An attribute subset receives a high merit score if it contains features that are highly correlated with the class and at the same time have low correlation among one another. An application of a best first search for the

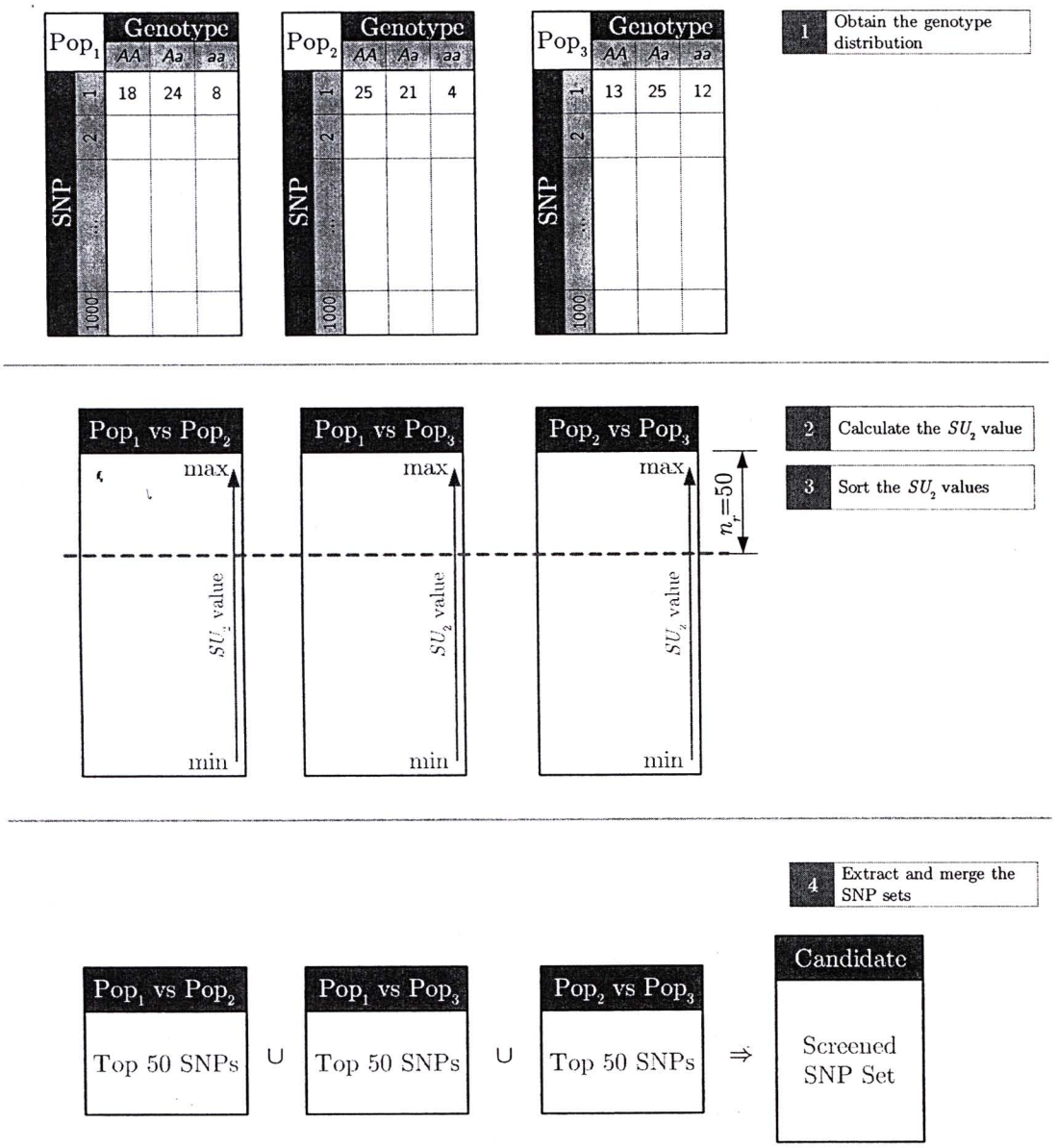

FIGURE 2-1 Outline of the SU_2 ranking. In this example, the three-population problem consists of balanced 150 samples and 1,000 SNPs. The genotype distribution of SNP_1 in all three populations is displayed. This leads to the SU_2 values of 0.016193, 0.009468 and 0.049025 for the population pairs (Pop_1, Pop_2), (Pop_1, Pop_3) and (Pop_2, Pop_3), respectively. After the calculation of SU_2 values for each SNP in every population pair is completed, SNPs are sorted according to their ranks. Three sets of top-ranked SNPs can be extracted from three population pairs. Only the top 50 SNPs are selected for each sorted set. The merging of three 50-SNP sets leads to the screened SNP set of size between 50 and 150.

best subset identification is carried out to avoid searching through all possible attribute subsets.

2.2.6 Wrapper

A wrapper refers to a category of attribute selection techniques in which the significance of an attribute subset is estimated from the resulting classification accuracy achieved by a classifier (Kohavi and John, 1997). In other words, the ability of a wrapper to identify necessary attributes or input features depends on the chosen classifier. Repeated five-fold cross-validation is implemented to provide an estimate of classification accuracy when an attribute subset is considered. Basically, the data samples are randomly divided into five folds where four folds of samples are used to train the classifier while the remaining fold of samples is used to test the classifier. The classifier training and testing procedure is carried out five times during one repetition of cross-validation where for each time a different sample fold is chosen as the testing fold. Hence, the samples in each fold are always used both to train and to test the classifier. Cross-validation is repeated as long as the standard deviation of classification accuracy over the repetitions is greater than one percent of the average classification accuracy or until the maximum of five repetitions is exhausted. The search for the best attribute subset is carried out via an application of a best first search and the chosen classifier for the wrapper is a naïve Bayes classifier.

2.2.7 Naïve Bayes Classifier

A naïve Bayes classifier is a classification system in which the prediction of the class output is based on the application of Bayes theorem (Mitchell, 1997). The naïve Bayes classifier can probabilistically predict the output class of an unknown sample using the available training samples to calculate the most probable output. The naïve Bayes classifier functions by assuming that the attribute values are conditionally independent given the output class. This assumption is particularly valid in this study because it is desirable to extract AIMs from different genomic regions, implying that the selected SNPs are most likely be uncorrelated.

TABLE 2-1 The number of SNP data partitions from each chromosome.

Chr	Number of SNP data partitions	Chr	Number of SNP data partitions	Chr	Number of SNP data partitions
1	59	9	35	16	21
2	63	10	40	17	17
3	48	11	38	18	23
4	46	12	36	19	11
5	47	13	30	20	23
6	52	14	23	21	10
7	41	15	20	22	11
8	41				

The number of partitions from Chromosome 2 is highest since there are more SNPs on this chromosome than other chromosomes. There are 735 partitions in total.

2.2.8 Implementation

The round robin symmetrical uncertainty ranking and F_{ST} extremity measurement programs are implemented in a C# programming language. The programs have been successfully tested for the execution under Windows operating systems. On the other hand, the simple symmetrical uncertainty ranking, the correlation-based feature selection technique, the wrapper embedded with a naïve Bayes classifier and the naïve Bayes classifier are available as parts of a WEKA package (Witten and Frank, 2005). All results included in the study are collected from the execution of the developed programs and WEKA in a personal computer. The computer is equipped with an Intel Core 2 Duo E6600 2.4 GHz processor and 2 GB of main memory. Windows XP is installed on the computer.

2.3 Results and Discussions

2.3.1 Benchmarking of Attribute Selection Techniques

There are many attribute selection techniques that can be used for AIM identification. A well-defined AIM panel should contain uncorrelated SNPs that lead to the highest population classification performance. Hence, a suitable attribute selection technique must be capable of extracting such an AIM panel from a SNP data set,

which contains both correlated and uncorrelated SNPs. Moreover, the computational time for the extraction process must be tractable. In this study, the candidate attribute selection techniques are the correlation-based feature selection (CFS) technique, the wrapper embedded with a naïve Bayes classifier (NB-Wrap), the simple symmetrical uncertainty ranking (simple SU ranking) and the newly proposed round robin symmetrical uncertainty ranking (SU_2 ranking). The classification performance is measured by applying selected attributes as inputs to a naïve Bayes classifier where ten-fold cross-validation is applied during the experiment. The values of n_r (number of top-ranked SNPs) for both simple SU and SU_2 ranking techniques are set to 50, 100, 200 and 300. Since the HapMap data contains a large amount of SNPs, the complete data set can be partitioned into a number of smaller data sets. Using multiple small data sets during the benchmarking of attribute selection techniques provides multiple results for statistical analysis. Moreover, it reduces the possibility of selecting false attributes that are unnecessary for the classification task (Park et al., 2007). Data partitioning is conducted on SNP data from every chromosome. Each partition consists of 5,000 positionally consecutive SNPs except for the last partition from each chromosome, which is allowed to contain less than 5,000 SNPs. The number of data partitions from each chromosome is summarised in Table 2-1. The search for AIMs is thus conducted by limiting the SNP inputs to those from the same partition. The distribution of classification accuracy obtained from all experiments is illustrated in Figure 2-2. It can be clearly seen that NB-Wrap produces the best screening result while CFS and the SU_2 ranking have the second best and third best results, respectively (a paired t-test on pair-wise algorithm comparison based on 735 experiments yields a p-value < 0.05). These results can be further interpreted as follows.

Generally, attribute selection can significantly improve the classification efficacy. Hall and Holmes (2003) have performed a benchmark test on a number of attribute selection techniques. Similar to the results from the current study, wrappers and CFS are also proven to be the best and second best techniques, respectively. This is deduced

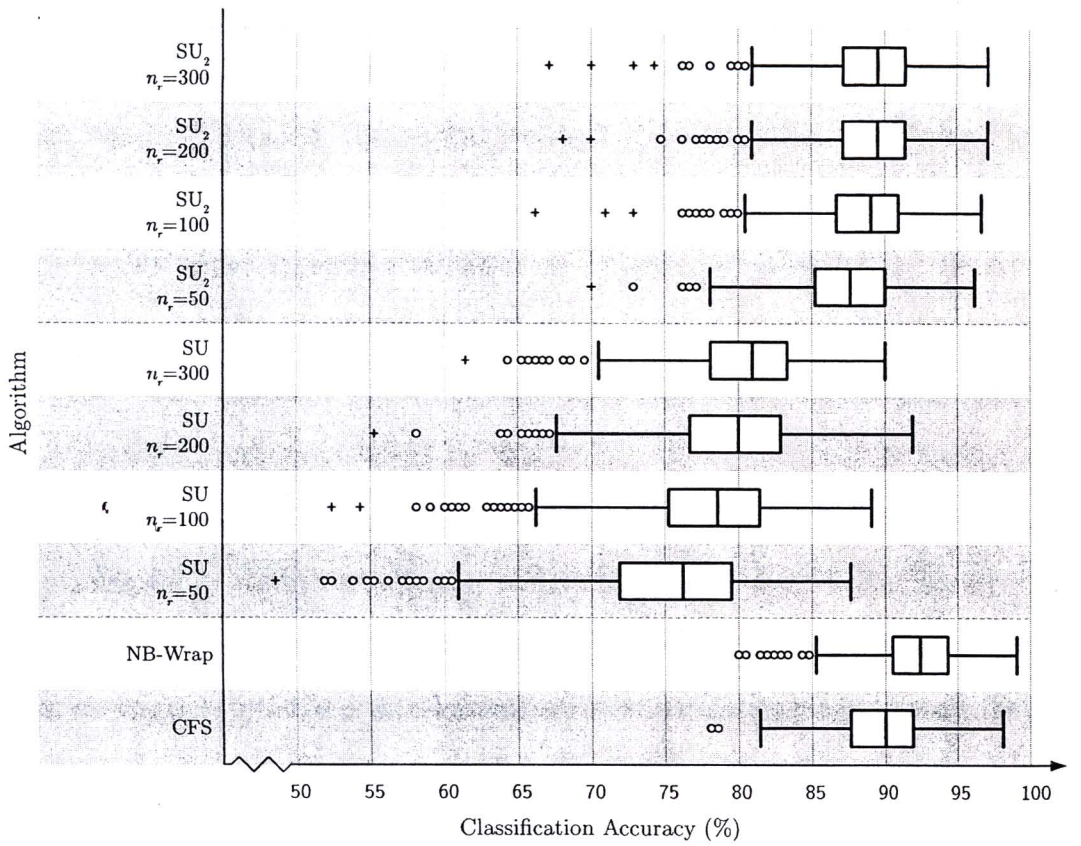

FIGURE 2-2 Performance of CFS, NB-Wrap, the simple SU ranking and the SU_2 ranking in conjunction with a naïve Bayes classifier.

from the overall classification performance across a range of benchmark problems in the UCI Machine Learning Repository in which the comparison is conducted by observing the performance of a naïve Bayes classifier before and after the attribute reduction. Wrappers and CFS appear to function well under moderate levels of attribute interaction. This is because both wrappers and CFS evaluate the significance of each attribute by considering the correlation between the attribute and the class while at the same time monitoring the inter-attribute correlation. Hence, a collection of attributes that together lead to high classification accuracy can often be conveniently identified.

The simple SU and SU_2 ranking techniques on the other hand consider only the correlation between each attribute and the class. Hence, the techniques are only able to identify the likelihood of an attribute being useful to the classification. In the absence

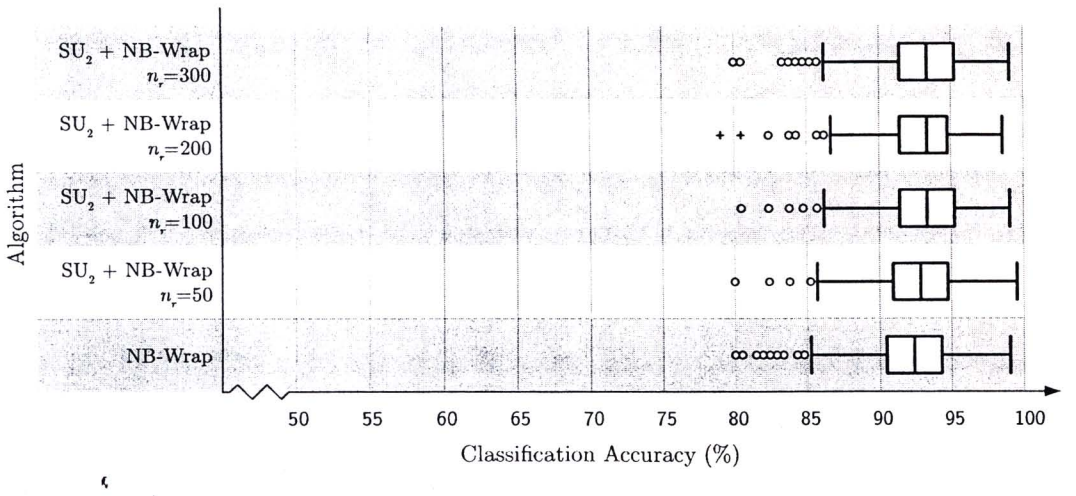

FIGURE 2-3 Performance of NB-Wrap and the two-stage approach, consisting of the SU_2 ranking and NB-Wrap, in conjunction with a naïve Bayes classifier.

of inter-attribute correlation monitoring, the presence of correlation among attributes can lead to performance degradation. The most probable source of attribute correlation is linkage disequilibrium, which exists among SNPs from the same localised region. The effect of linkage disequilibrium is most obvious when SNPs are screened by the simple SU ranking. Hastie and Tibshirani (1998) suggest that dividing a multi-class problem into a set of two-class problems can reduce the problem complexity. This approach has been adopted through the design and implementation of SU_2 ranking. By taking into account only a pair of populations at a time, each set of top-ranked SNPs generated prior to the set merging contains strong AIM candidates for separating two populations. Since a linkage disequilibrium pattern is specific to a population, a pattern difference is conveniently detectable when a pair-wise population comparison is conducted. This consequently leads to the reduction of linkage disequilibrium effect on the ranking mechanism as seen from the performance improvement exhibited by the SU_2 ranking over that from the simple SU ranking.

Although NB-Wrap produces the best screening result, its drawback is that a large computational effort is required to achieve this high performance. The computational time to finish the NB-Wrap calculation for each SNP partition on the computer is ap-

proximately 30 minutes while it takes less than one minute to complete the SU_2 ranking calculation. This is because the SU_2 ranking and NB-Wrap can tackle an n-attribute problem in linear and exponential time, respectively. Since the difference between the performance of both techniques is small, it is worth to explore the possibility of combining NB-Wrap and the SU_2 ranking. A similar two-stage approach for attribute selection has also been successfully applied in genetic association studies (Wongseree et al., 2009). Basically, the SU_2 ranking is first applied to the data. The screened SNPs are subsequently used as inputs for NB-Wrap. The classification accuracy is hence determined from a naïve Bayes classifier that takes inputs from the finally screened SNPs. Ten-fold cross-validation is still employed during the experiment. The distribution of classification accuracy in Figure 2-3 suggests that the two-stage approach is capable of maintaining the same level of performance achieved by NB-Wrap regardless of the n_r setting (a paired t-test on 735 experimental results yields a p-value > 0.05). Moreover, the two-stage approach with n_r = 50, 100, 200 and 300 leads to a reduction of computational time from 30 minutes to two minutes. This proves that the two-stage approach is highly suitable for AIM identification. Hence, the two-stage approach is selected for the attribute selection step of the AIM identification protocol. Since the n_r setting has no effect on the performance and the computational time of the two-stage approach, n_r = 50 is the chosen setting for the application of the AIM identification protocol to the genome-wide data in the next section.

2.3.2 Application of the AIM Identification Protocol to the HapMap Data

Candidate SNPs for inclusion in an AIM panel are SNPs from potential positive selection regions. These SNPs must have extreme F_{ST} values in which the F_{ST} extremity is estimated from empirical distribution. The empirical F_{ST} distribution calculated for every population pair using all SNPs in the HapMap data is illustrated in Figure 2-4. A similar F_{ST} distribution calculated from the HapMap data has also been reported (Fujimoto et al., 2008). The illustrated distribution describes different degrees of population subdivision for each population pair. For instance, the right tail

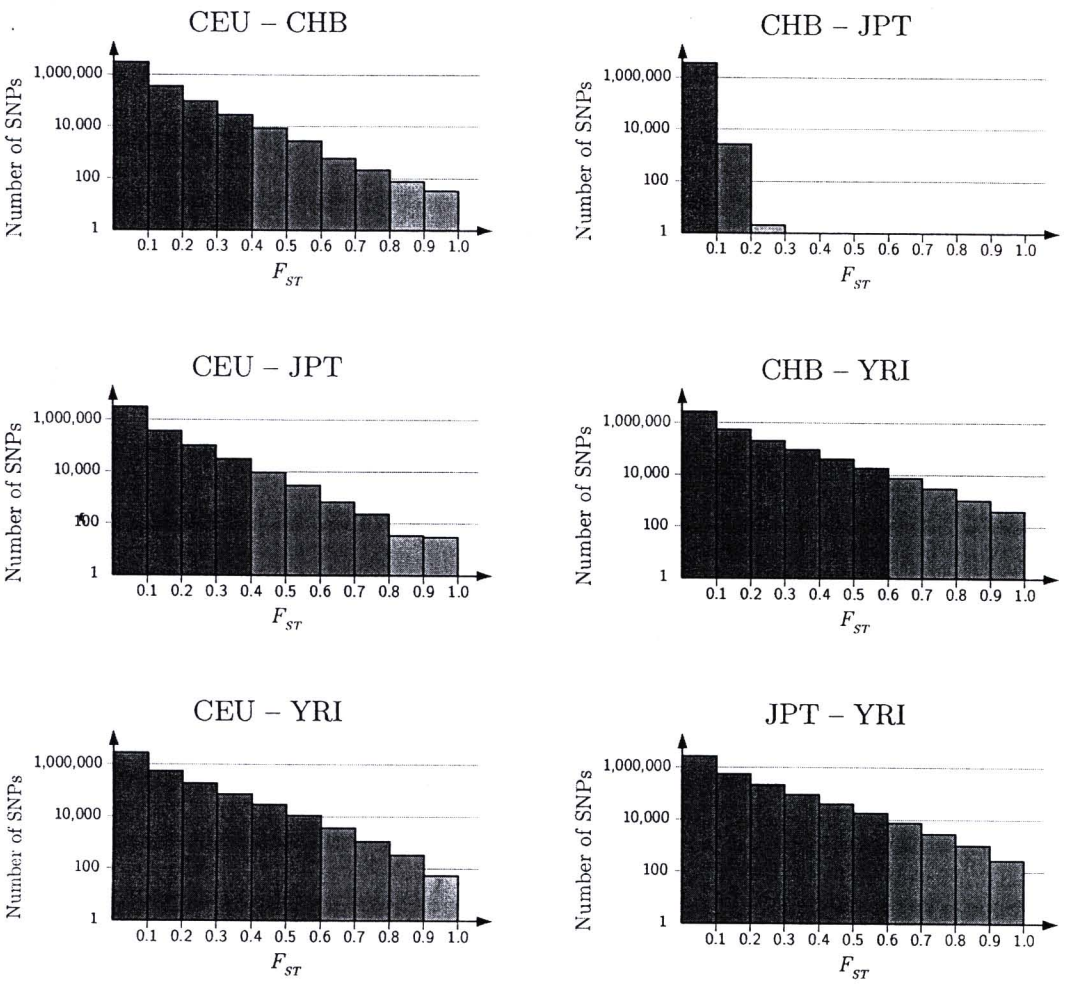

FIGURE 2-4 Empirical F_{ST} distribution for every population pair.

of the F_{ST} distribution for the CHB-JPT population comparison is located at a low numerical value, suggesting that these populations have recently begun to subdivide. On the other hand, the right tail of the F_{ST} distribution for the CEU-YRI, CHB-YRI and JPT-YRI population pairs is situated at a high numerical value. This implies that the emergence of newer populations from the African ancestors has taken place a long time ago. There are 31,465 SNPs with F_{ST} values in the top 0.3 percentile of the illustrated distribution. Each SNP possesses at least one extreme F_{ST} value among six F_{ST} values obtained from the pair-wise population comparison. SNPs with extreme F_{ST} values are subsequently screened by removing SNPs which are located neither inside nor close to

genes. The resulting candidate SNP set contains 13,328 SNPs within or near genes. A non-synonymous SNP set is also derived from the full candidate set and contains 230 SNPs. The full candidate and non-synonymous SNP sets are then subjected to two-stage attribute selection—the SU_2 ranking with $n_r = 50$ follows by NB-Wrap—and naïve Bayes classification.

The panel of 10 SNPs from the full candidate set given in Table 2-2 leads to complete classification between four populations. There are two SNPs which could signify positive selection: rs922452 and rs2269529. rs922452 is located on intron 4 of *EDAR*. This SNP and rs3827760 in the CHB samples are in linkage disequilibrium ($D' > 0.9$). rs3827760 is a non-synonymous missense SNP which is located on exon 12 of *EDAR* and causes a conservative substitution of valine by alanine. rs3827760 has been subjected to many investigations and is proven to be the source of positive selection of *EDAR* in Asian populations (Sabeti et al., 2007; Bryk et al., 2008; Fujimoto et al., 2008; Mou et al., 2008). The discovery of an extreme F_{ST} AIM in the proximity of rs3827760 conforms to the evidence that the average F_{ST} value across the *EDAR* SNPs is significantly higher than the genome-wide average F_{ST} value (Kelley et al., 2006). In contrast to rs922452, rs2269529 is a non-synonymous missense SNP which is located on exon 34 of *MYH9* and causes a conservative substitution of isoleusine by valine. The genotype distribution at this polymorphic locus suggests that positive selection may have occurred in the CEU population since the ancestral allele A is entirely replaced by the derived allele G. Moreover, an analysis of the HapMap data suggests that *MYH9* is located in a low heterozygosity region (Cheng et al., 2009). Nonetheless, further studies on the effect of rs2269529 on possible genotypic changes are required to confirm that positive selection has in fact occurred.

TABLE 2-2 Ten SNPs selected from the full candidate set containing 13,328 SNPs that have F_{ST} values in the top 0.3 percentile of the empirical distribution and lie within or near genes.

SNP	Population pair	F_{ST}	SNP location on the gene	Gene	Gene location on the chromosome
rs17408457	CHB-JPT	0.0974	Intron 7	*SLC30A7*	1p21.2
rs922452	CEU-CHB	0.9804	Intron 4	*EDAR*	2q11–q13
	CEU-JPT	0.7310			
	CHB-YRI	0.9063			
	JPT-YRI	0.6593			
rs12633912	CHB-JPT	0.1180	Intron 3	*FOXP1*	3p14.1
rs2693740	CEU-CHB	1.0000	Intron 1	*UBE2H*	7q32
	CEU-JPT	1.0000			
	CHB-YRI	1.0000			
	JPT-YRI	1.0000			
rs10758590	CHB-JPT	0.1704	Intron 2	*GLIS3*	9p24.2
rs3803464	CHB-JPT	0.0806	Intron 8	*MAN2C1*	15q11–q13
rs2238298	CHB-JPT	0.1100	Intron 18	*POLG*	15q25
rs2526371	CHB-JPT	0.1353	Intron 2	*RNF43*	17q22
rs6074677	CHB-JPT	0.1097	Intron 2	*MACROD2*	20p12.1
rs2269529	CEU-YRI	0.9355	Exon 34	*MYH9*	22q13.1

TABLE 2-3 Sixteen SNPs selected from 230 non-synonymous SNPs with F_{ST} values in the top 0.3 percentile of the empirical distribution.

SNP	Population pair	F_{ST}	Type of nonsynonymous missense SNP	Gene	Gene location on the chromosome
rs4915691	CEU-YRI	0.5428	Conservative	*DNAJC6*	1pter–q31.3
rs3795661	CHB-JPT	0.0918	Conservative	*ZNF697*	1p12
rs3827760	CEU-CHB	0.9247	Conservative	*EDAR*	2q11–q13
	CEU-JPT	0.6917			
	CHB-YRI	0.9247			
	JPT-YRI	0.6917			
rs1366842	CHB-YRI	0.6732	Non-conservative	*ZNF804A*	2q32.1
	JPT-YRI	0.8314			
rs482912	CHB-JPT	0.0790	Conservative	*LAMP3*	3q26.3–q27
rs2294008	CHB-JPT	0.1700	Non-conservative	*PSCA*	8q24.2
rs10989591	CHB-JPT	0.0932	Conservative	*GRIN3A*	9q31.1
rs284859	CHB-JPT	0.1049	Non-conservative	*C10orf26*	10q24.32
rs6265	CHB-JPT	0.0864	Conservative	*BDNF*	11p13
rs735295	CHB-JPT	0.1055	Non-conservative	*CCDC77*	12p13.33
rs2228224	CHB-JPT	0.0942	Non-conservative	*GLI1*	12q13.2–q13.3
rs9262	CHB-JPT	0.1060	Conservative	*C12orf29*	12q21.32
rs2273801	CHB-JPT	0.0806	Conservative	*WDR25*	14q32.2
rs2337127	CHB-JPT	0.1142	Non-conservative	*LOC100130736*	15q13.1
rs2236695	CHB-JPT	0.0940	Conservative	*PRDM15*	21q22.3
rs5764698	CHB-JPT	0.0942	Conservative	*SMC1B*	22q13.31

The panel of 16 SNPs from the non-synonymous SNP set given in Table 2-3 also leads to complete classification between four populations. Unsurprisingly, rs3827760, which is located in *EDAR*, is present in the panel. The obtained F_{ST} values support the presence of subdivision between CHB/JPT and CEU populations and that between CHB/JPT and YRI populations. This conforms to the evidence of positive selection of *EDAR* in Asian populations. In addition to rs3827760, rs1366842 is another SNP that indicates the subdivision between CHB and YRI populations and that between JPT and YRI populations. rs1366842 is located on exon 4 of *ZNF804A*, which is identified as a candidate gene for recent positive selection in Pima Indians (López Herráez et al., 2009). Both rs3827760 and rs1366842 are the only two SNPs in the panel that are useful for separating CHB and JPT populations from the other populations. In contrast, only rs4915691 is required to identify the subdivision between CEU and YRI populations. Since the genotype distribution in the CEU population at rs4915691 is similar to that at rs2269529 on *MYH9*, rs2269529 is not needed in this AIM panel.

The identification of non-synonymous missense SNPs, which are also AIMs, provides a direct correlation between possible phenotypic variations and ancestral origins. These possible phenotypic variations are the results of both conservative and non-conservative substitutions of amino acids. A non-conservative missense SNP is more likely to produce noticeable consequences since it causes the substitution of an amino acid with different properties. Nonetheless, the phenotypic effect of a non-synonymous missense SNP remains difficult to predict since it depends on how the substitution of an amino acid changes the structure and function of a protein (Hartwell, 2008).

2.3.3 Comparison with Other AIM Identification Techniques

In the present study, the genome-wide search for AIMs reveals that sets of 10 and 16 SNPs are sufficient for complete classification between four populations in the HapMap data. The sizes of AIM panels are at least four times smaller than those reported in the early works by Park et al. (2007), Paschou et al. (2007) and Zhou and Wang (2007). A summary of the sizes of AIM panels from the early works and the present

TABLE 2-4 Number of SNPs required for the classification of HapMap data.

Reference	Number of populations	Number of SNPs	Classification accuracy (%)
Present study	4	10	100.00
	4	16	100.00
Park et al. (2007)	3	82	100.00
Zhou and Wang (2007)	3	64	100.00
	4	100	90.00
Paschou et al. (2007)	3	14	100.00
	4	164	99.52
	4	64	98.57

The three-population problem is formulated by grouping JPT and CHB samples into the same class.

study is given in Table 2-4. Park et al. (2007) employ a nearest shrunken centroid method while Zhou and Wang (2007) develop a modified *t*-test for SNP screening. Both approaches are filter-based attribute selection techniques where each SNP is prioritised by identifying its usefulness for separating all population classes from one another. This is different from the strategy embedded in the SU_2 ranking in which each SNP is prioritised according to its usefulness for separating classes in each class pair. This strategic difference is most likely to be the cause of the reduction in the sizes of AIM panels from those reported in the works by Park et al. (2007) and Zhou and Wang (2007). Nonetheless, the strategy employed in the SU_2 ranking can be incorporated into both the nearest shrunken centroid method and the modified *t*-test. The modification should enhance the capability of both approaches, which could lead to the reduction in the sizes of AIM panels.

In contrast to Park et al. (2007) and Zhou and Wang (2007), Paschou et al. (2007) use a clustering technique to identify AIMs. In other words, the population labels are not considered during the SNP screening. As a result, larger AIM panels than those from the present study are selected to achieve the maximum distances between population clusters. Although the technique proposed by Paschou et al. (2007) may be less effective in the case of HapMap data, the technique is highly effective when the

population labels are not known a priori and the population boundary is determined solely via genetics.

In the present study, complete classification between four populations in the HapMap data is achieved using a naïve Bayes classifier. The predicted output class from the classifier is the class with the highest probability. This implies that if there are more than one class with equally high probabilities, a naïve Bayes classifier stands a chance of reporting a wrong class prediction. This means that the proposed AIM identification protocol in its present form may not be suitable to a classification problem that contains both ancestral populations and admixed populations derived from the ancestral populations. Nonetheless, the proposed protocol can be modified to accommodate this scenario by reporting each probability value for selecting a distinct class in the problem as the output instead of reporting only the output class with the highest probability. However, an additional criterion for determining the classification accuracy is also required.

The proposed AIM identification protocol relies on the ability to estimate F_{ST} extremity of each SNP in the data set. If the available SNPs do not cover enough genomic regions, the empirical F_{ST} distribution may significantly depart from the actual distribution. In addition to the constraint imposed by the F_{ST} extremity estimation, the number of populations in the classification problem also places a limitation on the functionality of the proposed protocol. This is because SU_2 values for each SNP are calculated for every pair-wise population comparison during the SU_2 ranking. Nonetheless, the computational time of SU_2 ranking is a quadratic function of the number of populations. This means that the computational time is still tractable for a reasonably large problem.

2.4 Conclusions

In this thesis, the identification of ancestry informative markers (AIMs) within potential positive selection regions has been conducted. The AIM identification proto-

col consists of three main steps: identification of SNPs with extreme F_{ST} values, SNP screening via attribute selection and classification model construction. SNPs are primarily screened according to their F_{ST} values. The F_{ST} extremity is estimated from the empirical F_{ST} distribution evaluated from all SNPs in the data set. SNPs with extreme F_{ST} values are subjected to further screening by two-stage attribute selection consisting of round robin symmetrical uncertainty ranking and a wrapper embedded with a naïve Bayes classifier. Finally, a classification model is built from the finally screened SNPs using a naïve Bayes classifier. Ten-fold cross-validation is applied during the AIM search. The proposed protocol is implemented and tested on the HapMap Phase II data set, which covers samples from four populations namely the CEU, CHB, JPT and YRI populations (The International HapMap Consortium, 2003, 2005, 2007). Two identified AIM panels are made up from lesser numbers of SNPs than those previously reported (Park et al., 2007; Paschou et al., 2007; Zhou and Wang, 2007). This suggests that a synergy between information extracted by data mining and that based on prior knowledge regarding population subdivision leads to more efficient AIM identification. The limitation of the proposed protocol and how it can be improved are also discussed.