GENETIC FEATURE SELECTION USING DIMENSIONALITY REDUCTION APPROACHES: A COMPARATIVE STUDY

by

Layan Imad Nahlawi

A thesis submitted to the School of Computing

In conformity with the requirements for

the degree of Masters of Science

Queen's University

Kingston, Ontario, Canada

(December, 2010)

Abstract

The recent decade has witnessed great advances in microarray and genotyping technologies which allow genome-wide single nucleotide polymorphism (SNP) data to be captured on a single chip. As a consequence, genome-wide association studies require the development of algorithms capable of manipulating ultra-large-scale SNP datasets. Towards this goal, this thesis proposes two SNP selection methods; the first using Independent Component Analysis (ICA) and the second based on a modified version of Fast Orthogonal Search.

The first proposed technique, based on ICA, is a filtering technique; it reduces the number of SNPs in a dataset, without the need for any class labels. The second proposed technique, orthogonal search based SNP selection, is a multivariate regression approach; it selects the most informative features in SNP data to accurately model the entire dataset.

The proposed methods are evaluated by applying them to publicly available gene SNP datasets, and comparing the accuracies of each method in reconstructing the datasets. In addition, the selection results are compared with those of another SNP selection method based on Principal Component Analysis (PCA), which was also applied to the same datasets.

The results demonstrate the ability of orthogonal search to capture a higher amount of information than ICA SNP selection approach, all while using a smaller number of SNPs. Furthermore, SNP reconstruction accuracies using the proposed ICA methodology demonstrated the ability to summarize a greater or equivalent amount of information in comparison with the amount of information captured by the PCA-based technique reported in the literature.

The execution time of the second developed methodology, mFOS, has paved the way for its application to large-scale genome wide datasets.

Co-Authorship

The research work presented in this thesis was performed under the supervision of Dr. Parvin Mousavi who offered general guidance and feedback, in addition to her efforts to review and provide modifications to the manuscript. However, the proposed methods and their implementations in the context of SNP selection were the original work of the author. All published methods and ideas belonging to other researchers were acknowledged and referenced according to the standard referencing practice.

The two proposed methods, ICA and mFOS, were published in the IEEE Engineering in medicine and biology conference EMBC' 2010 in two separate papers co-authored by Dr. Parvin Mousavi and written under her supervision:

1. Nahlawi, L; Mousavi, P: "Single Nucleotide Polymorphism Selection using Independent Component Analysis" Proceedings of the IEEE Engineering in Medicine and Biology conference pp. 6186 – 6189, 2010.
2. Nahlawi, L; Mousavi, P: "Fast Orthogonal Search for Genetic Feature Selection" Proceedings of the IEEE Engineering in Medicine and Biology conference pp. 1077-1080, 2010.

Acknowledgements

Before proceeding in acknowledging all whom deserve the acknowledgement, I humbly raise my praying hands to thank GOD, the mighty creator, who provided me with countless bounties and blessings and gave me the power to accomplish all what I have accomplished so far, and seek in the future.

My sincere gratitude is addressed to my lovely parents, Imad and Zeina, indeed the most precious to my heart. They constantly encourage and motivate me to excel in my studies and realize my dreams. They are the shelter I refuge to whenever obstacles burden my shoulders. No one has ever supported me in my decisions and provided unconditional guidance and help except my beloved parents. I would also like to express my sincere gratitude to my siblings, Adnan, Acile, and Mohammad and all my family who always nourish my soul with their continuous emotional feedback and support.

I am deeply indebted to my supervisor Dr. Parvin Mousavi and her devotion in guiding and assisting me throughout my studies. I am thankful to all her keen effort and long hours spent reading and correcting this manuscript as well as my other papers.

I am also grateful to my fiancé, Sharief Oteafy, who has always been a persistent supporter and inspirer in my research work. He had eagerly aided me in my writings and provided me with continuous encouragement. I also greatly appreciate Dr. Abd el Hamid Taha's effort in teaching me the essentials of research methods during the course I took with him. His criticism and feedback have immensely improved my technical writing skills.

Finally, I would like to thank all my friends, Gehan, Mervat, Mona, Samira, Shereen and Wisam who were always there for me; in support and sincere advice.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advent of high-throughput technologies, such as DNA microarrays, has paved the way for simultaneous measurement of hundreds of thousands of gene expressions and Single Nucleotide Polymorphisms (SNP). At the same time, analyzing and mining such data in order to identify fragments of informative features introduces major computational challenges. Informative genetic data dimensionality reduction and identification of significant processes in the data require thorough analysis. This thesis addresses the problem of SNP data dimentionality reduction by comparing the results of two well established methods in the signal processing literature; the Independent Component Analysis (ICA) and a modified version of Fast Orthogonal Search (mFOS).

1.1 Motivation

The analysis of SNP data is a key component of disease-gene association studies. With the aforementioned technological developments, high-throughput genotyping and sequencing techniques challenge researchers to analyze genome-wide sequence datasets with hundreds of thousands of SNPs. Due to the large size of these datasets, educated reduction of the number of SNPs is required in order to meet the computational demands of association studies. The possibility of performing dimension reduction without loss of information in the data, stems from the biological phenomenon of Linkage Disequilibrium (LD) [1].

LD describes the inhereted association of distant SNPs along a DNA sequence [6]. As a result of the nature of these associations, it is plausible for a small number of SNPs to summarize a significant amount of the information carried by the whole group of SNPs under observation. In other words, knowing the value of one SNP provides information about the value of its associated SNP. Hence, the information carried along a sequence of DNA can be compressed into a small group of SNPs which have the ability to predict the other SNPs in the studied sequence. A justifiaction of this predictive ability is the redundancy of information available in the group of associated SNPs. SNPs along the DNA sequence exhibit a form of redundancy in their information, which makes dimensionality reduction of SNP data a plausible process [2].

The need for dimensionality reduction techniques stems from the urgent need to perform disease-gene association studies. Several widespread and critical diseases such as cancer, multiple sclerosis and Alzheimer's are considered genetic diseases demonstrating associations with multi-genetic loci [3]-[6]. Identifying these loci raises the potential of early diagnosis of the studied disease, and enhances the ability to predict its course and developmental phases. Moreover, locating the association loci on the DNA sequence could potentially improve disease treatments by catering to customized therapeutic needs through personalized medicine.

Performing association studies on a genome-wide level requires a significant amount of resources, specifically computational power and processing memory, to be able to deal with the time and space complexity of the data. On the most basic level, an association study should inspect each mutation on the DNA sequence and look for a significant association with the susceptibility of certain diseases or the response to some disease treatment. However, this simplistic approach is computationally expensive due to the large number of SNPs, and is prone to missing combinatorial factors in the association studies. In addition, the number of false

positives in association studies performed using traditional statistical tests, such as Chi square and F- tests, is directly proportional to the size of the analyzed data. This proportionality is derived from the mere definition of statistical significance. Consequently, more sophisticated approaches which use complicated data mining techniques were developed, leading to a better inspection of the group of SNPs being studied. Unfortunately, these approaches significantly increased the computational overhead [7].

Dimensionality reduction, or more precisely, feature selection techniques, were adopted from data mining and signal processing literature to simplify the process of locating SNP loci associated with the studied disease [37]. This thesis discusses the application of two signal processing methods to reduce the dimension of SNP datasets. The following sections introduce the methods used and highlight the thesis contributions.

1.2 Feature Selection for SNP data

Feature selection techniques aim to reduce the number of dimensions in the studied dataset using the intrinsic characteristics or the correlation and association of each feature (in this case, each SNP) to a classification outcome. Methods focusing on the inherent characteristics of the SNPs are considered filtering techniques. Wrapper techniques are another group of methods that reduce the dimensionality according to associations between SNPs and class labelling available in the dataset. In other words, with a SNP dataset, the features are considered to be the actual SNPs and the feature selection process aims to reduce their number. This reduction depends on the ability to discard redundant information present among the SNPs.

Feature selection approaches select a group of SNPs carrying a sufficiently diverse amount of information, which effectively reflects the amount of information contained in the original group

of studied SNPs. The selected SNPs exhibiting non-redundant information are considered independent. The concept of the independence of features has led to the exploration of different statistical approaches designed to remove a group of SNPs which depend on other SNPs. In this thesis, two reduction techniques based on signal processing methods are adopted and applied to three different datasets.

The first method is the Independent Component Analysis (ICA); a matrix decomposition approach. ICA looks at a certain dataset as a group of mixed or combined signals. Its goal is to recover the mixing transform responsible for generating the observed signals, as well as the original unmixed independent signals known as the Independent Components (ICs). The mixing transform is usually a linear combination of the ICs. SNP values are considered mixed genetic signals where the redundancy of information and the dependence of specific SNPs on other SNPs are two main characteristics of the datasets. Hence, ICA is applied to the SNP datasets in order to recover the independent SNPs summarizing the information carried by the whole dataset. The ICs are calculated as linear combinations of the SNPs under investigation. Dimensionality reduction is performed by assessing the relevance of each SNP to a limited number of independent ICs.

The second method, mFOS, is a modification of FOS and can be described as a multivariate regression technique. Each studied SNP sequence in the dataset is regarded as an output of a genetic system aiming to express the information carried along the DNA as sequence of a group of selected SNPs. mFOS builds a model for the aforementioned genetic system by assessing the importance of each SNP in estimating the output of the system. This assessment depends on minimizing an error term of the estimated output in comparison with the original output

sequence. In summary, mFOS is applied to SNP datasets where it assesses the contribution of every SNP in the estimation of each genetic system output available in the dataset. Results of mFOS and ICA applied for SNP selection are compared with each other and with another dimension reduction approach, a second matrix decomposition method called Principal Component Analysis (PCA).

1.3 Thesis Objectives

This thesis aims to provide solutions for performing an educated dimension reduction without the loss of information carried along the studied SNP sequences. This goal of this thesis is achieved by proposing SNP selection techniques that:

- Reduce the dimension of a SNP dataset in order to facilitate and simplify the disease association studies.
- Perform the reduction by using only the intrinsic characteristics of the SNPs and without the need for a classification method or class labelling of the data.
- Reconstruct the entire sequence of SNPs from the selected group of SNPs with high accuracy.

1.4 Thesis Contributions

The main contributions of this thesis are as follows:

- Formulation of the ICA reduction technique as a filtering approach unlike the previously used application of ICA. In this thesis, ICA is presented as an unsupervised dimensionality reduction approach without requiring any class labels.
- Implementation of a modified version of FOS on SNP datasets in order to select the most informative SNPs.

- SNPs are chosen as the pool of candidate functions used for modeling SNP sequences using the mFOS method instead of using functions which are external to the system. This choice provides biological context and simplifies the interpretation of the results.
- Implementation of the SNP reconstruction technique based on CUR matrix decomposition in order to evaluate the selection results. The evaluation is reported according to the accuracy of reconstructing the studied dataset using only the selected SNPs.
- Re-implementation of PCA-based SNP selection in combination with the reconstruction technique for the purposes of comparison.

1.5 Thesis Structure

The remainder of this thesis is organized as follows: Chapter 2 provides the biological background of this thesis, a literature review of previous dimensionality reduction studies for SNP selection and previous applications of the ICA and mFOS methods. Chapter 3 presents the three datasets used in the thesis. In addition, it discusses the encoding approaches adopted for the transformation of the SNP sequences into numerical forms. Chapter 4 discusses the proposed dimension reduction techniques and their implementations. Chapter 5 reports the results of ICA for three SNP dataset as well as the results of mFOS on two SNP datasets and compares these results with previously published SNP selection results using PCA. Finally, Chapter 6 summarizes the work presented in this thesis along with ideas for future work.

Chapter 2

Background

In this chapter, a literature review of the biological foundation of this thesis work and the methodological background are provided. The chapter contains three sections. The first section presents the essential biological notions and concepts essential to the thesis. The second section surveys genetic data analysis literature for dimensionality reduction techniques and their different categories. The third section provides the fundamentals of the proposed methodologies, ICA and FOS, and reviews the literature for the application of these methodologies in similar contexts.

2.1 From DNA to Genome

2.1.1 The DNA

DNA (Deoxyribo-Nucleic Acid) is the reservoir of human genetic information. Genetic information is represented in the DNA in the form of a sequence of nucleotide base-pairs. A nucleotide base, or simply a base, is one of the following four molecules: Adenine (A), Guanine (G), Cytosine (C), or Thymine (T). The DNA exists in a double-stranded helix structure shown in Figure 2.1 where complementary nucleotides from both stands form the bonds between the two stands [1]. The complementarity between nucleotides is presented by the (A) - (T) bond and the (G) – (C) bond.

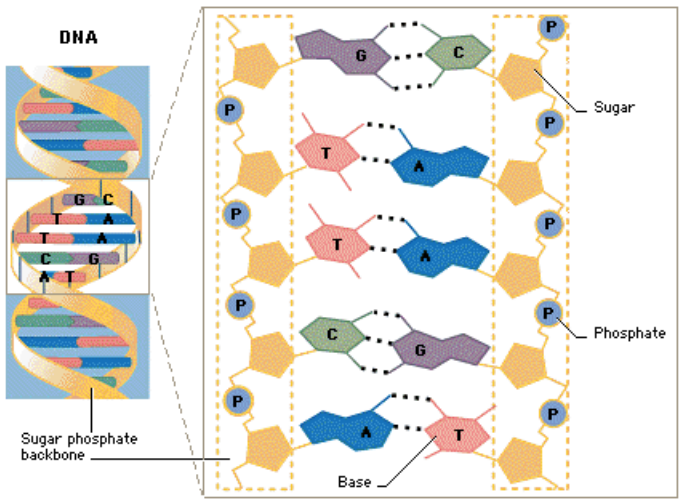

Figure 2.1 DNA double stranded helix structure shows the sugar phosphate backbone in addition to the different nucleotide bases. Picture adopted from: NUCLEIC ACIDS - DNA, RNA –GENETICS (http://biomolecules-world.blogspot.com/2008/12/deoxyribonucleic-acid-dna.html).

The DNA represents the coding material for all structural and functional proteins in the human body. Its sequence consists of a succession of coding (genes) and non-coding regions. The totality of the genetic material available in each human cell is known as the genome [9].

2.1.2 Genomic Diversity and Single Nucleotide Polymorphism

Despite the noticeable differences in human appearances, humans share more than 99% of their genome [10]. The diversities in the human DNA are due to chromosomal crossovers, which take place during reproduction, or external factors [13]. The differences generated by crossover are known as genetic polymorphism, whereas the differences resulting from external stimuli are called mutations [13]. In this research, we are interested in the analysis of a specific form of mutations, referred to as Single Nucleotide Polymorphism (SNP).

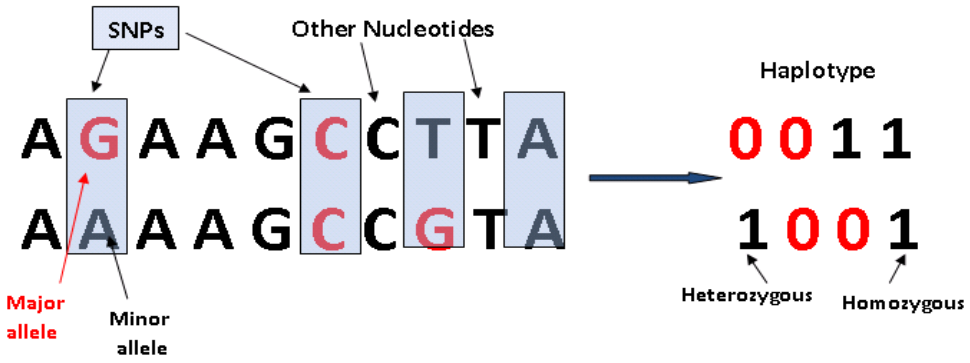

Figure 2.2 A Graphical explanation of SNPs and their related terms.

SNPs constitute 90% of the several different manifestations of mutations on human DNA [13]. A SNP is a mutation in a single base (Figure **2.2**), along the DNA sequence, which appears in at least 1% of the population [10]. SNPs have high frequencies; they occur every 100 to 300 bases [10] in either coding or non-coding genome regions. A SNP on a coding region may cause a modification in the coded protein, leading to abnormalities such as diseases and changes in responsiveness to drugs, among others. However, SNPs occurring in a non-coding region are considered as markers. As a result, SNP analysis is an essential component of studying disease-gene association

The majority of SNPs are bi-allelic, meaning that they have two possible values, with one value referred to as the major allele and the other referred to as the minor allele, according to their frequency of appearance in the studied population. Figure **2.2** demonstrates an example of a SNP. A SNP that has the same allele on both strands of the DNA is called homozygous while a SNP with different alleles on the DNA strands is called heterozygous [11]. A group of statistically related SNPs on a single chromosome is known as a haplotype [6], [11].

Technologies used to measure SNP values along the human genome, also known as genotyping technologies, have progressed rapidly in the last decade [12]. Recent ultra-high-throughput genotyping platforms are designed to assay for more than one million SNPs [12]. Currently, the available genotyping technologies are Affymetrix GeneChips©, Invader assays, Illumina's Infinium Beadchips©, and the Perlegen Genotyping Platform [12]. The Affymetrix© and Illumina© technologies are the most frequently used technologies due to their high accuracies in measurements [12]. Both of these technologies use direct hybridization of the assayed DNA, where DNA strands are associated with fluorescent complementary locus and allele specific probes [11]. The fluorescent intensities of the associated probes are collected and measured to designate the available alleles on the studied strands. However, the Affymetrix© technique hybridizes via locus and allele specific 25-mers of oligonucleotide probes and the Illumina© technique utilizes a bead array of 50 base-pair-long locus specific probes [12]. As a result of genotyping advancements, researchers are faced with ultra-large scale SNP datasets.

2.1.3 Genome Wide Association studies

Genome Wide Association Studies (GWAS) consist of studying the genome in a systematic manner to search for genetic variations (SNPs) associated with either a certain disease, or a certain response to a disease treatment [14]. In most reported GWAS, researchers use statistical tests such Chi-square, logistic regression and others to identify SNP associations with the disease or the studied trait [15]-[19]. Logistic regression, the most popular method used in association studies, is a predictive modeling technique. Its goal is to model the susceptibility or the response of a treatment of a certain disease as a log function of the SNP values available in the study. The identified SNPs provide researchers with the ability to better understand complex diseases,

perform early diagnosis and predict certain disease prognosis [20],[21]. Personalized medicine is one of the goals of GWAS; if achieved it can radically modify current healthcare practices and lead to improved disease treatments [22]. Genome Wide Association studies have been published for several complex diseases such as multiple sclerosis, cancer, and Alzheimer's disease among others [15],[23].

Seshadri *et al.* [23] performed a trend statistical test on genome wide SNP data and 35,000 individuals (case/control) in order to uncover the hidden association between specific SNPs and Alzheimer's disease (AD). They identified a strong association between the risk allele G of SNP rs2075650 and AD with a P-value $< 10^{-3}$. However, Kramer *et al.* [24] used a logistic regression to test the association of the genome wide SNP data and AD. They recognized a strong association between the disease and two SNPs: rs4298437 and rs11782819. On the other hand, Harold *et al.* [25] used a logistic regression on genome wide SNP data with 16,000 case/control individuals in an AD association study. They have replicated previously discovered associations between the disease and the SNP rs2075650. Moreover, they uncovered new associations with the disease and two other SNPs: rs11136000 and rs3851179.

Long *et al.* [26] conducted a multi-stage genome wide association study on 4,150 case/control individuals to discover SNP associations with breast cancer. Using logistic regression, they identified a strong association between breast cancer and the risk allele T of SNP rs4784227. Turnbull *et al*. [27] performed a breast cancer GWAS on 3,659 cases and 4,897 controls. Using the Cochran-Armitage trend test and Chi-square test, they reported associations between the cancer and five new SNPs on chromosomes 9, 10 and 11 with P-values ranging from 10^{-7} and 10^{-}

[15]. Furthermore, Thomas *et al*. [28] conducted a GWAS for breast cancer susceptibility on 1,145 cases and 1,142 controls by performing an unconditional logistic regression. They identified two new SNPs rs11249433 and rs999737 which were highly associated with breast cancer.

On the other hand, Nischwitz *et al*. [29] performed a GWAS using logistic regression to identify susceptibility loci for multiple sclerosis (MS). Their study included 590 cases and 825 controls and resulted in the identification of associations with the MS and SNPs on the two genes VAV2 and ZNF433. However, Sanna *et al*. [30] reported associations between MS and the SNP rs9657904 on the CBLB gene using conditional regression analysis. Their study was conducted on 882 cases and 872 controls with genome wide SNP data. Conversely, Jakkula *et al*. [31] conducted a GWAS for MS on 68 cases and 136 controls using a fisher exact test. They uncovered strong associations between the disease and SNP rs744166 on the STAT3 gene.

Hundreds of other studies analyzed associations with different diseases and reported new or replicated previously discovered results. This thesis is concerned with facilitating the use of techniques other than the statistical tests used in the above reported studies since larger GWAS datasets result in higher numbers of false positives in statistical test-based GWAS. In addition, some previously used statistical tests tend to miss combinatorial factors during association discovery, a fact that strongly affects GWAS results. Two approaches to systematically reduce the dimension of SNP datasets are provided in order to decrease the complexity of searching through DNA sequences for associations with diseases.

2.2 Dimensionality reduction for SNP Selection

By definition, dimensionality reduction refers to the identification of a lower dimension representation of a set of variables, which still captures the content and information available in the original representation of the variables [32]. In other words, given an *n* dimensional dataset, we seek to find a subset of *k* dimensions, where $k < n$, which contain all the information of the original data [32]**.**

The analysis of SNP data is key to disease-gene association studies [33]. The large size of these GWAS datasets requires a reduction of the number of SNPs in order to meet the computational requirements of association studies.

Achieving the desired reduction in the number of SNPs consists of searching within the SNP space and choosing a group of representative SNPs that summarize the information hidden in the entire sequence, while best predicting the non-selected SNPs in the studied sequence. Predicting the non-selected SNPs, residing on the haplotype in a study serves as a method of evaluating the results of a SNP selection approach. In bioinformatics jargon, this task is known as SNP selection or tagging SNPs. Tagging SNPs or htSNPs are equivalent terminologies indicating the set of chosen SNPs resulting from a SNP selection approach. HtSNPs have the ability to predict other SNPs in the haplotype. The non-selected SNPs are referred to as tagged SNPs and are not able to predict any other SNP on the chromosome.

SNP selection imposes algorithmic challenges, especially in terms of the nature and the complexity of search algorithms. Various dimensionality reduction approaches have been used to perform the selection of the most informative SNPs. These approaches can be categorized in two main groups: wrapper, and filter approaches, which are discussed separately in the following two subsections.

2.2.1 Wrapper approaches for SNP Selection

Wrapper techniques search through attribute (SNP) space for a group of SNPs that are capable of accurately classifying the data into the desired classes [37]. In other words, the dimensionality reduction is performed by assessing the relevance of each SNP to class attributes such as diseases or drug responses. Since data used in wrapper techniques require labeling and/or classification, they are regarded as supervised dimensionality reduction approaches.

Wrappers have been used for SNP selection from different SNP datasets with various classifiers. Shah *et al.* [43] proposed two wrapper approaches. The first one consisted of weighted decision trees which assigned SNP data into two classes good or bad responders for drug or placebo treatment. Different trees were weighted according to their SNP prediction and subject classification accuracies. The number of genes used in this study was reduced to 10 and 8 genes out of 32 genes for the drug and placebo datasets, respectively. Consequently, SNPs contributing to the highly weighted trees were selected. The second wrapper method in [43] selected SNPs based on a genetic algorithm. This algorithm evaluated sets of SNPs according to a fitness function reflecting their correlation with a desired classification outcome. The reduction in SNPs resulted in selecting 63 and 59 SNPs out of 172 SNPs for the drug and placebo datasets, respectively. Similarly, Yang and Zhang [44] used a genetic algorithm to select different sets of SNPs. Their fitness function was based on the classification accuracy of neural networks. Their method was applied to 10 SNPs belonging to 146 case/control dataset for Age-Related Macular Degeneration (AMD). Using this method, they reduced the number of studied SNPs to 4 SNPs capable of accurately classifying the data.

Zhang *et al.* [45] selected a set of SNPs based on their classification accuracy using a support vector machine. Dawy *et al.* [46] used Independent Component Analysis (ICA) to model SNP

expressions. Inspired by the blind source separation problem, they considered the SNPs to be transformed by unknown means to form some SNP expression which produces a phenotype or disease. They measured the relevance of independent components to the classification outcome of case/control data by applying a linear least square regression. Long *et al*. [47] presented two wrapper techniques. The first SNP selection technique was performed by choosing the SNPs which contributed to the distinction of different mortality means classes. The contribution was measured according to the information gained by adding a SNP to the classification model. Therefore, the higher the information gain of a SNP, the higher the association of that SNP with the classification attribute and the greater its ability to differentiate among the class values.

The second wrapper technique in [47] consisted of searching through subsets of SNPs and evaluating their ability to classify the studied data according to their cross-validation classification accuracies. The authors compared two search algorithms: Forward Selection (FS) and Backward Elimination (BE), and three classification methods: Naïve Bayes (NB), Bayesian Network (BN) and Neural Network (NN). FS begins with an empty set of SNPs and iteratively assesses the classification accuracy of each of the three above methods by adding one SNP at time. BE begins with the full set of SNPs and proceeds with the evaluation of SNP accuracies by eliminating one SNP with each iteration. The stopping criterion for both search algorithms is reached when the addition or elimination of a SNP does not cause a significant change in the classification accuracies. The above mentioned approaches were performed on separate chromosomes belonging to the Collaborative Study on the Genetics of Alcoholism (COGA) dataset which contains 894 individuals and 10,070 SNPs. Both methods reached an overall reduction of 264 SNPs.

Liu *et al.* in [48] developed a supervised method, namely Recursive Feature Addition (SRFA), to select the tag SNPs according to their contribution to the classification of individuals in the data using Support Vector Machines (SVMs). The authors in [48] presented their selection algorithm in three steps: First they ranked all SNPs according to their classification accuracy from the highest to lowest; then they selected the SNPs with highest accuracies; finally, each SNP was assessed according to its statistical similarity to the previously chosen SNPs as well as the increase in classification accuracy caused by the addition of the SNP.

Since the majority of the publicly available SNP datasets lack class labels, and since wrapper approaches require labels to perform SNP selection, filtering techniques are better suited for SNP selection. Furthermore, the use of wrappers is computationally more demanding due to the need to incorporate classification techniques.

2.2.2 Filter Approaches for SNP Selection

Filters or filtering techniques reduce the dimension of a dataset by assessing the relevance of attributes according to their intrinsic properties such as the correlation between attributes [37]. This kind of dimension reduction is performed without the need for labeling or classification of the data. Using machine learning terms, filters are considered unsupervised approaches. As previously mentioned, SNPs exhibit genetic associations that are inherited from one generation to the next. These associations are the intrinsic characteristics of the SNP data which enable data filtering.

Many filtering approaches have been adopted in the literature for SNP selection. Lee & Shatkay [38], have implemented a multi-stage SNP tagging technique using Bayesian networks (BN). BNs capture the probabilistic conditional dependencies among different SNPs using a directional acyclic graph. Directions of dependencies in the BN graph sort the studied SNPs in

topological order revealing a parent/child association among SNPs. Their technique starts with an empty set of htSNPs. SNPs are added to the htSNP set by assessing their relevance in predicting other SNPs in the dataset. The studied SNPs are assessed one at a time, following the order of the SNP topological sorting. They calculate the prediction accuracy of all SNPs using a heuristic sequential search through the sorted SNPs. All SNPs having prediction accuracies less than a certain predefined threshold are added to the list of htSNPs. The stopping criterion of their search is reached when either all nodes are visited or the prediction ability of the current htSNP set is higher or equal to an accepted threshold.

Lin and Altman [36] adopted a feature similarity measure as the criterion to discard redundant SNPs conditioned by the level of information loss. They used the Linkage Disequilibrium (LD) measure r^2 which reveals the correlation and similarity between two SNPs depending on the probability of major and minor alleles on these SNPs. Using r^2, they defined their distance measure, D, equivalent to $1 - r^2$. The higher the distance D, the more dissimilar the two SNPs are. Hence, the algorithm in [36] calculates the distance D between each SNP and its nearest dissimilar SNP. Afterwards, the SNP selection proceeds by finding the SNP with minimum distance to its neighboring SNPs and discarding the SNP's closest neighbor. Thus, their choice of the discarded SNP is based on minimum information loss. The algorithm continues by choosing a tighter neighborhood of SNPs and recalculates distance to discard the most similar SNPs neighbors. This filtering technique stops when the inspected dissimilarity neighborhood is bound only by adjacent SNPs.

Halldorsson *et al*. [39] devised a three step framework for filtering SNPs. The first step consists of building clusters of 13 to 21 consecutive SNPs. The second step measures the level of

"informativeness", I, between SNPs in a cluster. The measure indicates how well a set of SNPs can predict another SNP in the same cluster. "*I*" is defined in the following equation:

$$I = \frac{\#\ of\ haplotypes\ with\ different\ allele\ values\ at\ the\ predicted\ SNP}{\#\ of\ all\ allele\ pairs\ at\ the\ predicted\ SNP}$$

In the third step, SNP selection is performed by optimizing the informativeness measure inside each cluster achieved by searching for the best subset of SNPs that can predict other SNPs in the same cluster.

Halperin *et al.* [40] filtered SNPs by minimizing the tagged SNP prediction errors. To predict a biallelic SNP s with a/b as major/minor alleles, they used the two closest tagging SNPs to s on the haplotype, and SNPs $r_1\ and\ r_2$ with $a_1/b_1\ and\ a_2/b_2$ as their major/minor alleles. They calculated the conditional probability of $P_1(s\ to\ have\ a\ or\ b\ |r_1 with\ a_1 or\ b_1)$ and $P_2(s\ to\ have\ a\ or\ b\ |r_2 with\ a_2 or\ b_2)$ and performed a majority vote according to the calculated probability to decide on the predicted value for SNP s. Then, each SNP used in the prediction is given a score depending on its prediction error. Finally, the selection technique used dynamic programming to select the tagging SNPs responsible for minimizing the total prediction error, depending on the calculated scores.

In a later publication, Lin and Altman [41] implemented a filtering technique using Principal Component Analysis (PCA). Initially, the correlation matrix of the SNP dataset was decomposed into eigenvectors. Afterwards, SNPs were chosen according to their contribution to eigenvectors that explained a high percentage of the variance in the dataset (details to follow in section 4.4).

Chuang *et al*. [42] adopted a Genetic Algorithm (GA) with a K-nearest neighbor (KNN) evaluator in order to select the most informative SNPs. The chromosomes used for the GA iterations consist of a sequence of zeros (for non-selected SNPs) and ones (for selected SNPs). The lengths of these chromosomes are equal to the number of SNPs in study. All chromosomes

are randomly initialized. Each chromosome is evaluated using the KNN fitness function. Using the hamming distance between SNP loci, the fitness function locates the three nearest neighbors for each selected SNP. Prediction accuracy errors are then calculated for the different values of SNP alleles. SNPs causing minimal prediction accuracy errors are selected for the next iteration of the GA. Finally, the remaining steps of the genetic algorithm: selection, crossover and mutations are performed to prepare the next generation to be evaluated by the KNN fitness function.

In this thesis, both of the proposed SNP selection techniques, ICA and mFOS are categorized as filtering methods.

2.3 Independent Component Analysis

Herault and Jutten [49] were the first to introduce the concept of Independent Component Analysis. They discussed the process of extracting source signals from a group of captured mixed signals. The challenge was to be able to recover the mixed signals without any knowledge of the original signals or even the mixing model. The problem was discussed as a temporal or spatial analysis for neural networks and the term ICA was not used. Later in [50], the terminology was established, and an efficient algorithm was proposed to solve the source separation problem.

Independent Component Analysis, as defined in [50] and [51], is a linear transformation minimizing the statistical dependence among the extracted components. According to Fodor in [52], ICA generates linear projections under the condition of being as statistically independent as possible. ICA makes use of higher-order statistics to fulfill the previously mentioned condition. It is essential to differentiate between statistical independence and uncorrelatedness. Two random variables, x_1 and x_2, are uncorrelated if they have zero covariance *(Cov)* calculated as follows:

$$Cov(x_1, x_2) = E(x_1, x_2) - E(x_1)E(x_2) = 0 \qquad (2.1)$$

where $E(x_i)$ is the expected value of the random variable x_i and $E(x_i, x_j)$ is the expected value of the combined two random variables x_i and x_j.

Unlike statistical independence, uncorrelatedness depends only on second-order statistics measured by the covariance. Statistical independence is the ability of a multivariate probability distribution f with n dimensions to be factored into its constituting univariate distributions [52] as shown in the following equation:

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n) \qquad (2.2)$$

In the literature, there exist other linear transforms such as Principal Component Analysis (PCA) [53] which imposes Gaussianity on the analyzed data and extracts only uncorrelated components. Gaussianity is the characteristic "bell shape" probability distribution also known as a "normal distribution". Yet, one of the interesting characteristics of ICA which motivated its adoption for this thesis is the absence of a Gaussianity assumption for the data, since most of the biological data analyzed do not possess a Gaussian distribution. Achieving the strict definition of statistical independence stated in Eq. (2.2) is not feasible due to the lack of a linear transformation capable of performing such factorization [51]. Therefore, ICA transforms the data into components that exhibit independence as close as possible to the above mentioned condition. In order to maximize this statistical independence in the extracted components, several objective functions can be selected for optimization (minimization or maximization). The choice of the objective function depends on the implied characteristics of the sought independence, such as negative entropy, non-Gaussianity (measured by kurtosis) and mutual information [51]. The ICA approach consists of two major autonomous tasks. The first task is the selection of the desired independence measure and accordingly, the formulation of the objective function. The second

task is the selection of an optimization algorithm to maximize or minimize the previously selected objective function, leading to the maximization of component independence. Gradient ascent and descent methods are examples of optimization methods used in the literature [52].

Before surveying the different applications of ICA, we formulate the problem solved by ICA. Given a group of observed signals, ICA extracts the source signals, also known as Independent Components (ICs), exhibiting the least redundancy in information and the most independence among themselves. ICA calculates the coefficients of the transformation matrix whose application to the source signals leads to the recovery of the observed signals.

ICA has been used for several different applications in a variety of disciplines. It has been implemented for Blind Source Separation (BSS), feature extraction and feature selection. For BSS, ICA interprets the transformation matrix as a mixing matrix, causing the independent source signals to be perceived as mixed dependent signals.

In this context, the authors of [54] and [55] used ICA to analyze financial stock market data in order to recover the main factors affecting changes in stock prices. This analysis considers the gathered signals to be mixed with noise and other uninformative signals that camouflage the main factors that affect the stock market. Another application of ICA as BSS is in brain and heart signal analysis, for example Electroencephalogram (EEG) and Electrocardiogram (ECG) analysis. The authors of [56] and [57] used ICA to separate the ECG and/or noise signals from the collected EEG in order to achieve improvements in the quantitative EEG analysis. Lotsch *et al*. [58] applied ICA on a time series of normalized difference vegetation index (NDVI) imagery with the purpose of separating spatial and temporal manifestations of climatological and ecological processes. They segregated seasonal and non-seasonal signals using ICA in order to eliminate noise signals caused by instrumentation artifacts.

For feature extraction, ICA handles the terms of the transformation process differently. The columns of the transformation matrix are considered as features extracted from the observed signals. In addition, the matrix of source signals represents the coefficients of corresponding observed features. Hung *et al*. [59] used ICA to extract left and right brain map activity during the simulation of left or right finger lifting. These extracted features provided learning materials for a brain-computer interface (BCI). However, Serdaroglu *et al.* [60] extracted independent features from images of textiles. Their features contained mutually independent pixels and were used in detecting defective fabric textiles.

Feature selection is the final category of ICA applications in this review, and the approach is used in this thesis. The aim of this ICA application is to select the most informative features in the observed data, where a feature is an actual measured attribute in the data. Informativeness is designated by the contribution of an attribute in the calculated ICs. Dawy *et al*. [46] used this application of ICA to select the in most informative SNPs in the studied dataset as a wrapper approach (see section 2.2.1) . In addition, in this thesis, ICA is applied to SNP data for feature selection; the details of this application are presented in section 4.1.

2.4 Fast Orthogonal Search

Orthogonal search (OS) was originally devised for the identification of non-linear models for time series data [61]. The OS approach models non-linear systems by estimating their outputs. It searches through several orthogonal functions and assesses their relevance in minimizing the square error between the estimated and actual outputs [62]. This estimation is based on a multivariate regression scheme that aims to approximate the output by searching in a pool of candidate functions, and picking ones that generate a best fit for this output.

Given a system with a vector of outputs presented as a matrix X, where rows represent outputs and columns represent different time points, OS applies a QR transformation to X, where Q is the calculated orthogonal matrix [63]. The generation of matrix Q entails the orthogonalization of the different outputs in X. In other words, OS iteratively calculates the projections of the group of functions modeling an output into an orthogonal space using the Gram-Schmidt orthogonalization method [63], [64]. These orthogonal functions span the same space as the vector of outputs. The OS approach selects the terms of the output model by assessing the relevance of each orthogonalized function in reducing the model estimation error [64]. The most important implication of this orthogonalization process is the acquired independence of the orthogonal functions contributing to the output model. Hence, the selected model reflects a set of independent factors (functions) capable of carrying the information hidden in the output vector. The ability of OS to describe system outputs was the motivation for applying OS in different research areas.

The major drawback of the application of OS, in many research areas such as in biological time-series data analysis, is the increased computational complexity of orthogonalizing a larger number of candidate functions. Therefore, Korenberg [65] improved the OS method by decreasing the needed number of calculations in each iteration of the OS algorithm. The improved OS is significantly less time consuming as well as less demanding on memory, therefore it was named the Fast Orthogonal Search (FOS) [65]. FOS avoids the actual construction of the orthogonal functions by only calculating the correlation coefficients of the orthogonal functions and the chosen candidate functions.

FOS has been used to model systems in a wide spectrum of research areas. Korenberg *et al.* [66] demonstrated the ability of FOS to model temporal biological signals such as electroencephalogram (ECG) and electromyogram (EMG) data. However, Minz *et al.* [67] implemented FOS on molecular data by searching promoter regions of genes on DNA to identify specific motifs, and therefore build cooperative gene network models for the yeast cell-cycle [67]. Mostafavi *et al.* [68] incorporated FOS within a multivariate feature selection and classification framework to predict the response of multiple sclerosis patients to therapy, from gene expression data [68].

As feature selection is the main focus of this thesis, a modified version of FOS was applied to model the information carried by SNPs along DNA sequences. This modeling technique searches through the list of SNPs to identify independent SNPs capable of relaying the hidden information in a haplotype with high accuracy. The proposed modification of FOS is to accommodate predicting the hyplotypes directly using SNPs. This introduces a challenge where the dimensions of the output to be modeled are different from the candidate functions. Another approach for SNP selection would have been to use the candidate functions to predict other SNPs. Here, the former approach for SNP selection was used; this approach is discussed in detail in section 4.2.

Chapter 3

Single Nucleotide Polymorphism Data

This chapter presents the datasets used for the proposed feature selection approaches. In addition, it explains the data encoding techniques used prior to the application of the dimensionality reduction techniques. Three Single Nucleotide Polymorphism (SNP) datasets are used, where each dataset corresponds to one of the three genes: ACE, ACBC1 and IBD5q3 [41][1].

3.1 ACE Dataset

The ACE SNP dataset was originally published by Rieder *et al*. in [69] and it contains 78 un-phased SNP data for 11 individuals. These SNPs belong to the gene DCP1, which is extended over a 24 kb (kilo-base) long genomic region. This gene encodes for the Angiotensin I Converting Enzyme (ACE). ACE is a peptide contributing to the regulation of blood pressure by narrowing the blood vessels. It is also responsible for balancing fluids in the human body [70]. This gene has been included in many disease-gene association studies because of its important role in the renin-angiotensin system; a hormone system that regulates blood vessel pressure [69]. Several of these studies have revealed strong associations between the DCP1 gene and crucial cardiac functions, as well as the treatment of hypertension [70]. In addition, certain mutations on the DCP1 gene increase the susceptibility to cardiovascular diseases [69].

[1] All three gene datasets were used in [41] and provided by Dr. Zhen Lin, a previous member of the Helix research group at Stanford University.

The dataset, on which the selection approaches were applied, is a phased and trimmed version of the original. The phasing of the dataset was performed by determining the maternal and paternal alleles and segregating their values in two separate sequences. Trimming consisted of localizing the analysis on a specific genomic region smaller than the whole assayed region. The ACE phased dataset contained 52 bi-allelic SNPs and 22 phased SNP sequences for the original 11 individuals [41].

3.2 ABCB1 Dataset

The ABCB1 dataset originated from a study published by Kroetz *et al*. [71]. The dataset originally contained diploid (unphased haplotypes) genotypes for 48 SNPs available on the ABCB1 gene in 247 individuals. The ABCB1 gene encodes membrane-associated proteins which are members of ATP-Binding Cassette ABC transporters [73]. For instance the ABCB1 gene encodes the P-glycoprotein [72] which is a membrane protein arbitrating multidrug resistance [74], 56].

This dataset was phased by Kroetz *et al.* and Lin *et al.* [71], [1] where individual haplotypes were prepared for analysis. The phased dataset contained 27 bi-allelic SNPs for 484 haplotypes. This phased dataset was used in the dimensionality reduction application contained in this thesis.

3.3 IBD 5q3 Dataset

The IBD 5q3 dataset originated from an inflammatory bowel disease (IBD) study of father-mother-child trios by *Daly et al.* [76]. It represents diploid genotypes for 103 SNPs on chromosome 5q31 (IBD5 region) for 387 individuals [77]. The study of Daly *et al.* aimed to provide a high resolution analysis of the genomic region on the 5q31 chromosome. Many linkage

disequilibrium studies associate variations on this chromosomal locus with the inflammatory bowel disease (IBD) [77]. IBD is a chronic autoimmune disease characterized by relapsing symptoms of gastrointestinal tract disorders [79] . In addition, IBD 5q31 has been associated with Crohn's disease (CD) which is another digestive tract disorder causing the inflammation of the gastrointestinal tract [78].

The Daly *et al.* IBD 5q31 dataset was phased by Lin et al. in [41] into 774 haplotypes. The phased version of this dataset which consists of 103 SNPs in 774 haplotypes was used in the application of methodologies in this thesis.

3.4 Data Encoding

SNP data consist of sequences of nucleotides representing the mutations on a specific genomic region for the group of individuals under investigation. Each studied dataset is represented by a set of records (individuals) characterized by a set of fields (SNPs). In other words, the dataset is denoted by a matrix D_{mxn}, where *m* is the number of individuals in study and *n* is the number of studied SNPs. Table 3.2 shows a sample dataset. Each SNP dataset is accompanied by detailed frequency tables for every SNP value similar to Table 3.3, which shows the allele frequencies of the SNP dataset in Table 3.2. As explained in section 2.1.1, SNP values are usually biallelic, which means they can have one of two possible alleles (values). Hence, SNP values emerge as one of two possible values; each appearing in the studied population with a certain frequency. The allele with the higher frequency is referred to as the major allele. The remaining SNP allele is the minor allele.

Due to the numerical disposition of the SNP selection algorithms, an encoding technique was adopted. The alphabetic series of SNPs were encoded into a series of numbers according to the

major/minor allele frequencies following the algorithm in Table 3.1. Two encoding techniques were followed where all major alleles were encoded into either zeros or their corresponding major allele frequencies. Similarly, minor alleles were encoded into either ones or their respective minor allele frequencies. The difference in encoding into either 0/1 or major/minor frequencies is solely a matter of scale. Table 3.4 shows the 0/1 encoded SNP values for the dataset in Table 3.1.

The encoding technique for major/minor frequencies was chosen for the application of the mFOS-based selection method. The binary encoding technique was used for the ICA-based selection approach, in order to be able to compare the results with results generated by Lin *et al*. using PCA [41].

Algorithm 1: ***Encoding the data***

```
1.  for every haplotype (h_i)
2.      for every SNP (S_i)
3.          if value( S_i) = major allele
4.              Then S_i ← 0 or major allele frequency
5.          if value( S_i) = minor allele
6.              Then S_i ← 1 or minor allele frequency
7.          else missing_value
8.      end for
9.  end for
```

Table 3.1 The SNP data encoding algorithm

	SNP 1	*SNP 2*	*SNP 3*	*SNP 4*	*SNP 5*	*SNP 6*	*SNP 7*	*SNP 8*
Ind 1	C	T	C	G	C	C	C	T
Ind 2	C	A	T	A	G	C	C	T
Ind 3	C	T	C	G	C	C	T	T
Ind 4	C	T	C	G	C	C	C	C
Ind 5	C	A	T	A	G	C	C	C
Ind 6	C	A	T	A	G	C	C	T
Ind 7	C	A	T	A	G	C	C	C
Ind 8	C	A	T	A	G	C	T	T
Ind 9	T	A	T	A	G	C	C	C
Ind 10	C	A	T	G	G	T	C	T

Table 3.2 Sample SNP dataset showing 8 SNP values for 10 individuals.

	Major Allele	*Minor Allele*	*Major Allele Frequency*	*Minor Allele Frequency*
SNP 1	C	T	0.9	0.1
SNP 2	A	T	0.7	0.3
SNP 3	T	C	0.7	0.3
SNP 4	A	G	0.6	0.4
SNP 5	G	C	0.7	0.3
SNP 6	C	T	0.9	0.1
SNP7	C	T	0.8	0.2
SNP 8	T	C	0.6	0.4

Table 3.3 Major and minor allele frequencies as calculated according to the SNP sample in Table 3.2

	SNP 1	*SNP 2*	*SNP 3*	*SNP 4*	*SNP 5*	*SNP 6*	*SNP 7*	*SNP 8*
Ind 1	0	1	1	1	1	0	0	0
Ind 2	0	0	0	0	0	0	0	0
Ind 3	0	1	1	1	1	0	1	0
Ind 4	0	1	1	1	1	0	0	1
Ind 5	0	0	0	0	0	0	0	1
Ind 6	0	0	0	0	0	0	0	0
Ind 7	0	0	0	0	0	0	0	1
Ind 8	0	0	0	0	0	0	1	0
Ind 9	1	0	0	0	0	0	0	1
Ind 10	0	0	0	1	0	1	0	0

Table 3.4 The sample SNP dataset encoded into 0/1 according to the algorithm in Table 3.1 which depends on the major/minor allele frequencies shown in Table 3.3.

	SNP 1	*SNP 2*	*SNP 3*	*SNP 4*	*SNP 5*	*SNP 6*	*SNP 7*	*SNP 8*
Ind 1	0.9	0.3	0.3	0.4	0.3	0.9	0.8	0.6
Ind 2	0.9	0.7	0.7	0.6	0.7	0.9	0.8	0.6
Ind 3	0.9	0.3	0.3	0.4	0.3	0.9	0.2	0.6
Ind 4	0.9	0.3	0.3	0.4	0.3	0.9	0.8	0.4
Ind 5	0.9	0.7	0.7	0.6	0.7	0.9	0.8	0.4
Ind 6	0.9	0.7	0.7	0.6	0.7	0.9	0.8	0.6
Ind 7	0.9	0.7	0.7	0.6	0.7	0.9	0.8	0.4
Ind 8	0.9	0.7	0.7	0.6	0.7	0.9	0.2	0.6
Ind 9	0.1	0.7	0.7	0.6	0.7	0.9	0.8	0.4
Ind 10	0.9	0.7	0,7	0.4	0.7	0.1	0.8	0.6

Table 3.5 The sample SNP dataset encoded into major and minor allele frequencies presented in Table 3.3 and following the algorithm in Table 3.1

Chapter 4

SNP Selection

In this chapter, the proposed dimensionality reduction and feature selection methodologies for SNP data are explained. The purpose of this reduction is to eliminate the redundancy in the information carried along the genome. In order to realize dimensionality reduction, two techniques are proposed; they capture the information hidden in the studied SNP datasets by selecting a smaller subset of the most informative SNPs. Validation approaches adopted for the evaluation of the SNP selection are also presented in this chapter.

Independent Component Analysis is the first proposed technique and is considered a lossy transform. As mentioned before, ICA described a group of observed outputs as a linear combination of statistically independent components. Unlike the previous application of ICA on SNP datasets as a wrapper technique, the proposed implementation uses it as a filtering technique for unsupervised SNP selection. The second proposed technique is mFOS, an orthogonalization method which generates mathematical models for system outputs. FOS is a multivariate regression method devised to estimate observed outputs using selected basis functions. While previous implementations of FOS utilize functions, which are external to the system, for the estimation process, the proposed implementation on SNP datasets models individual haplotypes using the actual SNPs as basis functions. The proposed implementation promotes easier interpretation of the generated results. In addition to the proposed methodologies, SNP selection using (PCA), another lossy transform method, is briefly explained and implemented for comparative purposes.

4.1 Independent Component Analysis for SNP Selection

ICA is a signal processing technique that was originally developed to solve blind source separation problems. It aims to express a set of random variables as a linear combination of statistically independent component variables [81]. The significance of ICA lies in its ability to analyze data with non-Gaussian distributions. It is known that factor analysis and PCA also tend to calculate independent factors by decomposing the analyzed data; however, they impose an assumption of Gaussianity on studied data distributions, which is rarely valid for many biological datasets.

ICA considers a group of observed signals as mixed signals and proceeds to recover their original signals. To accomplish this goal, ICA estimates independent factors in multivariate data by decomposing the matrix of observed signals into a mixing matrix and the matrix of independent components. Thus, given a set of signals represented as a matrix X_{mxn} of random variables, one can estimate the matrix of original signals S_{kxn} by satisfying the following equation:

$$X_{mxn} = A_{mxk}\, S_{kxn} \qquad (1)$$

where A_{mxk} is the mixing matrix with real valued coefficients, m is the number of observed signals, k is the number of original signals ($k \leq m$) and n is the number of time points in each observed signal. The original signals S are the calculated ICs.

The decomposition of X into A and S can be approached using several optimization techniques discussed in detail in [82]. The original signals S are recovered by applying a linear transformation W on the observed signals X according to the following equation:

$$S = W\, X \qquad (2)$$

where W is the inverse of the mixing matrix A. The independent components are calculated by optimizing a measure of independence in the following linear combination of the observed signals:

$$S = \sum_{j=1}^{k} W_j X_l \qquad (3)$$

where W_j the linear transformation coefficients applied to the observed signal X_l.

The independent components, i.e. the linear combinations of the observed variables, are found at the actual local maxima of independence in all linear combinations of the analyzed data. Different ICA implementations use different measurements of independence. Mutual information and measure of peakedness of the signal (high order cumulant) are the most frequent independence measures used in the optimization process of the ICA. Each calculated local optima of a linear combination is considered as an independent component of the observed signals. Gradient methods are widely used for maximizing the independence among the different calculated ICs.

The calculation of the independent components guarantees that each pair of the estimated independent components are uncorrelated, as well as any non-linearly transformed versions of this pair of ICs [82]. Therefore, ICs are regarded as non-linearly uncorrelated components. The estimation of such components requires higher-order statistics. Thus, the covariance matrix which represents linear correlation among different variables is not useful for ICA, despite its critical necessity in other factor analysis methods.

In the proposed implementation, explained in detail in the coming section, the observed signals X are the group of studied haplotypes. While ICA considers the observed signals to be mixed, the haplotypes in the proposed implementation are considered to have redundant

information carried by associated SNPs. Hence, the overall information in the group of haplotypes is a mixture of redundant and unique. ICA is used to dissociate and de-mix this blend of information.

Figure 4.1 illustrates the workflow of the SNP selection framework using ICA. The first phase consists of encoding the SNP data into numerical format as explained in Chapter 3. The second

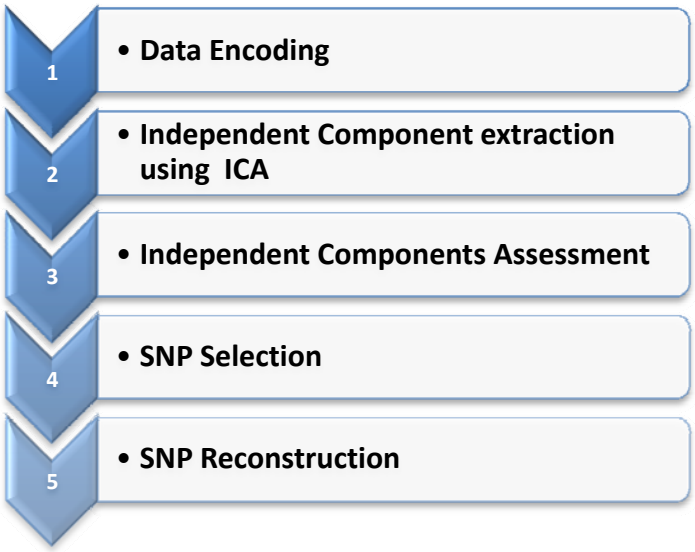

Figure 4.1 The workflow of SNP selection using ICA

phase comprises the application of ICA on the SNP dataset to extract the independent components. The third phase encompasses assessing the relevance of each extracted IC depending on the amount of information reflected by the combination coefficients. The fourth phase is the actual SNP selection. The selection method is performed by assessing the contribution of each SNP in the relevant group of ICs. Finally, in the fifth and final phase, the selection results are evaluated by reconstructing the original group of SNPs using the selected SNPs.

4.1.1 Independent Component Extraction

The uncorrelatedness and statistical independence of the calculated ICs are the actual motivation for applying ICA on SNP data. The ICs represent linear combinations of SNPs expressing maximal independence in the dataset. Accordingly, the calculated coefficients in these combinations are regarded as the individual contributions of each SNP to the ICs [82]. These contributions are represented in the following equation:

$$IC_j = \sum_{i=1}^{n} \alpha_i snp_i \qquad (4),$$

where α_i represents the coefficient value corresponding to the i^{th} SNP contribution in the j^{th} IC and n is the number of available SNPs in the dataset. As a result, a SNP with a higher coefficient in a certain IC is considered to play a main role in the extraction of independent factors. Such a SNP reveals the capability of summarizing information hidden in the dataset and furthermore representing other correlated or dependent SNPs. The lower contributions in the linear combinations of SNPs reflect redundancy in the information provided by the corresponding SNPs [82].

In this research, the fastICA[2] implementation is used to calculate the independent components in the three SNPs datasets. FastICA is a fixed point algorithm that finds the de-mixed matrix W, by maximizing the kurtosis measure of each IC [81]. In this implementation, the dataset is preprocessed before estimating the ICs [81]. The method decomposes the data using eigen vector decomposition and discards the vectors which explain low variance in the data. This step does not

[2] This implementation method is open source. The MATLAB version of the code available from: (http://www.cis.hut.fi/projects/ica/fastica/) was used.

alter the estimation of ICs, but rather assists in reducing the noise in the dataset as well as ensuring a faster convergence to an acceptable decomposition with fewer calculations [81].

4.1.2 Independent Component Assessment and SNP Selection

ICA calculates as many ICs as the number of variables (SNPs) in the dataset. Some ICs have homogenous coefficient values, thus none of the SNP contributions in the linear combination (IC) stands out in comparison with others. Such ICs exhibit less information and are considered less significant to the selection process. To assess the importance of each IC, the 4th order cumulant, also known as kurtosis [84] was used. Kurtosis is a statistical measure that represents whether the data distribution is peaked or flat relative to a normal distribution. It is calculated according to the following equation [63]:

$$Kurtosis = \sum \frac{(X-\mu)^4}{N\sigma^4} - 3 \qquad (5)$$

where X is the analyzed signal, μ is the mean of the signal X, σ is the standard deviation of the signal, and N is the number of observations in the signal.

A normal distribution has a zero kurtosis by definition [68]. The higher the value of kurtosis, the more peaked its distribution and the further it is from Gaussian or normal distribution. Kurtosis values can be either positive or negative. Negative kurtosis signifies a flatter distribution, whereas positive kurtosis characterizes more peaked distributions [64]. The kurtosis values of ICs explain the degree to which the distribution of the SNP contributions in each IC is peaked. An IC with a few significant SNP contributions has a high kurtosis value, whereas an IC with almost identical SNPs contributions has a low Kurtosis value. The more peaked the distribution of coefficients in an IC are, the clearer the contribution of the most informative SNP in this IC. In contrast, the flatter the distribution, the more ambiguous the SNPs contributions are. To identify a kurtosis threshold for choosing the group of significant ICs, the entire range of

calculated kurtosis values in a dataset were examined. Different kurtosis thresholds indicated by the P% coverage of the range of values were used. The P% coverage indicates the number of ICs selected starting with the upper bound of the kurtosis range and decreasing in kurtosis values until the P% of the range is selected. For example, given a group of kurtosis values ranging from 1 to 50 and at 85% threshold, all ICs having kurtosis values between 50 and 7.5 are added to the group of significant ICs.

Independent components are assessed as follows:

- Calculate kurtosis values for all ICs
- Decide on a threshold for kurtosis values satisfying the desired importance of SNP contributions in ICs.
- Select ICs with kurtosis values greater than or equal to the threshold.

After selecting the most significant ICs, the SNP with highest contribution in each IC is selected. This contribution is quantified by the coefficient value in the IC corresponding to the SNP position. Accordingly, SNPs with higher contributions, represented by the SNP coefficients, in one of the significant ICs are selected. SNPs with higher coefficients display greater involvement in the extraction of maximal independency factors. As a result, each IC is mapped to its most contributing SNP.

4.1.3 SNP Reconstruction and Selection Evaluation

After performing the SNP selection, it is necessary to prove that the selected group of SNPs (or htSNPs) is able to summarize the hidden information in the data set by predicting the non-selected SNPs. Therefore, a reconstruction algorithm was implemented in order to reconstruct the

analyzed datasets using the htSNPs. This reconstruction algorithm is based CUR-type matrix decomposition [86], [87].

Given a matrix D, the algorithm decomposes the matrix into a product of three matrices C, U and R. The matrix C consists of a small number of actual columns of D, whereas the second matrix R consists of a small number of actual rows of D. The third matrix U is a carefully constructed matrix that guarantees that the product of $C \times U \times R$ is as close as possible to D [87]. The reconstruction method basically expresses every column, in the original data matrix D, as a linear combination of the chosen columns, i.e. SNPs, by solving a least squares regression problem [86]. Subsequently, the results of regressions are rounded to approximate the original data [86].

The adopted reconstruction approach is applied after partitioning the data into training and testing components. SNPs are selected using the ICA framework which is applied on the training partition to predict the values of non-chosen SNPs in the testing partition. More specifically, given a training matrix D_1 of size $m_1 x n$ (which is the matrix R in the CUR definition) and testing matrix D_2 of size $m_2 x n$, the ICA framework is applied on the training set in order to select the informative SNPs (few columns of the data). Then, from the number of chosen SNPs, h, the matrix W of size $m_1 x h$ is constructed (W corresponds to U in the definition of CUR) and contains the values of the chosen SNPs in the training set. The informative SNPs (htSNPs) should always be assayed (i.e. genotyped) since their values are unpredictable by other groups of SNPs, which is a reason for being selected as htSNPs. A matrix C (as in the definition of CUR) of size $m_2 x h$ is then constructed from the values of htSNPs in the testing set D_2. Next, the predicted or reconstructed testing matrix is calculated according to the following equation:

$$D_{2_{m2\times n}} = C_{m2\times h} * {W^{+}}_{h\times m1} * D_{1_{m1\times n}} \qquad (6),$$

where W^{+} denotes the Moore–Penrose generalized inverse of W [87].

Since the data is encoded into a binary format, the reconstructed matrix with values rounded to either '0' or '1' is compared with the original data matrix in order to obtain the accuracy of SNP reconstruction. The accuracies represent the ratio of the number of correctly reconstructed SNPs to the number of all SNPs in the matrix. A higher ratio indicates a more accurate reconstruction. High reconstruction accuracy reflects that the selected SNPs capture a high portion of the information present in the data.

4.2 Modified Fast Orthogonal Search for SNP Selection

FOS is a multivariate regression method that performs a least square fit for a system output using predefined candidate functions [65]. The algorithm iteratively searches in a pool of functions to build a model for estimating the output of a system. The candidate functions are chosen so that the mean square error is minimized between the actual system output and the estimated measure from the model. One candidate term is added at a time to the model during error minimization. Models built using FOS are linear combinations of non-orthogonal candidate functions of the following form:

$$y(n) = \sum_{m=0}^{M} a_m \, p_m(n) + \, e(n) \qquad (7)$$

where M is the number of selected candidate functions, n represents time, $y(n)$ is the actual output of the model, p_m are the non-orthogonal candidate functions, a_m are the regression coefficients of the best fit of the model to the output, and $e(n)$ is the error between the actual and estimated system outputs.

Applying an orthogonal transform to the model presented in (7), we obtain a set of uncorrelated functions (w_m) that equivalently model the system output $y(n)$ as:

$$y(n) = \sum_{m=0}^{M} g_m \, w_m(n) + \, e(n) \qquad (8)$$

where M is the number of mutually orthogonal functions $w_m(n)$. FOS decreases the complexity of orthogonalizing the candidate function $p_m(n)$ by calculating the coefficients g_m of the correlation of orthogonal functions $w_m(n)$ and their corresponding $p_m(n)$. There are two methods of approaching the use of FOS for SNP selection. The first method is to have each SNP in the dataset (columns) as an output y(n) to be modeled, and applying FOS to identify the group of SNPs (other columns) that can best predict the corresponding y(n). The second method, implemented in this thesis, considers each haplotype (row) in the dataset as the output y(n) to be estimated and uses the SNPs (columns) for the estimation process. This latter approach for estimation imposes modifications to the original derivations of FOS. In this thesis it is referred to as the modified implementation of FOS, namely modified Fast Orthogonal Search (mFOS). The changes made to the original FOS are restricted to the calculation of the correlation coefficients, g_m. In the original derivations of FOS, the calculation of g_m entails a multiplication of the orthogonal function $w_m(n)$ and the $y(n)$ to be modeled as:

$$g_m = \frac{<y(n), w_m(n)>}{<w_m(n), w_m(n)>} \qquad (9)$$

In the implementation for mFOS, since the size of the output vector to be estimated, namely n, is larger than the size of the candidate function vector,l, the multiplication operation is not feasible. Therefore, multiplication is substituted by the convolution of the $y(n)$ and $p_m(l)$ where $n > l$. In FOS, the orthogonal functions $w_m(n)$ are derived from $p_m(n)$ and this derivation remains the

same in the proposed mFOS except for the dimension of p_m which is l instead of n, where in FOS $l = n$. The following equations present the derivations of w_m [66]:

$$w_0(l) = 1, \ \forall \ l \ time \ points \tag{10}$$

$$w_m(l) = \ p_m(l) - \ \sum_{r=0}^{m-1} \alpha_{mr} w_r(l) \ , \ \forall \ m \geq 1 \tag{11}$$

Assuming the calculations are performed according to a certain defined inner product space, inner products of $w_r(l)$ and $w_m(l)$ for all $r \ < \ m$ are equal to zero due to their orthogonality. The only non-zero product is the $w_m(l)$ with itself; therefore:

$$\alpha_{mr} = \frac{\overline{p_m(l) w_r(l)}}{\overline{{w_r}^2(l)}}, \forall \ r \in \{0, \dots, m-1\} \tag{12}$$

and where the bar represents the time average over all time points in corresponding vector.

FOS, and therefore the proposed mFOS method, is devised to reduce the computational overhead entailed by the actual orthogonalization of functions [65]. Previously, the orthogonal search (OS) was used to model system outputs, but it was computationally expensive due to the complexity of calculating the orthogonal functions, especially when the number of basis functions was very large. FOS is the fast implementation of OS which avoids the actual calculation of orthogonal functions and settles for the calculation of the g_m coefficients [66]. FOS's exact derivations of the previously mentioned coefficients are presented in detail in Appendix A.

The derivation of mFOS are described in the following set of equations:

$$g_m = \frac{\hat{C}(m)}{D(m,m)} \ , \ \forall \ m \ \in \{0, \dots, M\} \tag{13}$$

where Ds are calculated as follows:

$$D(0,0) = \ 1 \tag{14.1}$$

$$D(m,0) = \ \overline{p_m(l)}, \qquad \forall \ m \ \in \{0, \dots, M\} \tag{14.2}$$

$$D(m,r) = \overline{p_m(l)p_r(l)} - \sum_{i=0}^{r-1} \alpha_{ri} D(m,i), \quad \begin{cases} \forall\, m \in \{1,\dots,M\} \\ \forall\, r \in \{1,\dots,m\} \end{cases} \qquad (14.3)$$

and where αs are:

$$\alpha_{mr} = \frac{D(m,r)}{D(r,r)}, \qquad \begin{cases} \forall\, m \in \{1,\dots,M\} \\ \forall\, r \in \{0,\dots,m-1\} \end{cases} \qquad (15)$$

Furthermore, $\hat{C}$s are calculated according to the following equations:

$$\hat{C}(0) = \overline{y(n)} \qquad (16.1)$$

$$\hat{C}(m) = \overline{y(n) * p_m(l)} - \sum_{r=0}^{m-1} \alpha_{mr} \hat{C}(r), \quad \forall\, m \in \{1,\dots,M\} \qquad (16.2)$$

where $*$ denotes the convolution operation and the $y(n) * p_m(l)$ vector has $n + l - 1$ terms and its average is:

$$\overline{y(n) * p_m(l)} = \frac{\sum_{k=1}^{n+l-1} \sum_{j=\max(1,k+1-n)}^{l} y(k-j+1) p_m(j)}{n+l-1} \qquad (16.3)$$

$$= \frac{1}{n+l-1} \sum_{k=1}^{n} y(k) \sum_{j=1}^{l} p_m(j) \qquad (16.4)$$

Therefore $\hat{C}(m)$ is:

$$\hat{C}(m) = \frac{1}{n+l-1} \left[\sum_{k=1}^{n} y(k) \sum_{j=1}^{l} p_m(j) \right] - \frac{1}{n+l-1} \sum_{r=1}^{m-1} \alpha_{mr} \sum_{k=1}^{n} y(k) \sum_{j=1}^{l} p_r(j) - \alpha_{m0} \hat{C}(0)$$

$$\hat{C}(m) = \frac{n}{n+l-1} C - \alpha_{m0} \overline{y(n)} \qquad (16.5)$$

where C is the coefficient calculated using FOS (Appendix A, equation A.7) for an estimated output without D.C. component ($C(0) = 0$).

The proposed substitution of the multiplication operation by the convolution affects the calculation equation of C and Q (Appendix A, equations A.6-A. 10) where $\hat{C}$ and $\hat{Q}$ (the amount each orthogonal function deducts from the error) are the modified versions.

$\hat{Q}(m)$ is calculated in analogy with the $Q(m)$ calculated in FOS (Appendix A) and according to the following equations:

$$\hat{Q}(m) = \frac{\hat{C}^2(m)}{D(m,m)}, \quad \forall\, m \in \{1, \dots, M\} \qquad (17.1)$$

As a result, the estimation error is no longer the Mean Square Error (MSE) as defined in the original implementation of FOS in Appendix A (equation A.12).

Accordingly, the mFOS algorithm, in keeping with the original implementation of FOS, starts with an observation of the signal $\boldsymbol{y}(\boldsymbol{n})$ and a set of potential candidate functions. It iterates over the candidate functions calculating the reduction in error as a result of adding each function. It chooses to keep the candidate function causing maximum reduction in error The stopping criterion is reached when the largest reduction in error caused by the chosen candidate is less than the one that can be generated by white Gaussian noise or when all candidate functions have been inspected.

4.2.1 SNP Data Modeling using mFOS

As previously mentioned, the goal of applying mFOS on SNP data is to select a subset of SNPs capable of modeling and predicting the entire group of studied SNPs. The proposed SNP selection using mFOS was applied on two datasets: the ACE and ABCB1 genes datasets. Prior to

explaining the details of implementation, some terms used in the application of mFOS on the matrix D_{rxp} of SNP data are presented:

- The system output $y(n)$ → The values of all SNPs for a certain chromosome in the data; each row r of D is considered as $y(n)$.
- The candidate functions $w_m(l)$ → The group of SNPs in the studied dataset; each column p of D is considered as a candidate function.

In order to achieve the desired reduction in SNP dataset dimensions and following the original FOS algorithm, mFOS builds a model for every $y(n)$ in the dataset. Each model consists of a group of SNPs capturing the highest amount of information in the data. In other words, each column of data is treated as a candidate function. At each iteration of mFOS, the potential reduction in the error caused by the addition of a candidate SNP to the current model of $y(n)$ is calculated. The SNP causing the largest reduction in the error is kept and added to the group of already selected SNPs from previous iterations. After obtaining a model for every $y(n)$, the SNPs in all models are tallied to identify SNPs which contributed to most of the models. These SNPs are the most informative features in the data and are able to soundly model the data.

Consequently, the dimensionality of the data can be reduced by keeping only columns corresponding to this group of most frequent SNPs. Figure **4.2** provides a flowchart of the above procedure for modeling $y(n)$ using a subset of the available SNPs in the dataset. The process depicted in the flowchart is repeated for all $y(n)$ available in the dataset. The collective result of all these processes is obtained by selecting the group of htSNPs having the highest frequencies in contributing to the calculated models.

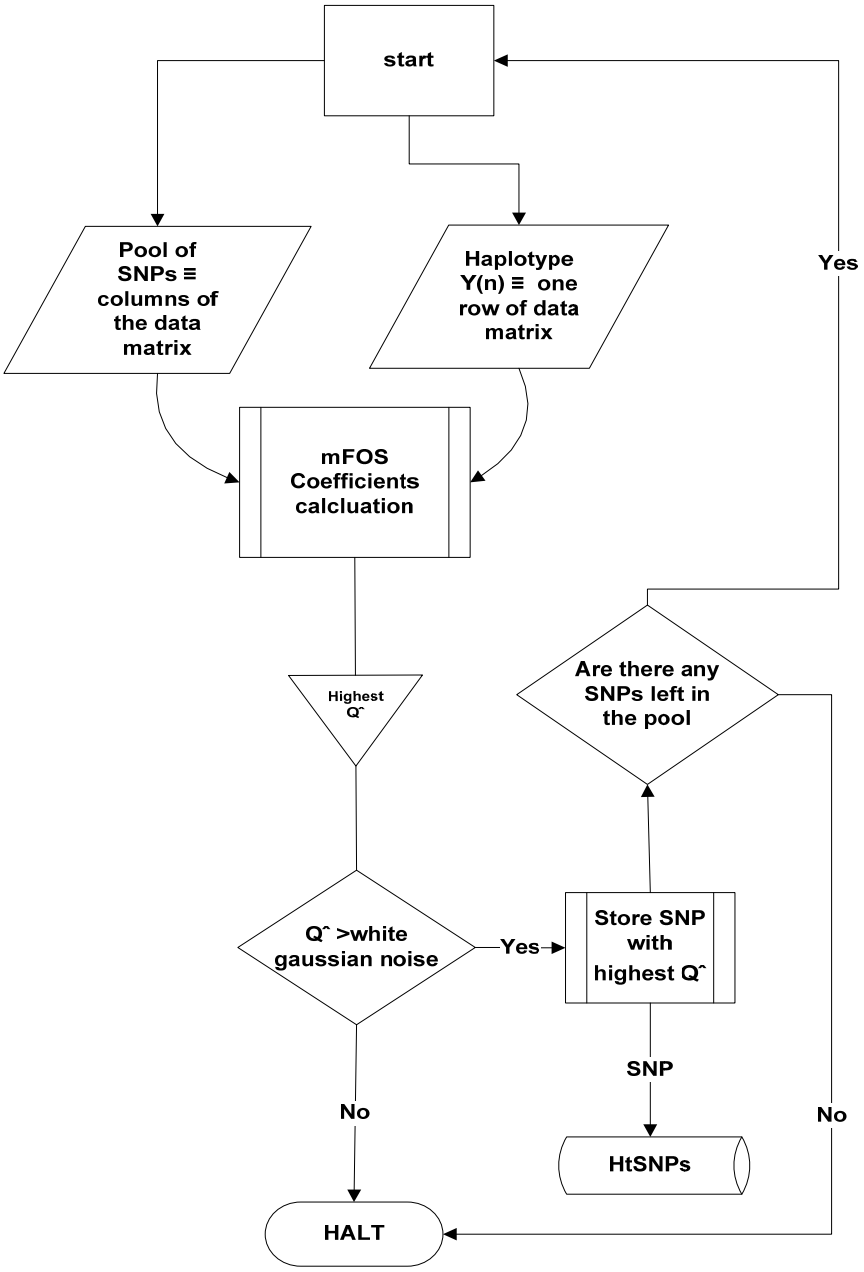

Figure 4.2 Flowchart of the application of mFOS for modeling the SNP values in one haplotype

4.2.2 Model Assessment

Each built model represents a group of htSNPs which are the most informative SNPs. In order to prove that the selected SNPs summarize the information carried along the studied haplotypes, and to prove that this group is also able to predict non-selected SNPs, the reconstruction algorithm explained in section 4.1.3 was implemented to reconstruct the studied dataset using the values of htSNPs.

This reconstruction approach is considered an assessment of the models built using mFOS as well as an evaluation technique for the SNP selection. After performing the regression to reconstruct the non-chosen SNPs, the difference between the original SNP values and the reconstructed values were calculated. The comparisons were based on encoding the calculated difference between the original SNP values and the reconstructed values into a binary form where "0" is used to indicate similar values and "1" is used to indicate different values. Accordingly, a calculated reconstruction difference higher than the tolerance threshold is set to "1" and a difference lower than the threshold is set to "0".

The difference tolerance threshold was set to 0.2. This tolerance threshold was selected by trial and error; however the histograms of the estimation errors and the bar charts shown in Figure **4.3**, Figure **4.4** and Appendix B provide an insight into the threshold choice. Figure **4.3** and Appendix B (figures B.1, 2, 3, 4, 5 and 6) show the histograms of estimation errors calculated throughout the experiment of modeling the different haplotypes in the ACE dataset. By inspecting the histograms, the 0.2 threshold appears as a significant threshold since very few errors are above this value (notice the drop point in the histogram at 0.2). Figure **4.4** and Appendix B (figures B.7, 8, 9, 10, 11 and 12) show the bar chart of allele frequencies for the ACE dataset and their estimation errors obtained by using 4, 5, 6, 7, 8 and 9 htSNPs. Please note that the

haplotypes that are predicted have are real numbers and the discrete frequency and error values represented in Figure 4.4 are only for visualization purposed. In Figure 4.4, the allele frequencies and the estimation errors are rounded to the highest first decimal digit in order to be able to plot the bar charts. For instance, all estimation error having values between 0 and 0.1 are shown in the bar charts as red bar of height 0.1 and allele frequencies having values between 0 and 0.1 are shown as blue bars of height 0.1 and similarly for other estimation error and allele frequencies. The number of htSNPs reported in Figure **4.4** resulted in the best reconstruction accuracy for the ACE dataset in the reconstruction procedure. In these bar charts, the heights of the blue bars represent the allele frequencies of the different SNPs to be reconstructed, and the heights of the red bars represent the estimation error in reconstructing the SNPs. The blue and red bars are grouped in order to have each allele frequency (blue bar) grouped with its corresponding estimation error (red bar). The dashed horizontal red line represents the tolerance threshold. According to the bar charts, there are only few allele frequencies which were reconstructed with an error of greater than 0.2. In addition, the estimation errors of greater than 0.2 do not result from reconstructing SNPs with allele frequencies falling in the range where such a threshold can alter the results. More specifically, the absence of a direct correlation between the estimation error of 0.2 and the frequencies (between 0.4 and 0.6) which might be affected by a tolerance threshold of 0.2 is noticeable. For the allele frequencies of 0.5, the estimation error is always 0.1 or less, indicating the ability to reconstruct these values closely. As shown in the bar chart, accurate reconstruction of the haplotypes is achieved for all allele frequencies. The labeling of alleles of frequency 0.5 is arbitrarily performed as major and minor alleles; if this approach were to be used for major and minor allele prediction, then at allele frequencies equal to 0.5, results can be interpreted as either the major or minor allele since the threshold value cannot resolve them.

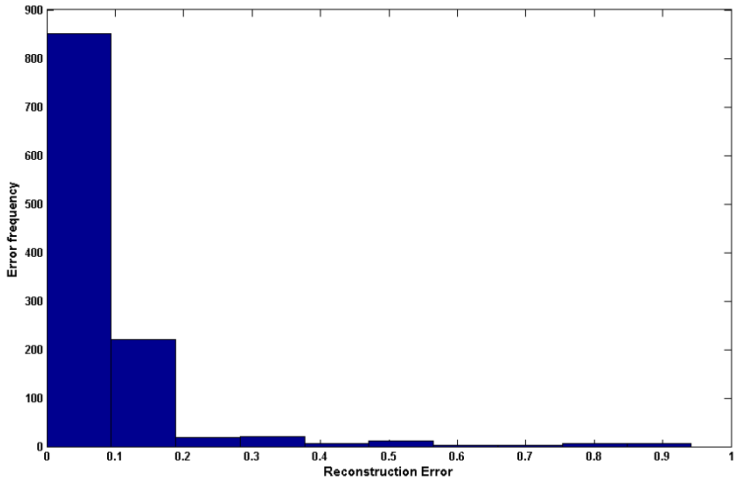

Figure 4.3 Histogram showing the frequency (y-axis) of obtaining the reconstruction errors on the x-axis using 4 htSNPs to predict the haplotypes of the ACE Dataset.

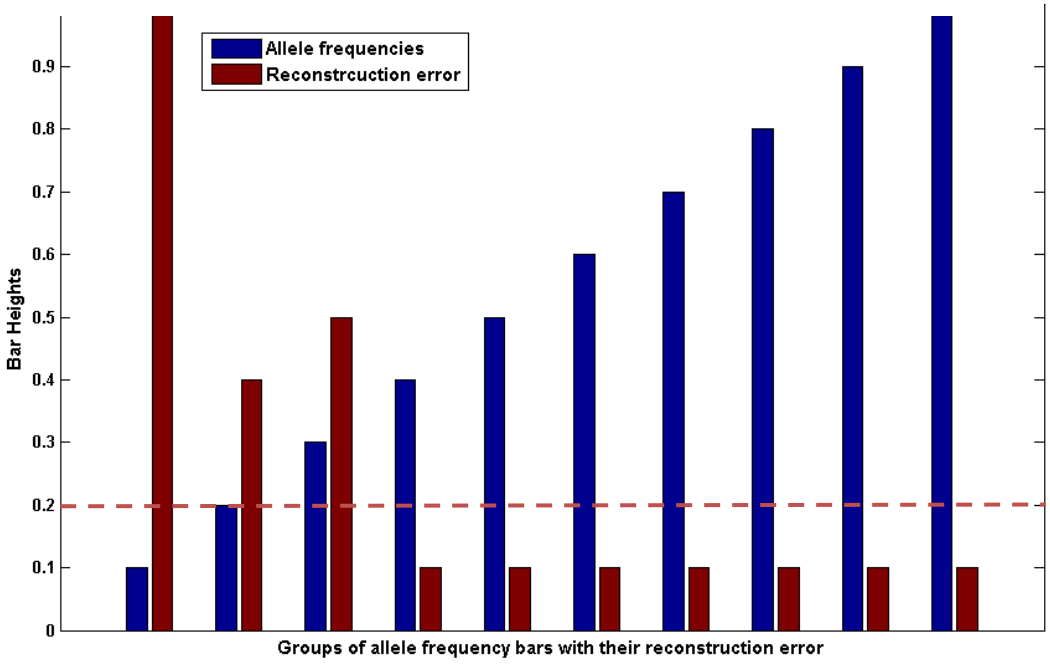

Figure 4.4 Bar chart showing the different allele frequencies of the ACE dataset and their reconstruction errors using 4 htSNPs. The heights of the blue bars represent rounded allele frequencies and the heights of the red bars represent the range of rounded error for those frequencies. Each allele frequency bar is grouped with its reconstruction error bar along the x-axis. The dashed red line shows the chosen threshold of 0.2.

The adopted reconstruction approach is applied after building mFOS models and choosing the most informative SNPs from the data. The accuracy is calculated as the ratio of the number of correctly reconstructed SNPs to the number of all SNPs in the matrix. The higher the ratio, the more accurate the reconstruction is. High reconstruction accuracy reflects the fact that the selected SNPs capture a high portion of the information present in the data, and suggests that the built models are accurate.

4.3 Validation Techniques

In order to have sufficient statistical confidence in the results generated by the two proposed SNP selection approaches, and in order to predict the extent of success of both approaches on unseen data, several cross-validation techniques were incorporated. Each studied dataset was partitioned into training and testing sections. The partitioning was performed by segregating the haplotypes in the study into training and testing partitions; partitioning was repeated many times to experiment with different combinations of haplotypes in both partitions. The training partition was used for the application of the proposed approaches and the testing partition was used to evaluate the results obtained from the training by performing the reconstruction algorithm. Two cross-validation methods were used: leave-one-out and k-fold. Leave-one-out cross-validation is characterized by leaving only one record from the data for testing and using the remainder for training. On the other hand, k-fold cross-validation segregates the data in k equal partitions, uses k-1 partitions for training and tests on the remaining partition.

4.3.1 Validation for SNP Selection Using Independent Component Analysis

First, a “k-fold” cross-validation method was adopted while applying reconstruction. In other words, the SNP selection approach was applied to different percentages of the data, after which

the remainder was reconstructed. Five different proportions were used for partitioning training and testing data, starting from 50% training and 50% testing data (2-fold cross-validation) and working up to 90% training and 10% testing data (10-fold cross-validation). The training/testing partitions were defined according to the rows (haplotypes) of the data matrix D. Results were averaged and grouped by the number of htSNPs. For every selected number of htSNPs, the average of all calculated reconstruction accuracies are reported. For example, all reconstruction accuracies generated at iterations where x htSNPs were selected were grouped, then the average of the accuracies in the group was calculated and one value for the group of x selected SNPs was reported.

Second, the application of the methodology was repeated on the three datasets for 300 times to ensure that different combinations of haplotypes were taken into account. Moreover, the repetitions also helped in overcoming any flawed randomization in the cross-validation implementations. The reported reconstruction accuracy results were averaged over the 300 repetitions.

4.3.2 Validation for SNP Selection Using Fast Orthogonal Search

K-fold cross validation was applied and implemented differently for the two datasets due to the difference in their dimensions. For the ACE dataset, leave-one-out cross validation was used; mFOS was applied to all rows of the dataset except for one, to build a model. The reason for choosing the leave-one-out method was because of the small number of available haplotypes in the dataset. Choosing k-fold cross-validation for this dataset may cause some loss of information in the candidate SNPs used for the application of mFOS. This process was repeated 22 times (the number of rows in SNP data matrix D), and a histogram of the frequency of all selected SNPs

was created. Next, different cut-offs for frequencies of selected SNPs were chosen in this histogram, and SNPs with frequencies higher than the threshold were kept as the group of selected SNPs. Finally, the data was reconstructed one row at time, using the selected SNPs and each frequency cut-off and the accuracy was reported.

For the ABCB1 dataset, 10-fold cross validation was performed. mFOS was applied to 9/10 of the data every time and the process was repeated 10 times to build 10 models. Similar to the procedure in the first dataset, a histogram of selected SNPs from the 10 models was generated and groups of SNPs corresponding to different cut-off points were built. Selected SNPs at various cutoffs were used to reconstruct the data and report the accuracies.

4.4 Principal Component Analysis for SNP selection

The results of SNP selection and reconstruction from mFOS and ICA were compared to a method devised by Lin *et al.* in [41] which uses Principal Component Analysis. PCA is a dimension reduction method for multivariate data. Geometrically, it rotates the data in order to project maximum variability on to orthogonal axes, according to a minimum-square-error (MSE) criterion [41]. This method essentially decorrelates the columns in the data by decomposing the covariance matrix into eigenvectors and eigenvalues. The decomposition of a given dataset X with a covariance matrix C is performed according to the following equation:

$$CV = \lambda V \qquad (19),$$

where V is the matrix of eigenvectors and λ is the corresponding eigenvalues.

The data's principal components are the eigenvectors of the covariance matrix. Principal components are linear combinations of the original features in the data and are characterized by their mutual uncorrelatedness. In addition, the eigenvalues represent the percentage of the data

variance explained by their corresponding eigenvectors. As a consequence, this decomposition provides a way to select the most informative linear combinations of the data's original features according to the amount of variance each combination explained, and represented by their corresponding eigenvalues. This selection is the main approach for dimensionality reduction using PCA.

The application of PCA on SNP datasets allows for a SNP selection process to be performed. Since the eigenvectors are linear combinations of features (SNPs), each coefficient in these vectors are considered as the SNP's contribution in that principal component. This notion of contribution was also used when implementing the ICA framework, above. However, in the ICA method, the notion has stronger implications since the components are not only uncorrelated but also statistically independent.

The authors in [41] adopted this PCA approach to select the most informative SNPs. Their selection algorithm was re-implemented in order to re-generate the results from Lin *et al.* and to compare them with the results from the proposed methodologies. Due to a lack of sufficient information on their implementation, specifically on their non-htSNPs prediction method, it was not possible to reproduce their exact results. Therefore, their PCA selections were combined with the reconstruction approach into one method called the combined PCA selection/reconstruction. This combined approach utilizes the greedy discard (GD) method in [41]. GD method removes the SNPs which have high coefficients in the group of eigenvectors that have low eigenvalues. After performing the selection technique, the analyzed dataset was reconstructed using the reconstruction algorithm explained in section 4.1.3.

Chapter 5

Results and Discussion

In this chapter, the application of the proposed SNP selection techniques is discussed and the detailed results obtained from different datasets are presented. The results are explained in two main sections. The first section presents SNP selection results using ICA and PCA. The second section discusses the results of applying mFOS on different datasets. In addition, a comparison of the ICA results with a SNP selection method based on PCA, as devised in [41] is provided as well as the modified implementation of PCA combined with the SNP reconstruction approach. Finally, a comparative discussion of all methods is presented.

5.1 Results of SNP Selection Using ICA

5.1.1 ACE Dataset Results

ICA was applied for SNP selection on a range of 50%-90% of training partitions of the ACE dataset, and reconstructed the haplotypes in the remaining partitions (testing data). Partitioning the data segregated the haplotypes (rows of the data matrix D) into training and testing haplotypes. An average reconstruction accuracy of all the test partitions was then calculated and reported. In order to choose the kurtosis threshold needed to identify the significant group of ICs, the partitioning experiment was repeated several times, ensuring the selection of different combinations of training and testing. The resulting histogram is shown in Figure **5.1**. The histogram illustrates the different kurtosis values for the calculated ICs and the frequency of obtaining each of the kurtosis values in all the iterations of the implementation. For every training partition, different kurtosis thresholds were tried.

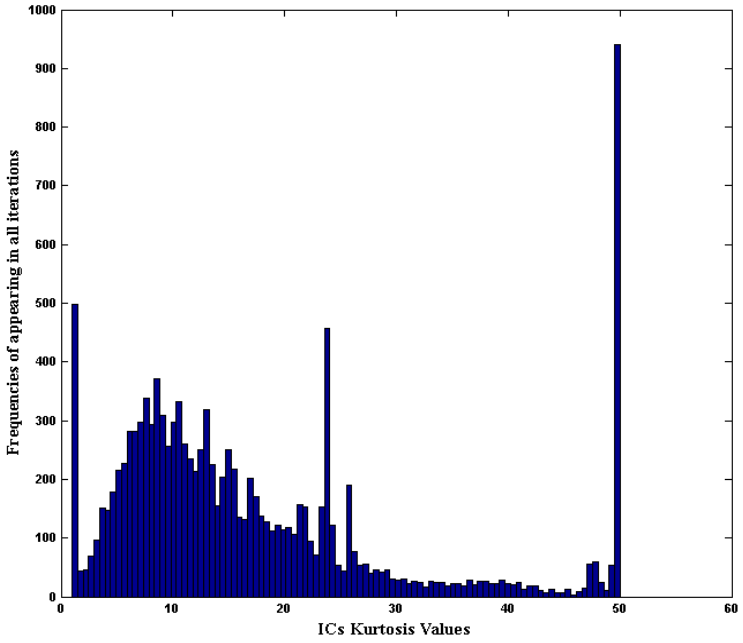

Figure 5.1 Kurtosis histogram of the ICs extracted from ACE dataset.

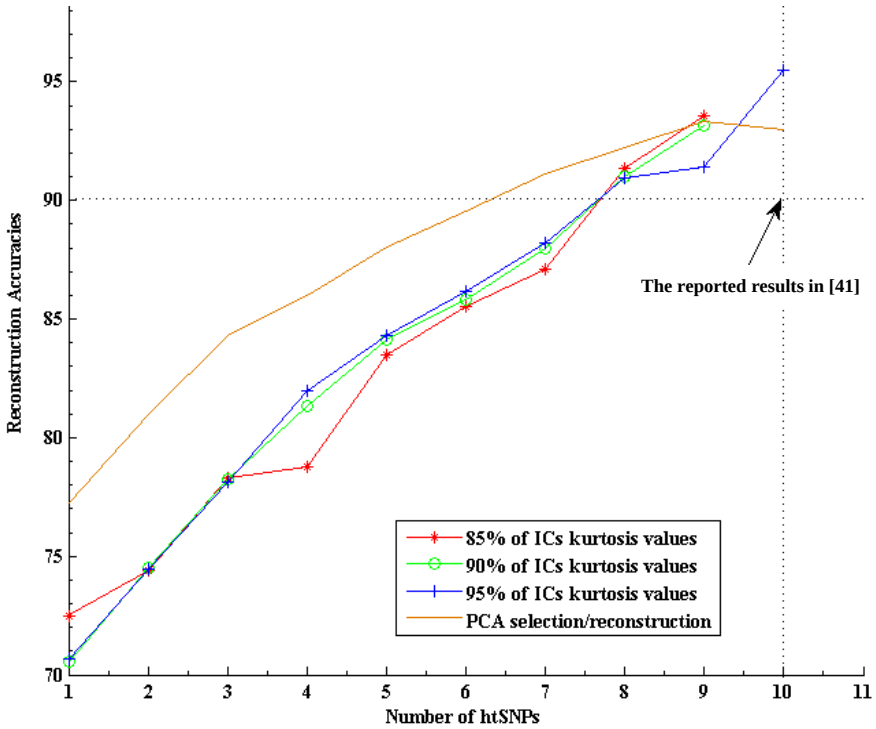

Figure 5.2 The results of the ICA framework on the ACE dataset.

For the purpose of demonstration, the results of SNPs contributing 85%, 90% and 95% of ICs kurtosis values are reported in the histogram in Figure **5.1**. Average reconstruction accuracies at the selected kurtosis threshold levels are shown in Figure **5.2**. The reported accuracies range from 70.6% to 95.5%. The horizontal line at 90% accuracy shows the result reported in [41]en whchosing10 ht-SNPs out of 52 SNPs in the dataset.

The line with star markers (*) (in Figure **5.2**) shows the average reconstruction accuracies for the 85% kurtosis threshold. The reconstruction accuracy reaches 90% when only 8 out of 52 SNPs are selected, compared to the selection of 10 SNPs reported in [41]. This difference in results demonstrates that the redundancy in information captured by the ICs is less than the redundancy captured by PCA only. Thus, the reconstruction method using the ICA selected SNPs is able to regenerate the non-chosen SNPs with more accuracy. The highest accuracy reached at the 85% kurtosis threshold is 93.5% with the selection of 9 out of 52 SNPs.

Also shown in Figure **5.2**, the line with circle markers depicts the results of the ICA framework with a 90% kurtosis threshold. Approximately half of the accuracies calculated at this threshold are higher than those at an 85% kurtosis threshold. However, the highest accuracy obtained with the selection of 9 SNPs is very similar for both kurtosis thresholds. These two sets of results suggest that the group of significant ICs selected at 85% and 90% kurtosis thresholds capture similar amounts of information.

The third line in Figure **5.2**, with '+' markers, represents the average reconstruction accuracies at a 95% kurtosis threshold. This line reflects a better performance in reconstruction compared with the other two thresholds. The highest accuracies here are reported at 95.5% accuracy with 10 SNPs selected. Again, by comparison with the results presented in [41], selecting 10 SNPs using PCA contained less information than the ICA method since reconstructing the non-chosen SNPs

with the same number of htSNPs yields higher reconstruction accuracies with SNP selection using ICA .

The fourth line in Figure **5.2** illustrates the results of the combined PCA selection/reconstruction approach. The PCA selection method used to generate the results in the fourth line is identical to the selection technique devised in [41]. However, the accuracies are reported according to our reconstruction technique using the htSNPs selected by PCA. The accuracies of reconstruction using this approach are higher than those reported by [41]. In comparison with the ICA framework, the highest accuracy reached by the combined PCA selection/reconstruction approach was 93% with 10 selected SNPs, which is lower than the best accuracy reached by ICA at 95.5%. However, this line reports an accuracy of 89.5% with only 6 selected SNPs. This later accuracy suggests that the modified PCA approach is capable of to selecting less SNPs and generating a more accurate reconstruction on this dataset.

A justification of PCA performance on this dataset is the Linkage Disequilibrium on genetic location of the ACE gene. As previously mentioned , SNP selection discards the redundancy in information due to the Linkage Disequilibrium associations among SNPs (see chapter 2.2). On the ACE gene, SNP associations might merely be correlations which can be dissociated using only PCA, and they might not exhibit any stronger associations that require the non-linear decorrelation approach of ICA in order to dissociate them. Another justification is the size of this dataset. The number of available SNPs in this dataset is not sufficient for the ICA to extract enough information from the dataset. In addition, the number of haplotypes is also small and does not provide enough training haplotypes for the extraction of ICs.

These results reveal the capability of the ICA framework to capture higher amounts of information, with less redundancy, compared to the PCA approach implemented in [41]

However, the results generated by the combined PCA selection/reconstruction method show better reconstruction accuracy with fewer selected SNPs from this dataset. The other two datasets demonstrate better performance for the ICA framework approach due to the larger size of available data used for training and testing.

5.1.2 ABCB1 Dataset Results

Partitioning percentages for training and testing of the ABCB1 dataset are identical to those applied on the ACE dataset. The reconstruction accuracies versus the size of the group of htSNPs used for their generation are reported in Figure **5.3**. These results are reported according to the same kurtosis thresholds used for the ACE dataset (85%, 90% and 95% kurtosis threshold).

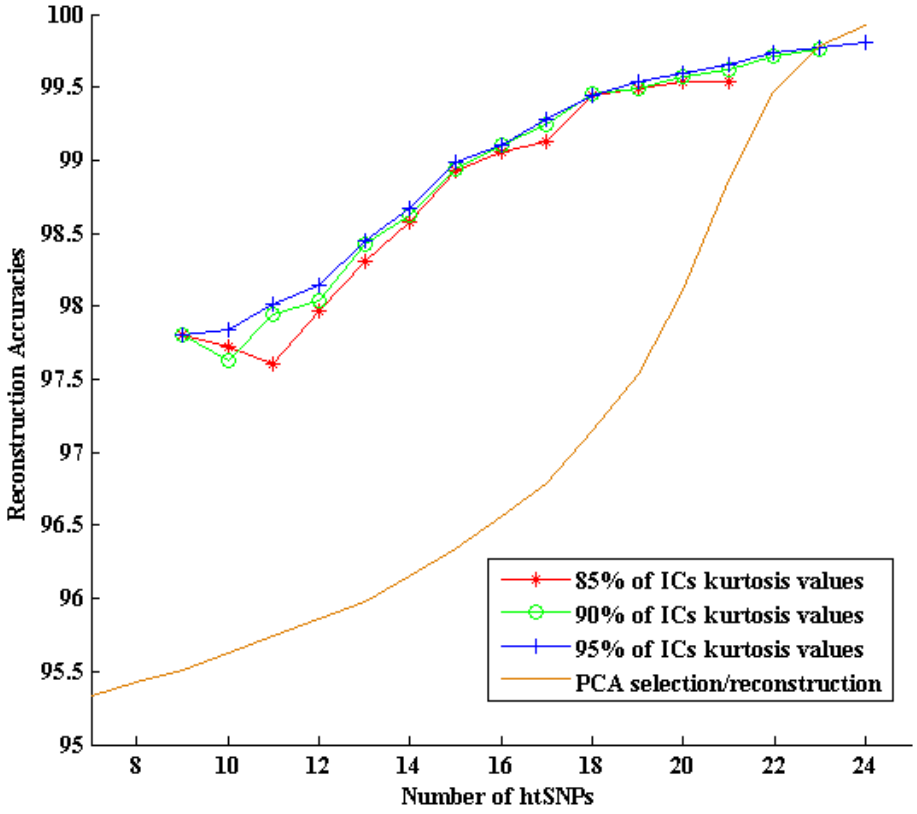

Figure 5.3 The reconstruction accuracies for ABCB1 at three kurtosis thresholds.

The 85% kurtosis threshold results are shown by the line with '*' markers. An accuracy of 97.5% is reported when selecting 10 out of 27 SNPs compared to an 80% accuracy reported in [41] with the same number of selected SNPs. The highest accuracy reached at this kurtosis threshold is 99.5% with 21 selected SNPs. In [41], the highest accuracy of 95 % was reported with 23 selected SNPs.

The other two lines with 'o' and '+' markers correspond to the results generated at 90% and 95% kurtosis thresholds, respectively. These results are slightly better than those reported at an 85% threshold. The reported accuracies in [41] are less than those reported in Figure 5.3. Moreover, the lowest reported accuracy for the 85%, 90% and 95 % kurtosis thresholds are 97.8% for 9 out of 27 selected SNP.

In addition to the difference in individual accuracies between the presented results and those in [41], it is observed that the ICA framework results generated by different numbers of selected SNPs vary over a smaller range of reconstruction accuracies in comparison with the larger range of accuracies generated in [41]. Overall, the results of the ICA framework are consistent and do not change with the kurtosis thresholds; they follow the same trend despite the change in the number of selected SNPs as seen in Figure **5.3**. The results in Figure **5.3** suggest that the information captured by the group of ICs selected at a lower kurtosis threshold is approximately the same as the information captured at higher thresholds. Therefore, a lower number of SNPs can be selected using the ICA framework and using smaller thresholds, without significantly affecting the reconstruction accuracies.

The results demonstrate that ICA is able to identify a small set of SNPs that capture most of the information in the ABCB1 dataset as shown by the high reconstruction accuracies.

The final line in Figure **5.3**, with no marker, depicts the accuracies generated by the combined PCA selection/reconstruction algorithm. The results are clearly less accurate than the ICA framework. However, compared to the reported results in [41], they show slight improvement. With 10 and 23 selected SNPs, we report 95.6% and 99.7% compared to 80% and 95% reported in [41], respectively. The source of this improvement may be in the reconstruction algorithm which was able to use the information captured by the htSNPs to more effectively rebuild the non-chosen SNPs.

It is important to note that the results with this dataset are better than those with the previous dataset due to the size of the data and the linkage disequilibrium at the ABCB1 gene location (for details, see section 5.1.1). Since the data size has a significant impact on training, a larger dataset may further improve the reported results on the previous dataset.

5.1.3 IBD 5q3 Dataset Results

Using this dataset, the ICA selection results with varying kurtosis threshold from 1% to 100% are reported in order to demonstrate that changing the kurtosis threshold does not provide significant changes in the reconstruction accuracies.

In Figure **5.4**, the solid line illustrates the ICA framework reconstruction accuracies generated while adjusting the kurtosis threshold from 1% to 100%. Each point on the line represents the average accuracy calculated for the corresponding number of SNPs. These results show a very similar trend in accuracies to those reported in [41]. Specifically, the calculated accuracies increase rapidly over the range of 1-23 selected htSNP, increasing from 79% to 96.5%. Similarly, the rapid increase in accuracies reported in [41] extends over the range of 1-30 htSNPs from 67%

to 95%. Furthermore, we reached an accuracy of 91.1% with 10 selected SNPs out of 103 SNPs in the dataset compared to 90% accuracy generated in [41] with the same number of htSNPs.

Reconstruction accuracies in this dataset did not show a significant improvement using the ICA framework over the PCA approach. One interpretation of this absence of significant improvement is that the information provided by this dataset doesn't hold stronger associations among SNPs other than the correlations detected by PCA. Therefore, even the combined PCA selection/reconstruction approach results, depicted using the dotted line in Figure **5.4**, follow the same trend as the other two approaches with almost identical accuracy ranges.

In addition, as previously mentioned, the applied ICA framework performs a preprocessing step, in order to aid in decreasing the dimensionality and to try to transform the data into a square matrix. The transformation of the shape of the dataset improves the speed of estimation of the ICs [81]. However, the IBD5q3 dataset partitions have a nearly square shape. As such, the preprocessing step could be eliminated for future improvement of the ICA framework on similar datasets.

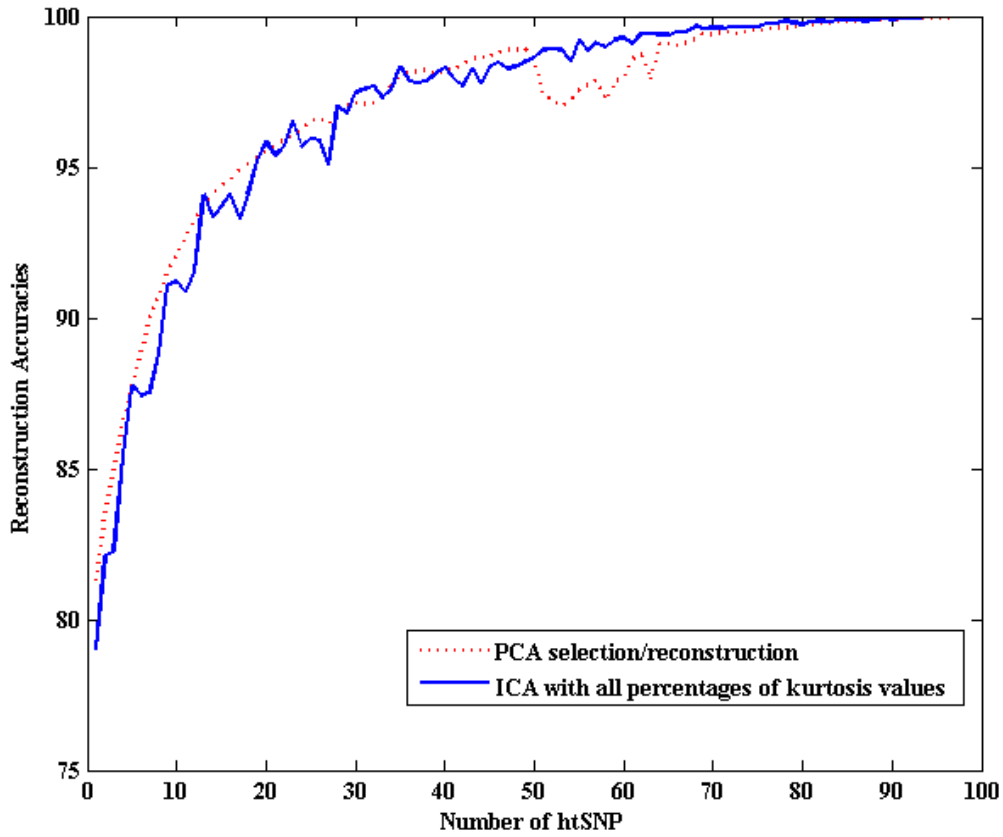

Figure 5.4 The IBD 5q3 ICA framework and combined PCA reconstruction/selection results

In summary, the results of the ICA framework have shown consistent improvements in reconstruction accuracies for SNP selection. However, some particular instances of these results were nearly identical to the other two PCA approaches which may be caused by the amount and type of linkage disequilibrium existing in the datasets under investigation. Furthermore, the higher reconstruction accuracies generated by the ICA framework, show that the selected htSNPs capture more information and are able to better regenerate the haplotypes in the study. In terms of speed, the ICA framework analyzed the three datasets on the order of seconds, but this was still slower than the implementation of mFOS .

5.2 Results of SNP Selection using mFOS

5.2.1 ACE Dataset Results

We applied mFOS and leave-one-out cross validation on this data in order to obtain a model for each of the 22 haplotypes in the dataset. The number of chosen SNPs for each model slightly differed from one chromosome to the other, yet there were many common SNPs chosen. The number of selected SNPs for each model ranged from 3 to 14 (out of 52 SNPs in the dataset). The histogram of the frequencies of the selected SNPs for all the models is shown in Figure **5.6**.

By inspecting this histogram, we recognize a significant number of SNPs were never selected by mFOS to model any genetic signal. These non-selected SNPs are more concentrated in the middle of the histogram in the range 19 to 40, a fact that reflects high redundancy in the information carried along this region. However, the highly frequent selected SNPs such as SNP 5, 8, 11 and 16 have greater contributions in the amount of information hidden in the studied genetic region.

We chose seven cut-off values for the frequencies of SNPs (30, 27, 25, 18, 13 and 11) on the histogram. At each cutoff, SNPs with frequencies above the threshold cutoffs were kept, and all non-chosen SNPs in the 22 haplotypes in the data set were reconstructed. The average accuracies associated with reconstructing all haplotypes are presented in Figure **5.5**.

The average accuracies ranged from 89.9% to 93.9% while using only 3 and 4 SNPs (out of 52), respectively. A slight decrease in reconstruction accuracies was observed when greater than 4 SNPs were chosen. One explanation for this reduction is that the information captured by additional SNPs after the first four are either redundant or they dilute the model. From Figure 2, it is obvious that mFOS was able to build accurate models to capture the information hidden in the group of 52 SNPs.

The models were able to capture at the very least 89.9% of the information in the dataset by using only 3 SNPs. This is a significant reduction in the dimensions of the dataset.

Comparing the reconstruction accuracies of mFOS on the ACE dataset with those generated by the SNP selection method using ICA, we see that the 3 htSNPs selected by mFOS were able to reconstruct 90% of the dataset; whereas the 3 htSNPs selected by ICA have reconstructed only 77.5% of the dataset. The highest reconstruction accuracy reported by mFOS was 94 % with 4 htSNPs and the same degree of accuracy was reached by ICA with 9 htSNPs. The performance of mFOS on this dataset suggests that modeling the haplotype has captured a higher amount of information compared to the information captured by the ICs.

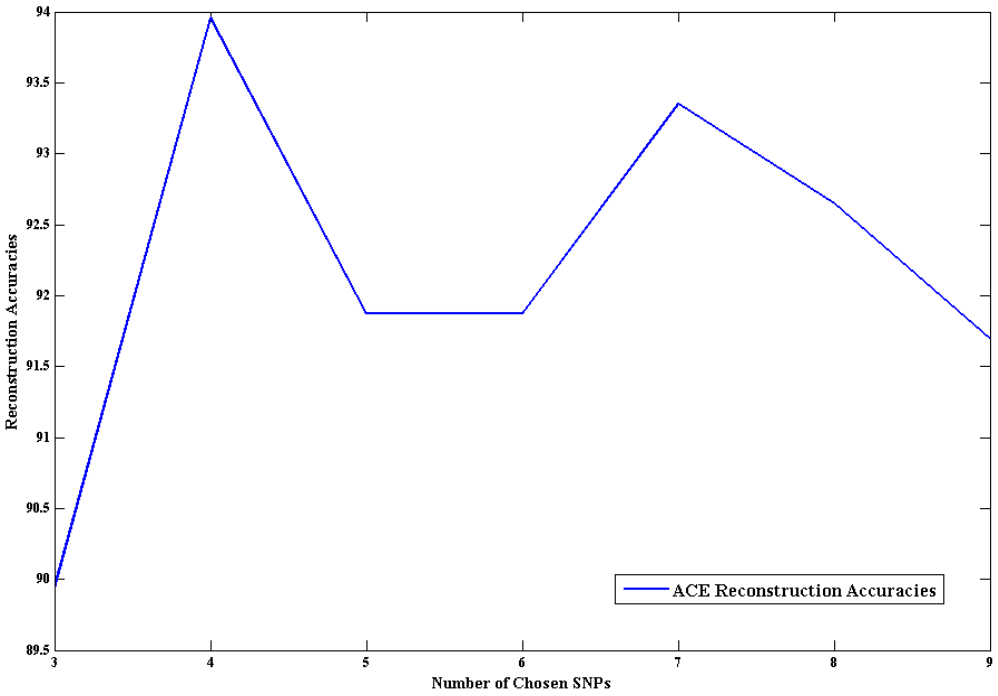

Figure 5.5 The ACE reconstruction accuracies using mFOS

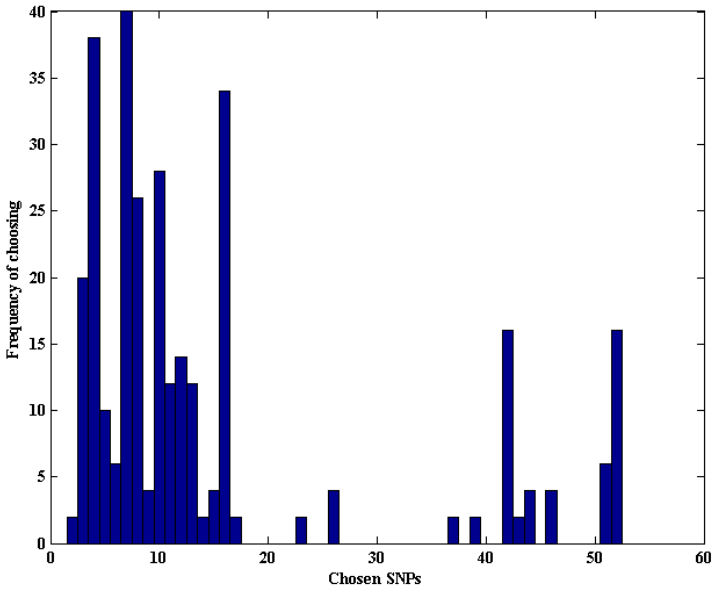

Figure 5.6 The ACE histogram of SNP frequencies using leave-one-out cross-validation.

5.2.2 ABCB1 Dataset Results

For the second dataset, we implemented mFOS with 10 fold cross-validation. The approach used with this dataset is similar to that used for the first dataset. The number of chosen SNPs by mFOS models ranged from 6 to 27 SNPs out of 27 available in the dataset. This implies that the information relayed by the SNPs in this dataset have less redundancy; as a result a higher number of SNPs are needed to model the haplotypes in this dataset compared to with the previous dataset.

Figure **5.7** presents a histogram of the frequencies of selected SNPs. Four frequency cut-offs (at 7, 6, 5, and 4) were selected for the histograms. For each cutoff, SNPs with frequencies higher than the threshold were selected and used to reconstruct the non-chosen SNPs. Average

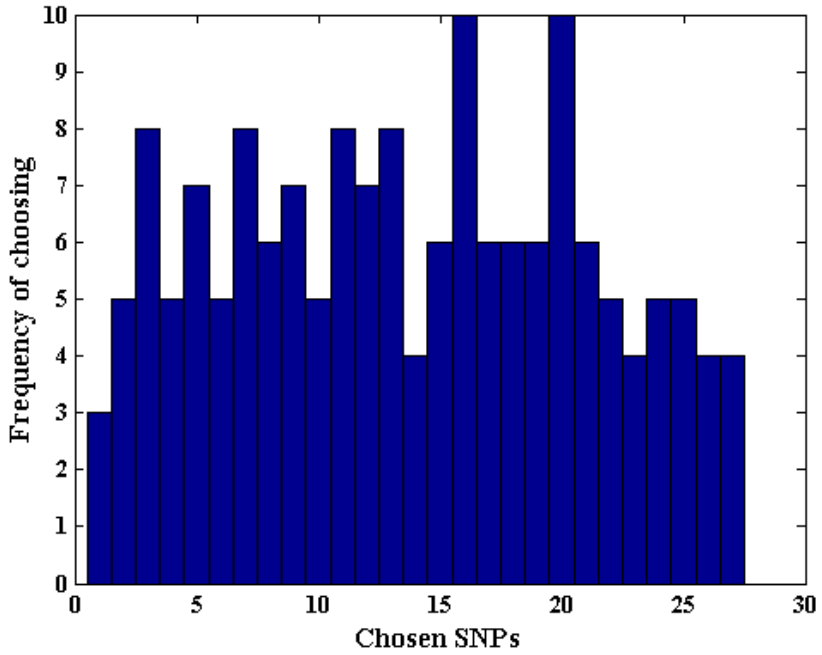

Figure 5.7 The ABCB1 histogram of SNPs frequencies during 10-fold cross-validation.

reconstruction accuracies for all haplotypes using a different number of selection SNPs are reported in Figure 5.8. The lowest reported accuracy is 94.9% when using 6 chosen SNPs out of the 27 available. The highest reported accuracy for was 99.3% with 18 chosen SNPs out of 27.

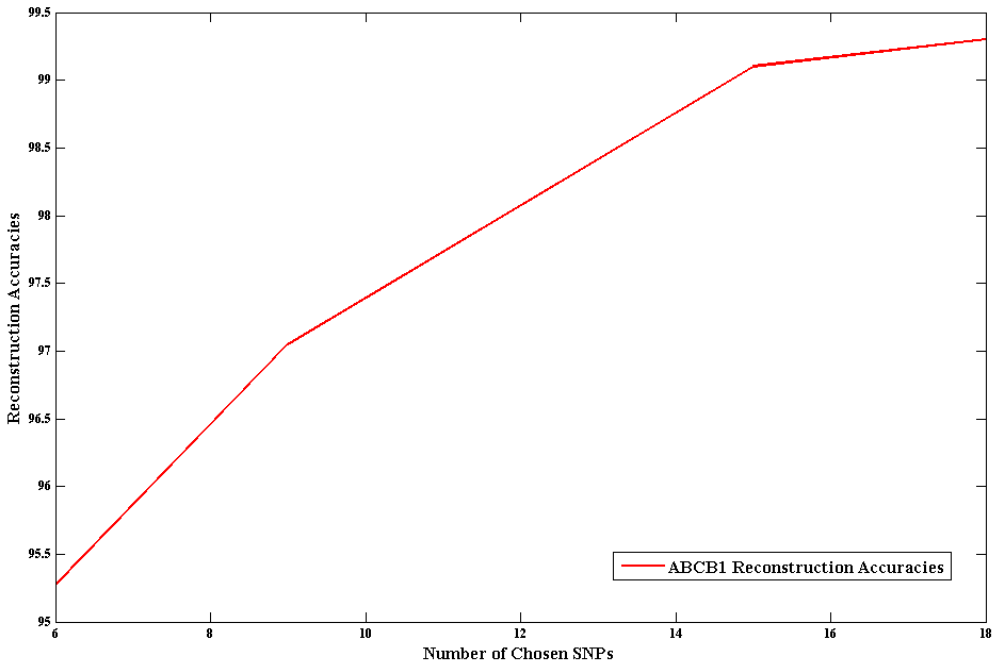

Figure 5.8 The ABCB1 reconstruction accuracies using mFOS

These accuracies reveal once again the ability of mFOS to model the information hidden in the dataset while reducing its dimensionality.

When compared to the reconstruction accuracies of the ABCB1 dataset generated by the SNP selection method using ICA, mFOS produces similar results. However, the lowest reduction in the dimension of the ABCB1 dataset was reported in the results generated by mFOS, where 6 htSNPs out 27 produced 95.4% accuracy of reconstruction. The smallest htSNP group reported by ICA consists of 9 SNPs out of 27 and this group was able to reconstruct 97.8% of the dataset. Therefore, a higher reduction in the dimension of the ABCB1 dataset was reached with the SNP selection method using mFOS. Accordingly, mFOS was able to capture a higher amount of information from this dataset using a fewer number of SNPs. As mentioned in this section, it is important to note that the accuracies reported are for haplotype reconstruction and not for major or minor allele prediction.

Chapter 6

Conclusion

In this thesis, two methodologies for SNP selection were proposed. The first method was based on ICA, which is a matrix decomposition technique borrowed from the signal processing literature. ICA calculated the statistically independent components of the SNP datasets by applying a linear transformation on the studied data. A SNP selection process was performed based on the contributions of each SNP in the most independent and non-Gaussian ICs. Consequently, the group of selected SNPs were characterized by high independence and were able to summarize the information hidden by the whole sequence of SNPs.

The second method investigated for SNP selection was mFOS; a modification of FOS which is a multivariate regression method devised for non-linear system identification. We considered each sequence of SNP values as a genetic system output and used mFOS to model this output for each individual in the study. mFOS searched through the group of studied SNPs and assessed the relevance of each SNP in modeling the output.

Three gene datasets were used for the application of ICA and two datasets were used for the application of mFOS. ICA was combined with a zero/one encoding algorithm to transform the alphabetical format of the SNP data into numerical values. However, the mFOS method was coupled with an encoding method which substitutes each SNP allele (value) by a real number that is the frequency of appearance of that SNP in the studied population.

After performing the SNP selection using either of the proposed methods, the selected group of SNPs was evaluated by measuring its ability to provide information about non-selected SNPs.

The evaluation technique consisted of predicting the values of the non-selected SNPs using the information relayed by the htSNPs. The chosen evaluation criterion was the accuracy of reconstruction and was measured by the ratio of the number of correctly predicted SNPs to the total size of the reconstructed SNP sequence. The reconstruction accuracies of both methods were compared with each other and to those of the PCA published in a previous work. The reported reconstruction accuracies are for the non-tagged SNPs in a haplotype. Major or minor allele prediction is not part of the goal of the proposed approaches.

The group of SNPs selected by mFOS exhibited the highest reconstruction accuracies using a smaller number of SNPs in comparison with the overall ICA and PCA results. For the ACE dataset, a group of 4 htSNPs selected by mFOS generated 94% of reconstruction accuracy, while a group of 10 htSNPs chosen by ICA reconstructed the dataset with 95% accuracy. The results of PCA as reported in the literature showed 90% accuracy with 10 selected SNPs. For the ABCB1 dataset, the smallest group of SNPs (consisting of 6 SNPs) chosen by mFOS was able to regenerate the dataset with more than 95% accuracy. The ICA was not able to choose the same number of chosen SNPs as mFOS, however, it selected a group of 9 htSNP and reached a reconstruction accuracy of more than 97%. The results of the PCA were lower than those generated by both the ICA and mFOS. For the third dataset, the results of ICA and PCA demonstrated a similar trend with nearly equivalent reconstruction accuracies.

mFOS and ICA were able to generate as good as or better results than the PCA reported results. Both of these approaches are considered to be unsupervised dimensionality reduction techniques. Their selection method is independent of any classification process or outcome; as such the computational overhead is decreased while using these techniques as a preprocessing step in a disease association study. The contributions of this thesis are as follows:

- Provided a comparative study of two proposed dimensionality reduction approaches, ICA and mFOS, and a previously published selection method using PCA.
- Presented a SNP filtering technique using ICA, unlike previous studies which used ICA as a wrapper technique. The proposed ICA selection approach was able to reduce the number of SNPs without the need for a classifier or class labeling.
- Offered an application of mFOS for selecting the most informative SNPs in a dataset. This implementation has limited the source of candidate functions used in the modeling of the genetic output system to the actual SNPs in study. Using mFOS, the sequences of SNPs values belonging to the different individuals were modeled according to the values of SNPs in the data.
- Provided a reconstruction approach based on the CUR matrix decomposition which was used for the evaluation of the proposed methods.

6.1 Future work

The proposed approaches were applied to individual gene datasets. Since the ultimate goal of dimensionality reduction techniques is the facilitation of disease association studies, it is essential to study the application of our methods to larger scale datasets. This application can enable searching for possible correlations between certain diseases and genetic multi loci. Therefore, investigating the applicability of our proposed methods on genome wide SNP datasets is an important continuation of the work presented in this thesis.

In addition to the above, an extension of this work could be a faster implementation of both proposed methods for SNP selection. The need to accelerate the proposed algorithms is crucial, especially when applied on genome-wide SNP datasets. One way of accelerating the methods is

the segregation of the genome-wide dataset into separate genes, before performing the selection on each gene. Another way of accelerating the application of the SNP selection on labeled genome-wide datasets is to survey the literature for previously reported associations between the class phenotype and certain genes; afterwards, the selection method will be applied only on the associated genes. Finally, a multistage SNP selection method can be implemented to combine the results obtained from the application of the methods on separate genes.

Another aspect of this work to investigate further is the application of the original implementation of FOS as mentioned earlier in Chapters 2 and 4. FOS can be applied to SNP selection to identify SNPs capable of predicting other SNPs in a dataset. In addition, different SNP encoding schemes could be investigated to enable major or minor allele prediction as another application.

Bibliography

[1] Constantine, Gurrin, McLaren, Bahlo, Anderson, Vulpe, Forrest, Allen, Gertig and the HealthIron Investigators: "SNP selection for genes of iron metabolism in a study of genetic modifiers of hemochromatosis", *BMC Medical Genetics 9 (18), 2008.*

[2] Takeuchi, Yanai, Morii, Ishinaga, Taniguchi-Yanai, Nagano and Kato: "Linkage Disequilibrium Grouping of Single Nucleotide Polymorphisms (SNPs) Reflecting Haplotype Phylogeny for Efficient Selection of Tag SNPs", *Genetics 170 (1) pp. 291-304, 2005.*

[3] McElroy and Oksenberg: "Multiple Sclerosis Genetics", *Advances In Multiple Sclerosis and Experimental Demyelinating Diseases Current Topics in Microbiology and Immunology 318 pp. 45-72, 2008.*

[4] Bertram, McQueen, Mullin, Blacker, Tanzi: "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database", *Nature Genetics 39(1) pp. 17 – 23, 2007.*

[5] Ponder: "Cancer genetics", *Nature 411(6835) pp. 336-341, 2001.*

[6] The International HapMap Consortium: "The International HapMap Project", *Nature 426(6968) pp. 789–79, 2003.*

[7] Thornton-Wellsa, Moorec, and Hainesc: "Genetics, statistics and human disease: analytical retooling for complexity", *Trends in Genetics 20(12) pp. 640-647, 2004.*

[8] Watson, Crick: "A Structure for Deoxyribose Nucleic Acid", *Nature 171 (4356) pp. 737–738, 1953.*

[9] Alberts, Johnson, Lewis, Raff, Roberts and Walters: Molecular Biology of the Cell, *Fourth Edition, New York and London: Garland Science, 2002.*

[10] Human Genome Project Information - SNP Fact-Sheet: (last accessed 10/15/2010) [www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml].

[11] Lawrence: Henderson's Dictionary of Biology, *Pearson Education limited 13th edition, England, 2005.*

[12] Ragoussis: "Genotyping Technologies for Genetic Research", *The Annual Review of Genomics and Human Genetics 10 pp. 117–33, 2009.*

[13] Smith: Genetic Polymorphism and SNPs: Genotyping, Haplotype Assembly Problem, Haplotype Map, functional Genomics and Proteomics *[online], 2002.*

[14] National Human Research Institute, National Institute of health - Genome Wide Association studies: (last accessed 10/15/2010) [www.genome.gov/20019523].

[15] Baranzini, Wang, *et al*.: "Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis", *Human Molecular Genetics 18(4) pp. 767-778, 2009*.

[16] The International Headache Genetics Consortium, Anttila, Stefansson, *et al*.: "Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1", *Nature Genetics 42(10) pp. 869–873, 2010*.

[17] Azzato, Pharoah, Harrington, Easton, Greenberg, Caporaso, Chanock, Hoover, Thomas, Hunter, and Kraft: "A Genome-Wide Association Study of Prognosis in Breast Cancer", *Cancer Epidemiol Biomarkers 19(4) pp. 1140-1143, 2010*.

[18] Mathias, Grant *et al*.: "A genome-wide association study on African-ancestry populations for asthma", *Journal of Allergy and Clinical Immunology 125(2) pp. 336-346.e4, 2010*.

[19] Eeles, Kote-Jarai, Al Olama, Giles, Guy, Severi, et al.: "Identification of seven new prostate cancersusceptibility loci through a genome-wide association study", *Nature Genetics 41(10) pp. 1116–1121, 2009*.

[20] Manolio: "Genomewide Association Studies and Assessment of the Risk of Disease", *New England Journal of Medicine 363(2) pp. 166-176, 2010.*

[21] Hardy, Singleton: “Genomewide Association Studies and Human Disease”, *New England Journal of Medicine 360(17) pp. 1759-1768, 2009*.

[22] Pearson, Manolio: “How to Interpret a Genome-wide Association Study”, T*he journal of American Medical Association 299(11) pp. 1335-1344, 2008*.

[23] Seshadri, Fitzpatrick *et al*.: “Genome-wide analysis of genetic loci associated with Alzheimer disease’, *The journal of the American Medical Association 303(18) pp. 832-40, 2010.*

[24] Kramer, Xu, *et al*.: “Alzheimer disease pathology in cognitively healthy elderly: A genome-wide study” *ElSevier, The Neurobiology of Aging, published online, 2010*.

[25] Harold , Abraham , *et al*.: “Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease”, *Nature genetics 41(10) pp.1088-93, 2009.*

[26] Long, Cai, *et al*.: “Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium”, *PLoS Genetics 6(6) pp. e1001002, 2010*.

[27] Turnbull, Ahmed, *et al*.: “Genome-wide association study identifies five new breast cancer susceptibility loci”, Nature Genetics *42(6) Pages 504-7, 2010*.

[28] Thomas, Jacobs, *et al*.: “A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1)”, *Nature Genetics 41(5) pp. 579-584, 2010*.

[29] Nischwitz, Cepok, *et al*.: “Evidence for VAV2 and ZNF433 as Susceptibility Genes for Multiple Sclerosis”, *Journal of Neuroimmunology, 227(1-2) pp. 162-6, 2010.*

[30] Sanna, Pitzalis, *et al*.: "Variants within The Immunoregulatory CBLB Gene are Associated with Multiple Sclerosis", *Nature Genetics 42(6) pp. 495-497, 2010*.

[31] Jakkula, Leppä, *et al*.: "Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene", *American Journal of Human Genetics 86(2) pp. 285-291, 2010*.

[32] I. Fodor: "A Survey of Dimension Reduction Techniques", *US Department of Energy - Lawrence Livermore National Laboratory Technical Reports, published online, 2010.*

[33] A science primer, Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources. SNPs: Variations on a Theme. (last accessed 10/15/2010) [www.ncbi.nlm.nih.gov/About/primer/snps.html]

[34] Affymetrix Inc." Data Sheet: Affymetrix® Genome-Wide Human SNP Array 6.0" *©2007 Affymetrix, Inc. Part No. 702509 Rev. 1*

[35] Chen, Lin, Sabatti: "Volume measures for linkage disequilibrium", *BMC Genetics 7 pp. 54, 2006.*

[36] Lin, Altman: "Choosing SNPs Using Feature Selection", *Proceedings of IEEE Computational Systems Bioinformatics pp. 301 – 309, 2005*.

[37] Saeys, Inza, Larranaga: "A review of feature selection techniques in bioinformatics", *Bioinformatics 23(19) pp. 2507-2517, 2007*.

[38] Lee, Shatkay: "BNTagger: improved tagging SNP selection using Bayesian networks", *Bioinformatics 22(14) pp. e211-e219, 2006*.

[39] Halldórsson, Bafna, Lippert, Schwartz, De La Vega, Clark, Istrail: "Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies", *Genome Research 14 pp. 1633-1640, 2004*.

[40] Halperin, Kimmel, Shamir: "Tag SNP selection in genotype data for maximizing SNP prediction accuracy", *Bioinformatics 21(Suppl 1) pp. i195–i203, 2005*.

[41] Z. Lin and R. B. Altman: "Finding Haplotype Tagging SNPs by use of Principal Component Analysis", *The American Journal of human Genetics 75(5) pp. 850-861, 2004.*

[42] Chuang, Hou Jr, Yang: "A Novel Prediction Method for Tag SNP Selection using Genetic Algorithm based on KNN", *International Journal of Chemical and Biomolecular Engineering 3(1) pp. 12, 2010.*

[43] Shah, Kusiak: "Data mining and genetic algorithm based gene/SNP selection", *Artificial Intelligence in Medicine 31(3) pp. 83-196, 2004*.

[44] Yang, Zhang: "A Hybrid Approach to Selecting Susceptible Single Nucleotide Polymorphisms", *in the proceedings of International Conference on BioMedical Engineering and Informatics, 2008*.

[45] Zhang, Li, Viktorovich, Lei: "A Heuristic Approach for Target SNP Mining Based on Genome-Wide IBD Profile", *in the proceedings of Third International Conference on Natural Computation ,2007.*

[46] Dawy, Sarkis, Hagenauer, Mueller: "A Novel Gene Mapping Algorithm Based Om Independent Component Analysis", *in the proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*.

[47] Long, Gianola, Rosa, Weigel, Avendano: "Comparison of Classification Methods for Detecting Associations between SNPs and Chick Mortality", *Genetics Selection Evolution 41 pp. 18-32, 2009.*

[48] Liu, Yang, Chen, Yang, Sung, Huang: "Supervised learning-based tagSNP selection for genome-wide disease classifications", *in the proceedings of the International Conference on Bioinformatics & Computational Biology, 2008*.

[49] Herault, Jutten: "Space or time adaptive signal processing by neural network models", *American Institute of Physics Conference Proceedings 151 on Neural Networks for Computing, 1987*.

[50] Comon: "Independent Component Analysis, a New Concept?" *Signal Processing Elsevier 36(3) pp. 287-314, 1994*.

[51] Hyvarinen: "Survey on Independent Component Analysis", *Neural Computing Surveys 2 pp. 94-128, 1999.*

[52] Fodor: "A survey of dimension reduction techniques", *Livermore, CA: US DOE Office of Scientific and Technical Information 18, 2002.*

[53] Pearson: "On Lines and Planes of Closest Fit to Systems of Points in Space", *Philosophical Magazine 2(6) pp. 559–572, 1901*.

[54] Cha, Chan: "Applying independent component analysis to factor model in finance", *Intelligent Data Engineering and Automated Learning—IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents, Springer pp. 161-173, 2000*.

[55] Back: "A First Application of Independent Component Analysis to Extracting Structure from Stock Returns", *International Journal of Neural Systems (Special Issue on Data Mining in Finance) 8 (5) pp. 473-484, 1997.*

[56] Tonga, Bezerianosa, Paula, Zhub, Thako: "Removal of ECG interference from the EEG Recordings in Small Animals using Independent Component Analysis", *Journal of Neuroscience Methods 108(1) pp. 11-17, 2001.*

[57] Barbati, Porcaro, Zappasodi, Rossini, Tecchio: "Optimization of an independent component analysis approach for artifact identification and removal in magnetoencephalographic signals", *Clinical Neurophysiology 115(5) pp. 1220-1232, 2004.*

[58] Lotsch, Friedl, Pinzón: "Spatio-temporal deconvolution of NDVI image sequences using independent component analysis", *IEEE Transactions on Geoscience and Remote Sensing 41 (12) pp. 2938–2942, 2003.*

[59] Hung, Lee, Wu, Chen, Yeh, Hsieh: "Recognition of Motor Imagery Electroencephalography Using Independent Component Analysis and Machine Classifiers", *Annals of Biomedical Engineering 33(8) pp. 1053–1070, 2005*.

[60] Serdaroglu, Ertuzun, Ercil: "Defect Detection in Textile Fabric Images Using Wavelet Transforms and Independent Component Analysis", *Pattern Recognition and Image Analysis 16(1) pp. 61–64, 2006.*

[61] Korenberg: "Orthogonal identification of nonlinear difference equation models", *in the Proceedings Midwest Symposium Circuit Systems, 1985.*

[62] Chen, Billings, Luo: "Orthogonal least squares methods and their application to non-linear system identification", *International Journal of Control 50(5) pp. 1873-1896, 1989*.

[63] Mao: "Fast Orthogonal Forward Selection Algorithm for Feature Subset Selection" *IEEE Transactions on Neural Networks 13(5) pp. 218 - 1224, 2002*.

[64] Wu, Sun, Krieger, Sclabassi: "Comparison of Orthogonal Search and Canonical Variate Analysis for the Identification of Neurobiological Systems", *Annals of Biomedical Engineering 27(5) pp. 592–606, 1999*.

[65] Korenberg: "Identifying Nonlinear Difference Equation and Functional Expansion Representations: The Fast Orthogonal Algorithm", *Annals of Biomedical Engineering 16(1) pp. 123-142, 1988.*

[66] Korenberg, Paarmann: "Applications of Fast Orthogonal Search: Time-Series Analysis and Resolution of Signals in Noise", *Annals of Biomedical Engineering, 17(3) pp. 219-231, 1989.*

[67] Minz, Korenberg: "Modeling Cooperative Gene Regulation Using Fast Orthogonal Search", *The Open Bioinformatics Journal 2 pp. 80-89, 2008.*

[68] Mostafavi, Baranzini, Oksenberg, Mousavi: "A Fast Multivariate Feature-Selection/Classification Approach for Prediction of Therapy Response in Multiple Sclerosis*", in the proceedings of IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'06),2006*

[69] Rieder, Taylor, Clark, Nickerson: "Sequence Variation in the Human Angiotensin Converting Enzyme" *Nature Genetics 22 (1) pp. 59-62, 1999.*

[70] J. Riordan: "Angiotensin-I-converting enzyme and its relatives", *Genome Biology 4(8) pp. 225, 2003.*

[71] Kroetz, Pauli-Magnus, Hodges, Huang, Kawamoto, Johns, Stryke, Ferrin, DeYoung, Taylor, Carlson, Herskowitz, Giacomini, Clark; Pharmacogenetics of Membrane Transporters Investigators: "Sequence diversity and haplotype structure in the human ABCB1 MDR1, multidrug resistance transporter gene", *Pharmacogenetics and Genomics 13(8) pp. 481- 494, 2003.*

[72] Ueda, Clark, Chen, Roninson, Gottesman, Pastan: "The human multidrug resistance (mdr1) gene. cDNA cloning and transcription initiation", *The Journal of Biological Chemistry 262 pp. 505-508, 1987* .

[73] Ruiz, Choi, von Hoff, Roninson, Wahl: "Autonomously replicating episomes contain mdr1 genes in a multidrug-resistant human cell line", *Molecular and cellular biology 9(1) pp. 109-115, 1989*.

[74] Riordan, Deuchars, Kartner, Alon, Trent, Ling: "Amplification of P-glycoprotein genes in multidrug-resistant mammalian cell lines", *Nature 316 pp. 817 – 819, 1985.*

[75] Dean, Hamon, Chimini: "The Human ATP-Binding Cassette (ABC) Transporter Superfamily", *The Journal of Lipid Research 42 pp. 1007-1017, 2001*.

[76] Daly, Rioux, Schaffner, Hudson, Lander: "High-resolution haplotype structure in the human genome", *Nature Genetics 29 pp. 229–232, 2001.*

[77] Richard Steven Blumberg and Markus Neurath: "Immune mechanisms in inflammatory bowel disease", *Advances in Experimental Medicine and Biology 479 pp. 26, 2006.*

[78] Armuzzi, Ahmad, Ling, de Silva, Cullen, van Heel, Orchard, Welsh, Marshall, Jewell: "Genotype-phenotype analysis of the Crohn's disease susceptibility haplotype on chromosome 5q31", *GUT an International Journal of Gastroenterology and Hepatology 52 pp. 1133–1139, 2003.*

[79] Giallourakis, Stoll, Miller, Hampe, Lander, Daly, Schreiber, Rioux: "IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of a novel association with ulcerative colitis", *American Journal of Human Genetics 73(1) pp. 205-211, 2003.*

[80] David Lane: “HyperStat Online Statistics Textbook”, (last accessed 10/15/2010) [davidmlane.com/hyperstat/index.html].

[81] Hyvarinen, Oja: “A Fast Fixed-Point Algorithm for Independent Component Analysis” *MIT Press Journal, Neural Computation 9(7) pp.1483–1492, 1997.*

[82] Hyvarinen, Karhunen, Oja: Independent Component Analysis, *a Wiley-Interscience Publication, New York, 2001.*

[83] Li, Chen, Juan, Fleisher, Reiman, Yao, and Wu: “Combining fMRI and SNP Data to Investigate Connections between Brain Function and Genetics Using Parallel ICA”, *Human Brain Mapping 30(1) pp. 241–255, 2009.*

[84] Scholz, Gatzek, Sterling, Fiehn, Selbig: “Metabolite fingerprinting: detecting biological features by independent component analysis”, *Bioinformatics 20(15) pp. 2447–2454, 2004.*

[85] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2010 *(online).*

[86] Paschou, Mahoney, Javed, Kidd, Pakstis, Gu, Kidd, Drineas: “Intra- and interpopulation genotype reconstruction from tagging SNPs”, *Genome Research 17(1) pp. 96-107, 2007.*

[87] Mahoneya, Drineas: “CUR matrix decompositions for improved data analysis”, *in the Proceedings of the National Academy of Science 106(3) pp. 697–702, 2009.*

Appendix A
FOS Coefficient Derivation

FOS is the fast implementation of OS which avoids the actual calculation of orthogonal functions and settles for the calculation of the simplified coefficients D, C, α and Q according to the following set of equations [66]:

$$g_m = \frac{C(m)}{D(m,m)}\ ,\ \forall\ m\ \in \{0, \dots, M\} \tag{A. 1}$$

where Ds are calculated as follows:

$$D(0,0) = 1 \tag{A. 2}$$

$$\boldsymbol{D(m,0) = \overline{p_m(n)},\ \forall\ m\ \in \{0, \dots, M\}} \tag{A. 3}$$

$$\boldsymbol{D(m,r) = \overline{p_m(n)p_r(n)} - \sum_{i=0}^{r-1} \alpha_{ri} D(m,i),\quad \begin{cases} \forall\ m\ \in \{1, \dots, M\} \\ \forall\ r\ \in \{1, \dots, m\} \end{cases}} \tag{A. 4}$$

and where αs are:

$$\alpha_{mr} = \frac{D(m,r)}{D(r,r)},\quad \begin{cases} \forall\ m\ \in \{1, \dots, M\} \\ \forall\ r\ \in \{0, \dots, m-1\} \end{cases} \tag{A. 5}$$

Furthermore, Cs are calculated according to the next equations:

$$\boldsymbol{C(0) = \overline{y(n)}} \tag{A. 6}$$

$$\boldsymbol{C(m) = \overline{y(n)p_m(n)} - \sum_{r=0}^{m-1} \alpha_{mr} C(r),\quad \forall\ m\ \in \{1, \dots, M\}} \tag{A. 7}$$

$Q(m)$, the amount each orthogonal function deducts from MSE and is calculated in the corresponding inner product space according to the following equations:

$$Q(m) = {g_m}^2 \, \overline{{w_m}^2(n)}, \quad \forall\, m \in \{1, \dots, M\} \tag{A. 8}$$

$$\boldsymbol{Q(m) = {g_m}^2 D(m, m)}, \quad \boldsymbol{\forall\, m \in \{1, \dots, M\}} \tag{A. 9}$$

$$\boldsymbol{Q(m) = \frac{C^2(m)}{D(m,m)}}, \quad \boldsymbol{\forall\, m \in \{1, \dots, M\}} \tag{A. 10}$$

As a result orthogonal functions having highest $Q(m)s$ are added to the system output model as contributors ensuring the best fit of the estimation which minimizes the mean square error (MSE) between the actual and estimated system outputs as follows:

$$\boldsymbol{MSE = \overline{\left[y(n) - \sum_{m=0}^{M} g_m \, w_m(n)\right]^2}} \tag{A. 11}$$

$$\boldsymbol{MSE = \overline{y(n)^2} - \sum_{m=0}^{M} Q_m} \tag{A. 12}$$

Appendix B

Tolerance Threshold of Reconstruction Error Bars

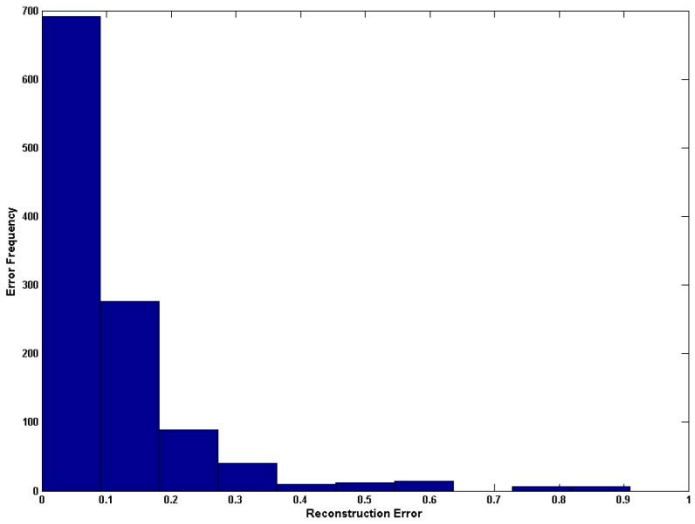

Figure B.1 Histogram showing the frequency on y-axis of getting the reconstruction errors on the x-axis using 3 htSNPs for the ACE Dataset

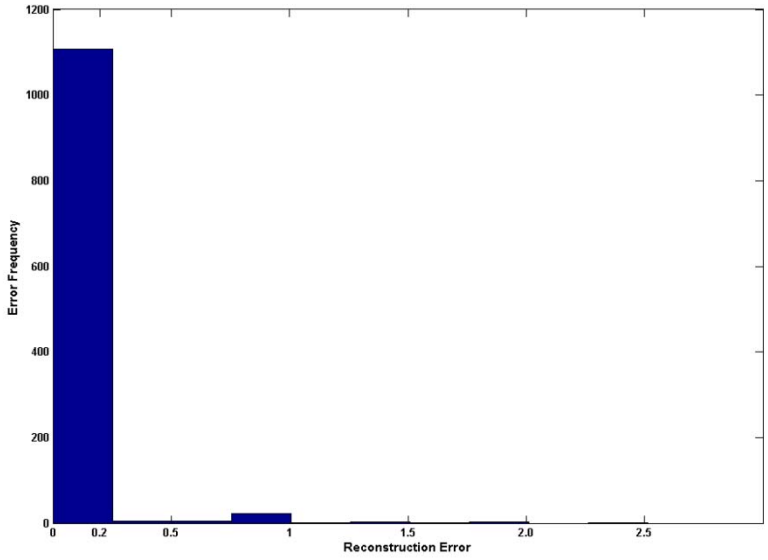

Figure B.2 Histogram showing the frequency on y-axis of getting the reconstruction errors on the x-axis using 5 htSNPs for the ACE Dataset

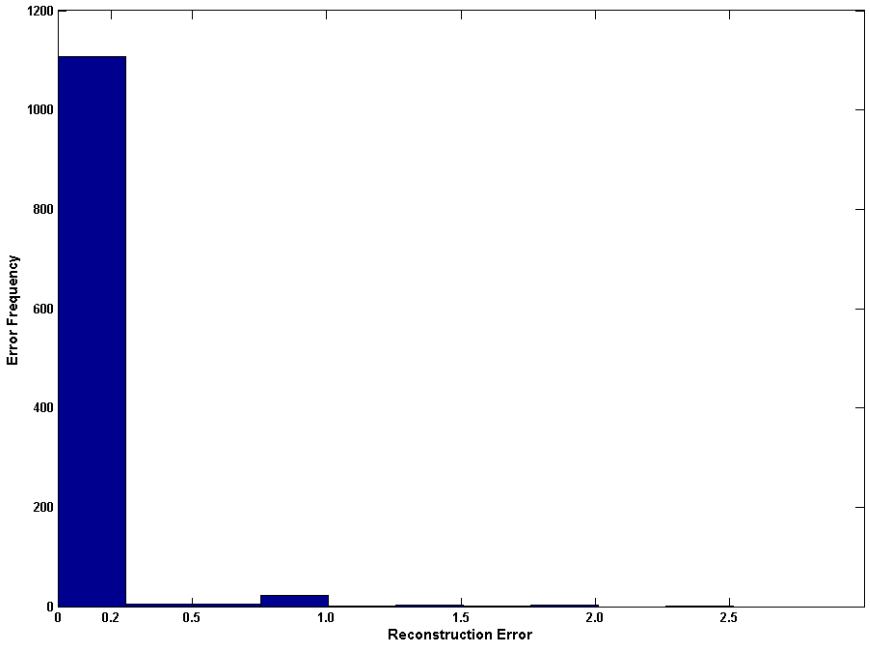

Figure B.3 Histogram showing the frequency on y-axis of getting the reconstruction errors on the x-axis using 6 htSNPs for the ACE Dataset

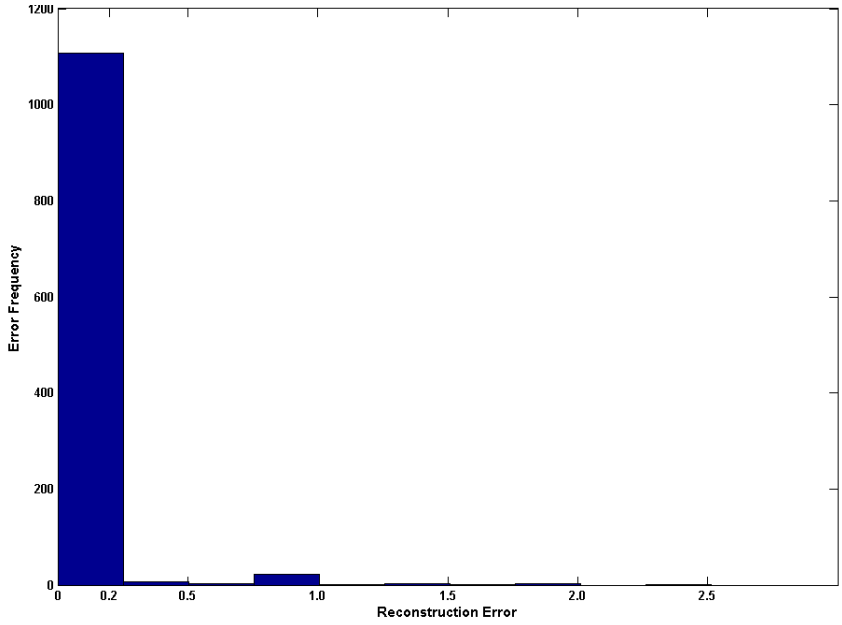

Figure B.4 Histogram showing the frequency on y-axis of getting the reconstruction errors on the x-axis using 7 htSNPs for the ACE Dataset

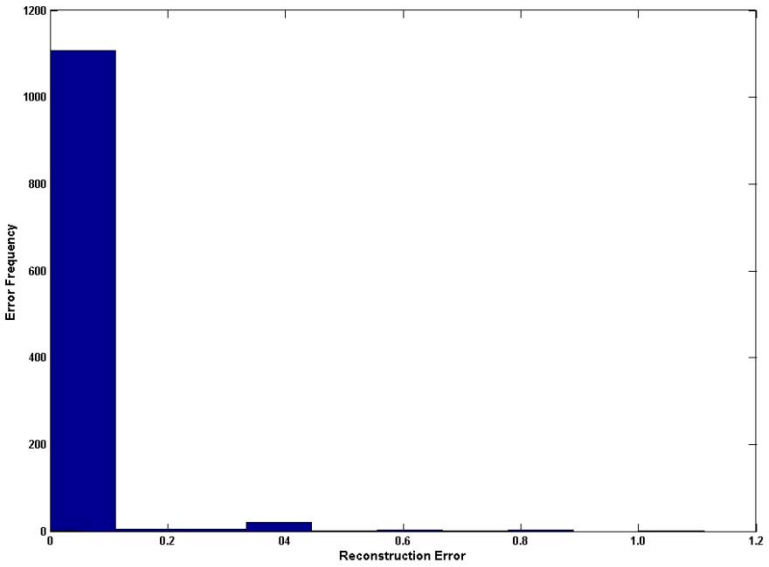

Figure B.5 Histogram showing the frequency on y-axis of getting the reconstruction errors on the x-axis using 8 htSNPs for the ACE Dataset

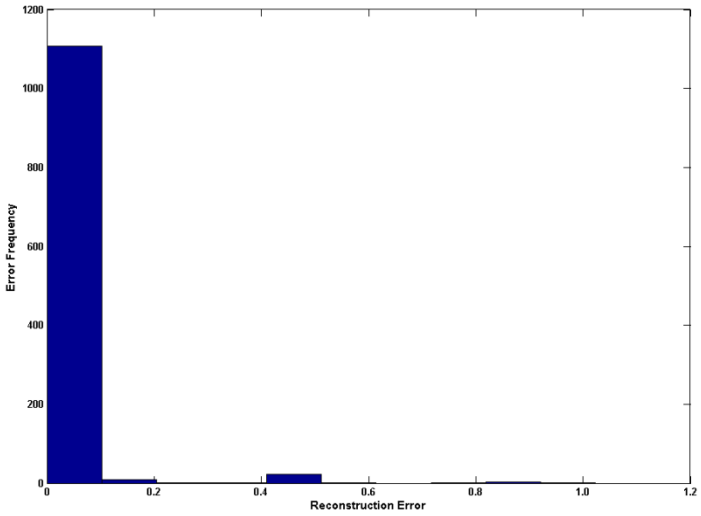

Figure B.6 Histogram showing the frequency on y-axis of getting the reconstruction errors on the x-axis using 9 htSNPs for the ACE Dataset

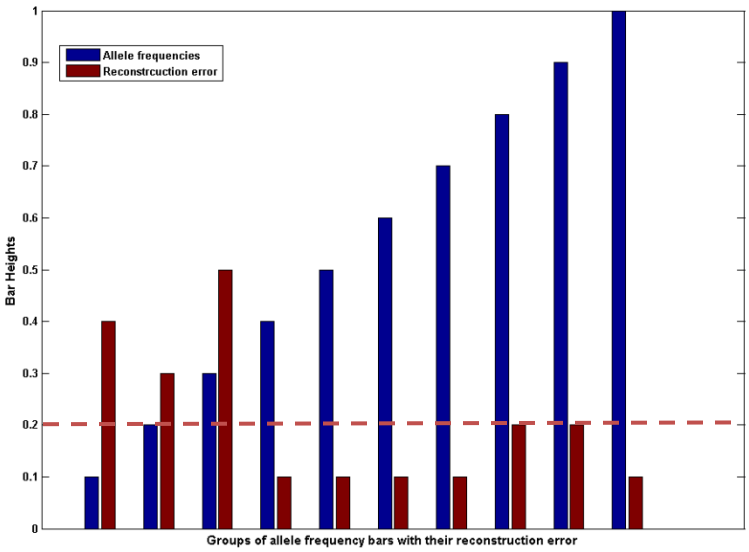

Figure B.7 Bar chart showing the different allele frequencies of the ACE dataset and their reconstruction errors using 3 htSNPs. The heights of the blue bars represent the allele frequencies and the heights of the red bars represent their errors. Each allele frequency bar is grouped with its reconstruction error bar in the same location on the x-axis.

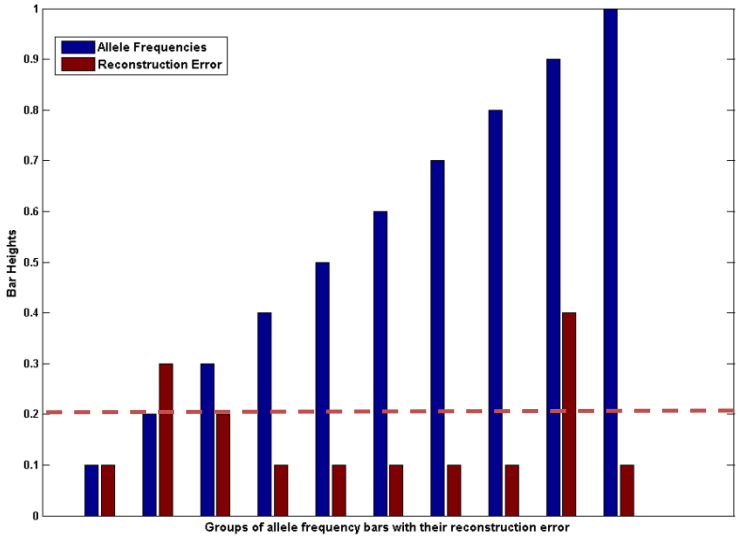

Figure B.8 Bar chart showing the different allele frequencies of the ACE dataset and their reconstruction errors using 5 htSNPs. The heights of the blue bars represent the allele frequencies and the heights of the red bars represent their errors. Each allele frequency bar is grouped with its reconstruction error bar in the same location on the x-axis.

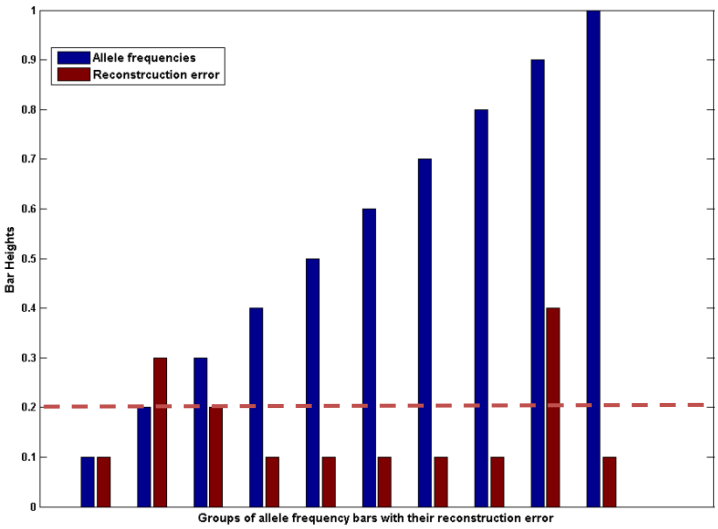

Figure B.9 Bar chart showing the different allele frequencies of the ACE dataset and their reconstruction errors using 6 htSNPs. The heights of the blue bars represent the allele frequencies and the heights of the red bars represent their errors. Each allele frequency bar is grouped with its reconstruction error bar in the same location on the x-axis.

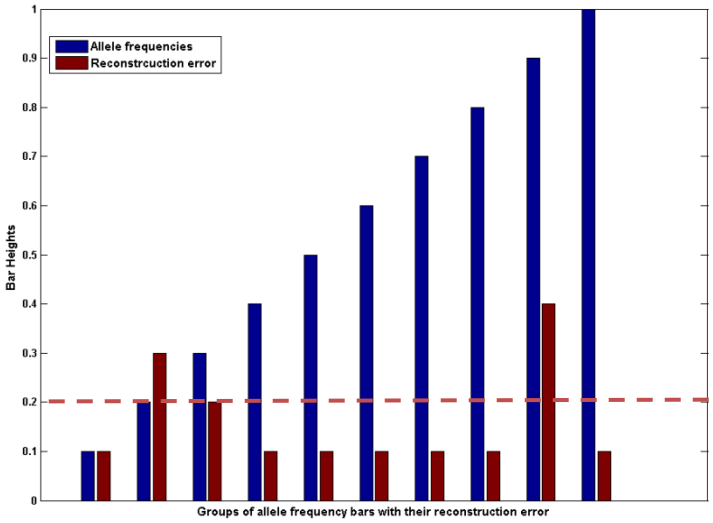

Figure B.10 Bar chart showing the different allele frequencies of the ACE dataset and their reconstruction errors using 7 htSNPs. The heights of the blue bars represent the allele frequencies and the heights of the red bars represent their errors. Each allele frequency bar is grouped with its reconstruction error bar in the same location on the x-axis.

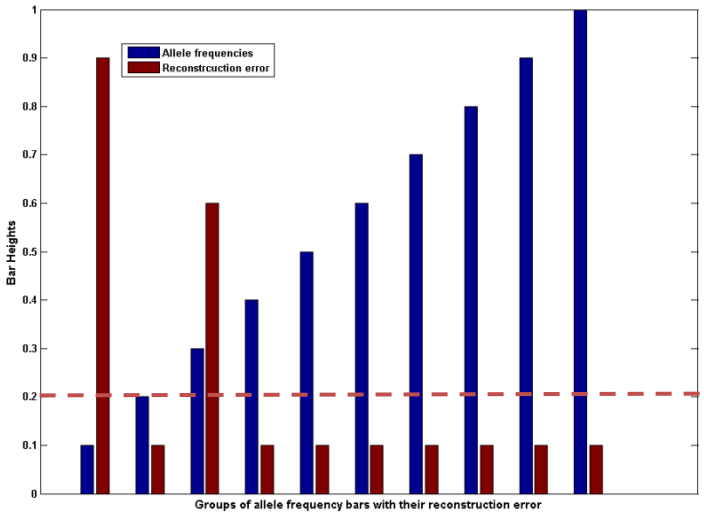

Figure B.11 Bar chart showing the different allele frequencies of the ACE dataset and their reconstruction errors using 8 htSNPs. The heights of the blue bars represent the allele frequencies and the heights of the red bars represent their errors. Each allele frequency bar is grouped with its reconstruction error bar in the same location on the x-axis.

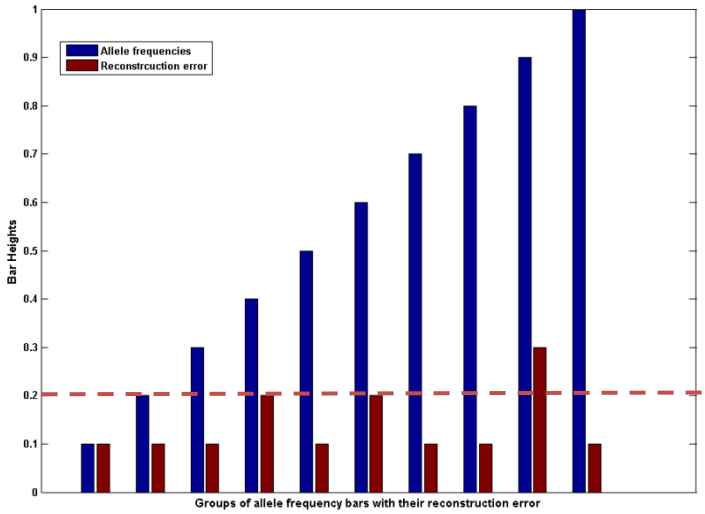

Figure B.12 Bar chart showing the different allele frequencies of the ACE dataset and their reconstruction errors using 9 htSNPs. The heights of the blue bars represent the allele frequencies and the heights of the red bars represent their errors. Each allele frequency bar is grouped with its reconstruction error bar in the same location on the x-axis.