University of Alberta

ASSESSING THE FEASIBILITY OF LEARNING BIOMEDICAL PHENOTYPE PATTERNS USING HIGH-THROUGHPUT OMICS PROFILES

by

Mohsen Hajiloo

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

Spring 2014
Edmonton, Alberta

Abstract

A decade after the completion of the human genome project, the rapid advancement of the high-throughput measurement technologies has made omics (genomics, epigenomics, transcriptomics, metabolomics) profiling feasible. The availability of such omics profiles has raised the hope for the development of more accurate disease models that will help improve the existing clinical strategies for disease prevention, diagnosis, prognosis, and treatment. Revealing the hidden pattern of diseases based on high-throughput omics profiles is only feasible if we choose the appropriate informatics techniques. While the basic univariate statistical analysis techniques are applicable to some extent within the reductionist paradigm of disease studies, supervised machine learning techniques are relevant in the systems biology paradigm of disease studies. This dissertation utilizes such machine learning techniques and foundations to analyze, experimentally and analytically, the feasibility of learning breast cancer and ancestral origins based on a genome wide scan of single nucleotide polymorphisms. In the former task, using a dataset from Alberta with 696 samples (348 breast cancer cases and 348 controls) over 900K features, we achieved 59.55% leave-one-out cross validation accuracy in breast cancer susceptibility prediction, after examining a wide range of supervised learning methods. In the latter task, using the international HapMap project phase II and III dataset with hundreds of samples with different continental and subcontinental ancestral origins over 900K or 1450K features, we developed a novel learning method, ETHNOPRED, that achieved over 90% 10-fold cross validation accuracies in

various continental, and subcontinental population identification problems. Our sample complexity analysis (in the probably approximately correct learning framework) suggests that the ancestral origin prediction task is a case of realizable learning with many irrelevant features and so requires only a relatively small number of instances, while the breast cancer prediction task appears to be a case of unrealizable learning with relevant hidden features and hidden subclasses, explaining why it requires a large number of instances to be learned effectively, which we suspect is why the results here were not as good.

Acknowledgements

I could never travel this long path and finish my doctoral studies without the ongoing support of the exceptional people surrounding me.

I would like to acknowledge my joint-supervisors, *Dr. Russell Greiner* and *Dr. Sambasivarao Damaraju*, from departments of Computing Science and Laboratory Medicine and Pathology of University of Alberta who meticulously guided me to conduct a fruitful interdisciplinary research.

Dr. Greiner opened the gateway of a lot of fascinating opportunities for me in life by accepting me to be a Ph.D. student in his group, spending at least an hour weekly in face-to-face supervisory meetings with me, providing critical feedbacks that improved my writing and presentation skills, and funding me to travel to relevant conferences and workshop each year. I will never forget his morality, talent, and thoughtfulness.

Dr. Damaraju invited me to be a member of his research group, provided me the chance to do interdisciplinary research via the funding he introduced, offered me access to the genome wide profile of study subjects in my experimental studies, encouraged me to publish the findings of our research study, and conveyed the structure of this dissertation. I will never forget his dedication, passion, and kindness.

I would also like to acknowledge the outstanding members of my supervisory committee, *Dr. David Wishart* and *Dr. Dale Shuurmans*, my final oral examining committee, *Dr. Igor Jurisica* and *Dr. Paul Stothard*, and my candidacy examining

committee, *Dr. John Mackey* and *Dr. Osmar Zaiane*, for their constructive feedbacks that improved the technical quality of my work.

This research study would not be successful without the one of a kind Interdisciplinary Graduate Studentship award granted to me by Faculty of Science and Faculty of Medicine & Dentistry of University of Alberta.

I would like to thank my family for their enduring support and endless love. This dissertation is dedicated to my lifelong supporters and the best parents ever, *Keshvar* and *Yadollah*, and my warm-hearted parents-in-law, *Parisa* and *Davoud*. My sincere thank goes to my beloved wife and soulmate, *Shadi*, who accompanied me patiently in the freezing winters and the hot summers of Edmonton and filled my moments with joy. I take my hat off to my first teacher in reading and writing at the age of four, my big-hearted brother, *Abdollah*. I should emphasize that it would be impossible to pursue my studies without the encouragement and support of my kindhearted sister and brother, *Rana* and *Mehdi*.

Last but not the least, I should mention that I am grateful to be supported by a group of friends and colleagues who filled my moments with cheerfulness in Edmonton specifically *Dr. Hamid Moghaddas*.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation utilizes the supervised learning and the computational learning theory frameworks to provide explicit answers for the following research questions:

I. Is it feasible to use a labeled training dataset, with hundreds of samples and hundreds of thousands of features, in the form of single nucleotide polymorphisms (SNPs) captured from the genome-wide scan of germline DNA, to develop an accurate predictive model for an individual's susceptibility to breast cancer?

II. Is it feasible to use a labeled training dataset, with hundreds of samples and hundreds of thousands of features, in the form of single nucleotide polymorphisms (SNPs) captured from the genome-wide scan of germline DNA, to develop accurate predictive models for an individual's continental and subcontinental ancestral origins?

Chapter 1 is in an introduction to a number of key concepts frequently used in this dissertation. Chapters 2 and 3 answer the first and second research questions using the supervised learning techniques. Chapter 4 addresses these research questions using the computational learning theory techniques to gain further insights. Chapter 5 concludes with a review of the pros and cons of this research study and possible future directions.

While the interested readers who want to know more about different biomedical phenotypes, heritability, omics profiling, supervised learning framework, and computational learning theory framework might proceed with the rest of this chapter, the others can skip it.

1.1 Biomedical Phenotypes and Heritability

In general terms, a phenotype is defined as a feature of an organism that emerges as a result of interactions between its genetic and non-genetic (commonly identified as environmental) factors [1]. The combination of different phenotypes of an individual makes him unique. These phenotypes can be categorized to two large groups: 1) disease-associated phenotypes such as an individual's susceptibility to different diseases like breast cancer, and 2) non-disease associated phenotypes such as an individual's ancestral origins, height, and eye color. Different phenotypes differ from each other in the number of associated genotype and environmental factors and in the complexity of the function that combines these factors into a specific phenotype.

Heritability is defined as the degree to which individual genetic variation accounts for phenotypic variation seen in a population. Heritability index (h^2) is a metric to express the extent of heritability of a certain phenotype:

$$h^2 = \frac{Var(Genetics)}{Var(Genetics)+Var(Environment)} \tag{1}$$

Var(genetics) and Var(Environment) are the phenotypic variance caused by genetic and environmental factors respectively.

To elucidate the differences between different phenotypes, we review the differences between monogenic and multifactorial diseases.

1.1.1 Monogenic Diseases

Mendel's study of inheritance patterns in pea plants provided a solid foundation for our current understanding of single-gene diseases in humans [2]. Heritable disorders caused by a single defective gene such as Thalassaemia, Haemophilia, and Huntington's disease are known as monogenic diseases or Mendelian disorders. Most of these monogenic diseases are quite rare, affecting only one person in several thousands or millions [3]. Public access to the latest information on all known monogenic diseases is provided by the Genetic and Rare Diseases Information Center (GARD) of National Institute of Health (NIH) [4].

1.1.2 Multifactorial Diseases

Multifactorial diseases are disorders that do not obey the single-gene patterns of Mendelian disorders. Each multifactorial disease is complex as it involves a set of genetic/heritable predispositions within populations and non-genetic (environmental and lifestyle) components interacting over time [5]. Most common human disorders including cancer, cardiovascular diseases, stroke, chronic lower respiratory diseases, diabetes, Alzheimer's disease, Parkinson's disease, and Kidney diseases are multifactorial [5]. An intrinsic property of a multifactorial disease, which complicates disease modeling, is the heterogeneity of the disease – i.e., there are a number of different genetic and environmental alterations that produce diseases that are similar enough to have been traditionally grouped together under one diagnostic term [5].

1.2 Omics Profiling

From the earliest time that human beings started living on the earth, diseases that cause pain, dysfunction, distress, social problems, and death were also present. Many people suffered and died for millennia from thousands of illnesses. Although progress in the medical sciences in the recent centuries has shed light on some of these diseases and has provided high level definitions for them using an unaided eye, the ability to explore areas of micrometer (μm), nanometer (nm), and picometer (pm) size has enabled scientists to identify new cellular and molecular players in the disease environment. The field of human genetics has progressed substantially with the discovery of the structure of DNA in 1953 [6] and the completion of the Human Genome Project in 2003 [7]. Different fields of biomedical study ending in the *–omics* suffix, such as genomics, transcriptomics, epigenomics, proteomics, and metabolomics have emerged in the 21st century. The global awareness of the findings in these different omics disciplines has increased the hope that this rapidly growing science will help to prevent, diagnose, prognose, and treat life threatening diseases. References such as Genetics Home Reference [8] give the public access to the latest health related discoveries including the omics-related ones for patients, families, and health care providers.

1.2.1 Genome

The Genome (DNA) contains the genetic instructions for the development and functioning of all known living organisms, except RNA viruses. In human cells, DNA is stored in 23 pairs of chromosomes (22 autosomal pairs and 1 sex pair) inside the cell nucleus and also a small amount in the mitochondria. It is known that the DNA molecule comprises two complementary and anti-parallel helical chains, each coiled around the same axis [6]. The basic units of these strands are approximately 3 billion nucleotides bases of Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Each DNA segment that carries genetic information is called a gene, but other DNA segments have structural purposes or are involved in regulating the use of genes. The Human Genome Project found that the total number of human genes is around 20,000 to 25,000 and that less than 2% of the genome codes for genes [7]. To carry out the duties specified by the information encoded in genes, DNA segments must be transcribed into RNAs molecules many of which are later translated into proteins, which are the functional entities within the cells [9].

1.2.2 Point Mutations vs. Single Nucleotide Polymorphisms

Mutations include base substitutions, insertions, or deletions with the DNA. Mutations might happen because of exposure of DNA to radiation, viruses, transposons, and mutagenic chemicals, as well as errors that occur during meiosis or DNA replication [9]. Different mutations have different effects, depending on their location and type of nucleotide change [10]. The nucleotide at a position in which a mutation occurs is called an allele. The allele with the higher frequency of occurrence within a population is called the major allele (represented as “A” allele), while those occurring less frequently are called the minor alleles (represented as “B” allele). For each mutation, the two allelic variations (A and B) can give rise to three possible genotypes. When both parents contribute an “A” allele (same major allele), the genotype is referred to as wild type homozygous ("AA"); when both parents contribute a "B" allele (same minor allele) the genotype is referred to as variant (mutant) homozygous ("BB"), and when the two alleles are different, the genotype is referred to as heterozygous ("AB"). Single nucleotide changes in the DNA sequence with a minor allele frequency of less than 1% are called point mutations. Single nucleotide changes in the DNA sequence with a minor allele frequency of more than 1% at a specific human population are called SNPs. To date, millions of common SNPs have been identified and are accessible in public databases, such as dbSNP [11] or Ensembl [12].

1.2.3 Copy Number Variations and Structural Chromosome Variations

While most of the initial studies of genetic variation concentrated on individual nucleotide differences like point mutations and SNPs, large scale changes in the form of copy number variations (such as insertions, deletions, and amplifications) also occur in many locations throughout the genome. These CNVs contradict the common belief that two copies of a gene (one on each chromosome) are almost always present, one in each copy inherited from each parent. Furthermore, in some instances, CNVs change the physical arrangement of genes on chromosomes [13]. The Database of Genomic Variants (DGV) provides a comprehensive summary of structural variations found in the form of CNVs in the human genome [14]. In addition to CNVs, other changes such as structural abnormalities of chromosomes reflected in chromosomal rearrangements like inversion, intra- and inter-chromosomal translocations, are known to be accountable for certain types of diseases, such as Down syndrome, Turner syndrome, and certain cancers [15-17].

1.2.4 Germline vs. Somatic

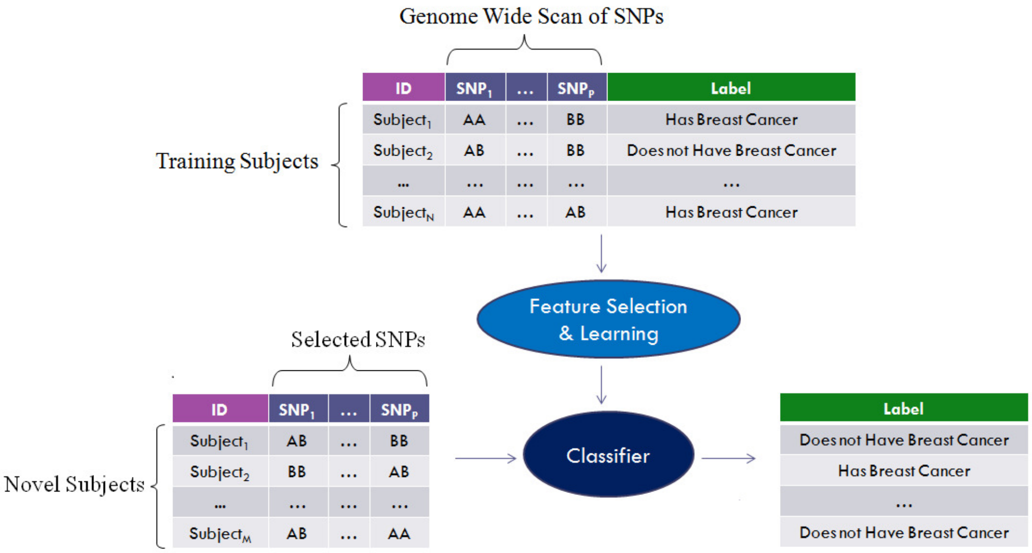
Blood cells, urine, and various body tissues are three common sources for extraction of omics profile of individuals. These omics signatures are either germline or somatic. A germline variation (usually extractable from blood cells) is inherited from an individual's parents and a somatic variation (usually extractable from various body tissues or cells shed in to body fluids such as urine and serum) is acquired during an individual's lifetime [18]. The somatic omics profiling in the form of epigenomics (which studies DNA modifications such as DNA methylations, and histone modifications), transcriptomics (which studies mRNA expressions and miRNA expressions), proteomics (which studies the proteome, the set of all proteins), and metabolomics (which studies the metabolome, the set of all metabolites) has a higher biological relevance for studying a disease since these signatures are proxy to the evolution of events culminating in a disease [18]. However, there are two challenges for applicability of somatic omics profiling in disease studies. The first challenge is the high variability of these signatures over time. The second challenge is the infeasibility of high-throughput measurement of some of these signatures such as proteomics and metabolomics due in part to limitations in today's technology [18]. As of today, the germline omics profiles in the form of genomics (which studies the genome variations such as mutations, SNPs, copy number variations (CNVs), structural genome variations) are attractive candidates for studying diseases due to signature stability and high-throughput measurement technology availability.

1.2.5 Microarrays vs. Next Generation Sequencing

Advancement of our knowledge in biology and of the relevant technologies has led to the fabrication of microarrays that assay large amounts of biological material in a high throughput screening method. There are different microarray platforms, which differ in fabrication, mechanism, accuracy, efficiency, and cost [19]. Although gene expression microarrays, which measure the mRNA levels of thousands of genes simultaneously, are the most well-known microarrays used, specific microarrays for detecting SNPs and CNVs also exist. While miniaturized gene expression microarrays have been used since 1995 [20], other microarrays such as SNP arrays and CNV arrays became applicable for conducting research in the recent years. Due to limitations of microarrays, they are being augmented by the next generation sequencers, which provide fast parallelized reads of thousands to millions of sequences at a reasonable price [21]. Whole genome sequencing provides the means of studying not only SNPs and CNVs but also mutations and structural genome variations. The raw data generated by either technology (microarrays or sequencing machines) needs to be preprocessed using appropriate quality control algorithms, to produce data amenable for downstream analysis.

1.3 Supervised Learning Framework

Different types of studies using different types of data analysis techniques are applicable to datasets produced by omics measurements. As the objective of this



ID	SNP_1	...	SNP_P	Label
$Subject_1$	AA	...	BB	Has Breast Cancer
$Subject_2$	AB	...	BB	Does not Have Breast Cancer
...	...	...	...	...
$Subject_N$	AA	...	AB	Has Breast Cancer

ID	SNP_1	...	SNP_P
$Subject_1$	AB	...	BB
$Subject_2$	BB	...	AB
...	...	...	...
$Subject_M$	AB	...	AA

Label
Does not Have Breast Cancer
Has Breast Cancer
...
Does not Have Breast Cancer

Figure 1.1: A schema of predictive studies of phenotypes using supervised learning framework.

Given a labeled training dataset of subjects each described by a genome wide scan of SNPs, feature selection and learning methods are applied to learn a classifier that can predict the labels of novel subjects.

dissertation is to produce new *predictive* models for phenotypes such as breast cancer susceptibility and ancestral origins, here we present the general framework of predictive studies. Furthermore, we compare predictive studies with association studies and risk assessment studies, which are more well-known in the biomedical community [22].

1.3.1 Predictive Studies

Predictive studies begin with a dataset of labeled subjects, each represented by a set of features (in our case, usually coming from high-throughput omics measurements), where each is “labeled” – that is, includes the phenotype of that subject. We can apply algorithms from machine learning (a field in the intersection of statistics and artificial intelligence, closely related to data mining and pattern recognition) to build a predictive model based on this dataset. This predictor may be later used to predict the class label of an unlabeled subject. There are a variety of statistical, probabilistic and optimization techniques that allow computers to learn classifiers from such datasets of labeled subjects [23-25]. These tools have been applied to produce effective predictors in many areas of biology and medicine [26-28]. The goal of the learning process is a performance system that uses a description of a novel subject, to predict some important characteristic of that person. The learning process starts with "labeled training data" – here, full or partial genomics measurements, for a set of subjects, each labeled with his/her phenotype. It then preprocesses the data and runs some predictive modeling system on this pre-processed data, to learn a pattern (a classification model) for that phenotype based on the assessed genotypes. Figure 1.1 represents the learning process steps. After producing such a prediction model, we can use this model to classify a new subject into one of the predefined labeled groups, based on a description of these subjects (selected parts of their genome).

Predictive studies are capable to answer a number of significant questions as follows if designed properly:

1. *Preventive question*: is an individual susceptible to a disease?
2. *Diagnostic question*: does an individual have a disease?
3. *Treatment question*: what is the best treatment for an individual diagnosed with a disease?

4. *Prognostic question*: how long will an individual survive from a disease in the absence or presence of a specific treatment?

1.3.2 Risk Assessment Studies

In the medical community, the term "predictive model" often refers to "risk assessment" or "risk prediction" [29-30]. Such risk assessment studies use a small set of pre-defined features (perhaps some subsets of clinical and/or omics features) to sort the subjects into a small set of bins, based on some combination of the values of these features - e.g., the Gail model used for estimating breast cancer risk uses seven features to produce a small number of bins - and then records the risk of each bin by simply calculating of the percentage of occurrence of each phenotype in each bin. Afterwards, to estimate the risk a new subject will face, the tool uses the subject's values for those relevant features to sort that subject into the proper bin, and returns the associated risk [31]. As this approach bases its assessment on only a small number of pre-specified features, it might not be sufficient to usefully characterize the subjects, especially if the hand-picked features are not adequate.

1.3.3 Association Studies

In the biomedical community, the most common type of study that uses high-throughput omics measurements is the association study. An association study takes as input a dataset of labeled subjects (cases and controls of a phenotype), each represented by a specific omics profile and attempts to identify the features that are mostly correlated with that phenotype. Some association studies seek differentially expressed genes in microarrays and others, called "genome wide association studies” (GWASs), use SNPs or CNVs. Different association studies use different metrics for ranking features, including t-test, ANOVA, and chi-square test [32-35]. Although association studies are applicable for identifying highly correlated features with a specific phenotype and potentially offer biological or mechanistic insights, they are not designed to produce predictive models of that phenotype.

1.4 Computational Learning Theory Framework

Computational learning theory is a subfield of machine learning that includes many theorems that explain the required computational and sample complexity of learning a classifier [36-37]. The focus of our research is on the sample complexity of learning, and this dissertation does not address the computational complexity of learning. The probably approximately correct learning (also known as PAC learning) concept [38] and Vapnik-Chervonenkis dimension (also known as VC dimension) concept [39] are the core foundations of the computational learning theory field. Here we first introduce these two concepts and then present a definition of the sample complexity upper-bound and lower-bound definitions in the PAC learning setting.

1.4.1 PAC Learning

Let a target concept (pattern) c be a member of the concept class C over an input space X_p (where X_p is $\{0, 1\}^p$ or p-dimensional Euclidean space R^p, etc). Supervised learning can be considered as searching in a chosen hypothesis class H for a hypothesis $h \in H$ that matches the training dataset. A concept class C is PAC

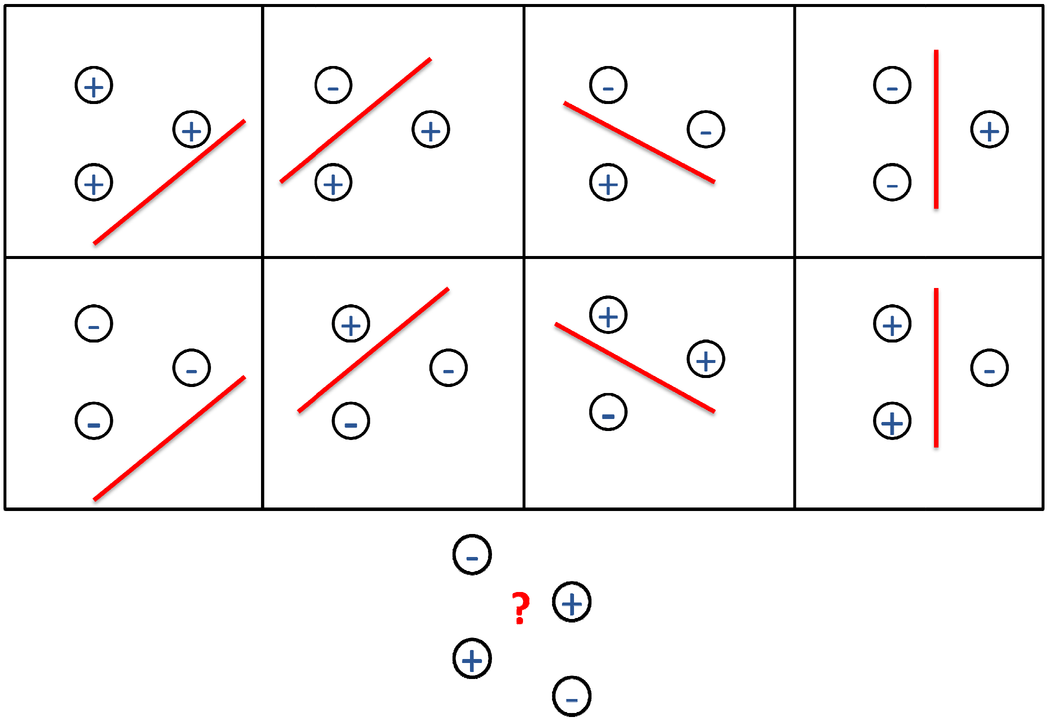


Figure 1.2: Representation of the VC dimension concept of linear classifiers in the 2-dimensional input space.

The VC dimension of $C_{linear, 2D}$ is 3 as, (1) for every possible labeling of 3 points, there is a linear classifier that correctly classifies these points, and (2) there is a set of 4 points that cannot be shattered using linear classifiers.

learnable using hypothesis class H if there exists an algorithm L with the following property: for every target concept $c \epsilon C$, for every distribution D on X_p, and for all values of the error parameters ε and the confidence parameters δ where $0 < \varepsilon < 0.5$ and where $0 < \delta < 0.5$, if L is given access to EX(c, D) (where EX(c, D) is a procedure that runs in unit time and on each call returns a labeled example $\langle x, c(x) \rangle$ where x is drawn randomly and independently according to D) and inputs ε and δ, then with probability at least $1\text{-}\delta$, L outputs a hypothesis $h \in H$, satisfying $error(h) \leq \varepsilon$ where $error(h) = Pr_{x \epsilon D}[c(x) \neq h(x)]$. This probability is taken over the random examples drawn by calls to EX(c, D), and any internal randomization of L. The hypothesis h ϵ H returned by the PAC learning algorithm is thus approximately correct with high probability, hence the name Probably Approximately Correct learning [40].

1.4.2 VC Dimension

A set of instances $\{x_1,\ldots,x_d\}$ is said to be shattered by C if for any labeling $(b_1,\ldots, b_d) \in \{\pm 1\}^d$, there is a concept c ϵ C such that $(c(x_1), \ldots, c(x_d)) = (b_1,\ldots,b_d)$. The VC dimension of a concept class C, which measures the complexity of that class, is the cardinality d of the largest set S shattered by C. If a learner sees the labels of only d-1 instances of the shattered set, the remaining instance would be unconstrained and could have either label. To better understand this concept, consider the concept class of linear classifiers over a two dimensional input space. As suggested by Figure 1.2, the VC dimension of this concept class is 3 since the perfect classification of 3 points using a linear classifier is always possible in the two dimensional space irrespective of their labeling, whereas there is a possible labeling for 4 points that a linear classifier cannot separate perfectly.

1.4.3 Sample Complexity Bounds in the PAC Learning Setting

Given the PAC learning setting, the computational learning theory literature considers two types of sample complexity bounds: the sample complexity *upper-bound* and the sample complexity *lower-bound* [40].

Sample complexity *upper-bound* for PAC learning a concept class C is the number training examples that are *sufficient* for finding a hypothesis h that is (ε, δ)-close to the target concept c in the hypothesis class H. PAC Learning would be feasible having a training dataset with greater than or equal training examples of the sample complexity upper-bound, in the worst case scenario. Here, the worst case refers to the worst possible choice of the target concept and the worst possible distribution of training examples. To represent the sample complexity upper-bounds, we usually use the O-notation. The definition of this notation is as follows:

f(n) = O(g(n)) if and only if: $\exists\ c, n_0 \in \Re^+$ such that $0 \leq f(n) \leq c \times g(n)$ for all $n \geq n_0$.

Sample complexity *lower-bound* for PAC learning a concept class C is the number training examples that are *necessary* for finding a hypothesis h that is (ε, δ)-close to the target concept c in the hypothesis class H. PAC learning would be infeasible having a training dataset with less training examples than the sample complexity lower-bound in the worst case scenario. To represent the sample complexity lower-bounds, we usually use the Ω-notation. The definition of this notation is as follows:

f(n) = Ω(g(n)) if and only if: $\exists\ c, n_0 \in \Re^+$ such that $0 \leq g(n) \leq c \times f(n)$ for all $n \geq n_0$.

To understand the sample complexity upper-bound and lower-bound concepts in the PAC learning setting, consider the example of buying a house in Edmonton with cash. Assume that we do not know how much the most expensive house in Edmonton is. However, we want to know how much money would buy any house for us in Edmonton. We can estimate the required amount of money for buying the most expensive house by finding lower-bound and upper-bound on this value.

If in reality, the most expensive house costs $1M, then $10K would be a very weak lower-bound and $10M would be a very weak upper-bound. Clearly, our estimate would be more precise, if these bounds are tighter. A lower-bound of $750K and an upper-bound of $2M are tighter bounds for the price of the most expensive house.

1.5 References

1. Lewontin R: **The Genotype/Phenotype Distinction.** In *The Stanford Encyclopedia of Philosophy, summer* 2011 edition. Edited by Zalta EN. 2011.
2. Mendel JG: **Versuche über Pflanzenhybriden Verhandlungen des naturforschenden Vereines in Brünn**, Bd. *IV für das Jahr* 1886 *, Abhandlungen*, 3-47.
3. Griffiths AJ, Miller JH, Suzuki DT, Lewotin R, Gelbart WM: *An Introduction to Genetic Analysis* (7 ed.). San Francisco, CA: Freeman, W. H. and Company 2008.
4. *The Genetic and Rare Diseases Information Center* (2004). Retrieved Sep 30, 2013, from http://rarediseases.info.nih.gov/gard.
5. Wright A, Hastie N: *Genes and Common Diseases.* New York, NY: Cambridge University Press 2007.
6. Watson JD, Crick FH: **A structure for Deoxyribose Nucleic Acid**. *Nature* 1953*,* **171**(4365): 373-378.
7. Collins FS, Morgan M, Patrinos A: **The human genome project: lessons from large-scale biology**. *Science* 2003*,* **300**(5617): 286-290.

8. Mitchell JA, Fun J, McCray AT: **Design of Genetics Home Reference: A new NLM consumer health resource**. *Journal of the American Medical Informatics Association* 2004*,* **11**(6): 439-447.

9. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD: *Molecular Biology of the Cell* (5 ed.). New York, NY: Garland Publishing 2007.

10. Bertram J: **The molecular biology of cancer**. *Molecular Aspects of Medicine* 2000*,* **21**(6): 167-223.

11. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al.: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Research* 2001*,* **29**(1): 308-311.

12. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, et al.: **Ensembl 2007**. *Nucleic Acids Research* 2007*,* **35**(Database Issue): D610-D617.

13. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al.: **Origins and functional impact of copy number variation in the human genome**. *Nature* 2009, **464**(7289): 704-712.

14. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al.: **Detection of large-scale variation in the human genome**. *Nature Genetics* 2004*,* **36**(9): 949-951.

15. Strachan T, Read AP: *Human Molecular Genetics*. New York, NY: John Wiley & Sons Inc 1999.

16. Gardner RJ, Sutherland G: *Chromosome Abnormalities and Genetic Counselling*, 3rd edition. New York, NY: Oxford University Press 2004.

17. Greaves MF, Wiemels J: **Origins of chromosome translocations in childhood leukaemia**. *Nature Reviews Cancer* 2003*,* **3**(9): 639-649.

18. Cho WC: *An Omics Perspective on Cancer Research.* New York, NY: Springer 2009.

19. *Microarray Factsheet* (2007). Retrieved Sep 27, 2013, from National Center for Biotechnology: http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html

20. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235): 467-470.

21. Shendure J: **The beginning of the end for microarrays?** *Nature Methods* 2008*,* **5**(7): 585-587.

22. Azuaje F: *Bioinformatics and Biomarker Discovery.* Singapore: Wiley-Blackwell 2010.

23. Mitchell T: *Machine Learning.* New York: McGraw Hill 1997.

24. Duda RO, Hart PE, Stork DG: *Pattern classification.* 2nd edition. New York: Wiley 2001.

25. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edition. New York: Springer 2009.

26. Baldi P, Brunak S: *Bioinformatics: The Machine Learning Approach, 2nd edition.* Cambridge, Massachusetts: The MIT Press 2001.

27. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles A: **Machine learning in bioinformatics.** *Briefings in Bioinformatics* 2006, 7(1)**:**86-112.

28. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S: **Machine learning and its applications to biology**. *PLoS Computational Biology* 2007, **3**(6):e116.

29. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, Van Duijn CM: **Predictive testing for complex diseases using multiple genes: fact or fiction?** *Genetics in Medicine* 2006*,* **8**(7): 395-400.

30. Kraft P, Hunter DJ: **Genetic risk prediction - are we there yet?** *New England Journal of Medicine* 2009*,* **360**(17): 1701-1703.

31. Decarli A, Calza S, Masala G, Specchia C, alli D, Gail MH: **Gail model for prediction of absolute risk of invasive breast cancer: independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition cohort**. *Journal of National Cancer Institute* 2006*,* **98**(23): 1686-1689.

32. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments**. *Statistica Sinica* 2002*,* **2**(1): 111-139.

33. Hardy J, Singleton A: **Genomewide association studies and human disease**. *New England Journal of Medicine* 2009*,* **360**(17): 1759-1768.

34. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al.: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007*,* **447**(7145): 661-678.

35. Craddock N, Hurles M, Cardin N, Pearson RD, Plagnol V, Robson S, et al.: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls**. *Nature* 2010*,* **464**(7289): 713-720.

36. Angluin D: **Computational learning theory: survey and selected bibliography**. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing* 1992, pp. 351-369.

37. Bousquet O, Boucheron S, Lugosi G: **Introduction to statistical learning theory**. In *Advanced Lectures on Machine Learning* 2004, pp. 169-207. Springer Berlin Heidelberg.

38. Valiant LG: **A theory of the learnable.** *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of* computing 1984.

39. Vapnik V, Chervonenkis A: **On the uniform convergence of relative frequencies of events to their probabilities**. *Theory of Probability and its Applications* 1971, **16**(2): 264-280.

40. Kearns M, Vazirani UV: *An introduction to computational learning theory*. The MIT Press 1994.

Chapter 2

Learning Breast Cancer Pattern Using Germline Genome Wide Scan of Single Nucleotide Polymorphisms

2.1 Summary

This chapter[1] introduces and applies a genome wide predictive study to learn a model that predicts whether a new subject will develop breast cancer or not, based on her SNP profile. We first genotyped 696 female subjects (348 breast cancer cases and 348 apparently healthy controls), predominantly of Caucasian origin from Alberta, Canada using Affymetrix Human SNP 6.0 arrays. Then, we applied EIGENSTRAT population stratification correction method to remove 73 subjects not belonging to the Caucasian population. Then, we filtered any SNP that had any missing calls, whose genotype frequency was deviated from Hardy-Weinberg equilibrium, or whose minor allele frequency was less than 5%. Finally, we applied a combination of MeanDiff feature selection method and KNN learning method to this filtered dataset to produce a breast cancer prediction model. LOOCV accuracy of this classifier is 59.55%. Random permutation tests show that this result is significantly better than the baseline accuracy of 51.52%. Sensitivity analysis shows that the classifier is fairly robust to the number of

[1] This chapter is prepared based on the following paper: Hajiloo M, Damavandi B, Hooshsadat M, Sangi F, Cass CE, Mackey JR, Greiner R, Damaraju S: **Using genome wide single nucleotide polymorphism data to learn a model for breast cancer prediction**, *BMC Bioinformatics* 2013, **14**(S13): S3.

MeanDiff-selected SNPs. External validation on the CGEMS breast cancer dataset, the only other publicly available breast cancer dataset, shows that this combination of MeanDiff and KNN leads to a LOOCV accuracy of 60.25%, which is significantly better than its baseline of 50.06%. We then considered a dozen different combinations of feature selection and learning method, but found that none of these combinations produces a better predictive model than our model. We also considered various biological feature selection methods like selecting SNPs reported in recent genome wide association studies to be associated with breast cancer, selecting SNPs in genes associated with KEGG cancer pathways, or selecting SNPs associated with breast cancer in the F-SNP database to produce predictive models, but again found that none of these models achieved accuracy better than baseline. We anticipate producing more accurate breast cancer prediction models by recruiting more study subjects, providing more accurate labeling of phenotypes (to accommodate the heterogeneity of breast cancer), measuring other genomic alterations such as point mutations and copy number variations, and incorporating non-genetic information about subjects such as environmental and lifestyle factors.

2.2 Background

Cancer is a complex disease, characterized by multiple molecular alterations triggered by genetic, environmental and lifestyle effects. Cancer cells typically accumulate alterations disrupting the cell's life cycle of growth, proliferation, and death [1]. Genomic changes that can eventually lead to cancer include mutations (<1% in frequency), single nucleotide polymorphisms (SNPs, >1% in frequency), insertion and deletion polymorphisms and structural changes in chromosomes. SNPs are the most common type of inherited genomic variation and recent advances in high-throughput technologies have led to whole-genome SNP arrays; datasets of such profiles over many subjects provide a valuable way to discover the relationship between SNPs and diseases such as cancer [2].

A genome wide association study (GWAS) compares the SNP profiles, over a wide range of SNPs, of two groups of participants: e.g., people with the disease (cases) versus people without the disease (controls). Each individual SNP whose values are significantly different between these groups (typically based on chi-square test between the values observed for the two groups) is said to be *associated* with the disease [3]. Of course, the resulting associated SNPs – even those with high statistical significance using genome-wide corrections for multiple hypothesis testing – are at best proxies for truly causal information, which can only be obtained through further deep sequencing of the associated loci and well-designed appropriate wet-lab studies. The database of Genotypes and Phenotypes (dbGaP) archives and distributes the results of studies that have investigated the interaction of a genotype and phenotype in GWASs [4]. However, while GWASs can help the researchers better understand diseases, genes and pathways, they are not designed to predict whether a currently undiagnosed subject is likely to develop the disease.

This chapter introduces Genome Wide Predictive Studies (GWPSs), which take the same input as a GWAS (the SNP arrays for a set of subjects, each labelled as a case or a control) but outputs a *classification model* that can be used later to predict the class label of a previously undiagnosed person, based on his/her SNP profile. The field of machine learning provides a variety of statistical, probabilistic and optimization techniques that allow computers to learn such classifiers from these datasets of labelled patients. Machine learning has been applied successfully in many areas of biology and medicine, often to produce effective predictors. Baldi and Brunak [5], Larranga et al. [6], Tarca et al. [7], Cruz and Wishart [8] each surveyed various applications of machine learning in biology, including gene finding [9], eukaryote promoter recognition [10], protein structure prediction [11], pattern recognition in microarrays [12], gene regulatory response prediction [13], protein/gene identification in text [14], and gene expression microarray based cancer diagnosis and prognosis [8]. We consider a way to learn a predictor ("who has breast cancer?"), for a dataset that specifies all available SNPs about each subject.

Our "genome wide" approach differs from research that attempts to learn predictors from only a pre-defined set of candidate SNPs. As an example of such a candidate SNP study, Listgarten et al. [15] applied a machine learning tool (support vector machine, SVM) to a pre-defined set of 98 SNPs, distributed over 45 genes of potential relevance to breast cancer, to develop a predictive model with 63% accuracy for predicting breast cancer. Ban et al. [16] applied a SVM to analyze 408 SNPs in 87 genes involved in type 2 diabetes (T2D) related pathways, and achieved 65% accuracy in T2D disease prediction. Wei et al. [17] studied type 1 diabetes (T1D) and reported 84% area under curve (AUC) using an SVM.

Our approach also differs from the conventional risk modeling/prediction studies. Those studies also begin with a small set of pre-defined features: they first sort the training subjects into a small set of bins, based on the values of these features – e.g., the Gail model uses 7 features – and record the percentage in each bin with the phenotype (here breast cancer) [18-19]. Afterwards, to estimate the risk a new subject will face, this tool uses the subject's values for those relevant features to sort that subject into the proper bin, and returns the associated probability (called risk). Hence this approach bases its assessment on only a small number of pre-specified features. Note this might not be sufficient to usefully characterize the subjects, especially if the hand-picked features are not adequate. On the other hand, our machine learning (ML) approach lets the data dictate on the possible combination of features that are relevant. (While the ML model described in this chapter returns a specific prediction for the individual – here breast cancer or not – there are other ML models that will return the probability that the individual will have the disease P(disease | feature_values), which is basically risk). Our general goal is to develop a tool to help screen women, by predicting which of the apparently healthy subjects sampled in a population will eventually develop breast cancer. This cannot be done by gene expression-based microarray analyses. Gene expression microarray analyses require biopsies of tissues from organs or tumours and are only relevant to individuals with suspect tissues. Therefore, they are not effective at identifying individuals at risk in a general population, before the onset of the disease, and so cannot be used for the purpose of early detection. The standard breast cancer risk assessment model (the Gail model [18-19], described

above) is designed to help with early detection; however, it has only limited clinical value. Note that researchers recently extended this Gail model by including 7 or 10 SNPs associated with breast cancer susceptibility (from GWASs); however, this led to only marginally improved accuracy [20-21].
This chapter presents a method to learn, from a dataset containing genome-wide SNPs of a cohort of subjects (cases and controls), a classifier that can predict whether a new subject is predisposed to the phenotype of breast cancer. (Note this classifier differs from the Gail model, as it can assign each individual subject to a label, potentially based on all of the features describing that subject.) We describe the challenges of addressing this high-dimensional data and show that a learner is capable of producing a classifier that can identify, with 59.55% accuracy, whether the subject has breast cancer, based only on her SNP profile. While this might not be clinically relevant, this performance is statistically significantly better than the baseline (of just predicting the majority class), which demonstrates that (1) there is information relevant to breast cancer in a patient's SNP values (note our method uses only SNPs, but not demographic data, nor other environmental data) and (2) that today's machine learning tools are capable of finding this important information.

2.3 Methods

In general, a Genome Wide Predictive Study (GWPS) takes as input the SNP profiles of a set of N individuals (including both cases and controls) and outputs a classifier, which can later be used to predict the class label of a new individual, based on his/her SNP profile; see Figure 1.1. Here, we used a dataset of N=696 subjects including 348 breast cancer cases (late onset of disease, i.e., of sporadic nature) and 348 controls (disease free at the time of recruitment and with no family history of breast cancer), accessed from a previous study on sporadic breast cancer wherein breast cancer predisposition in women is not related to mutations in the known high penetrance breast cancer genes (eg, BRCA) nor other genes of moderate penetrance, described in earlier studies [22]. Germline DNA was isolated from peripheral blood lymphocytes. Genotyping profiles were generated using Affymetrix Human SNP 6.0 array platform (906,600 SNPs on each array). The study subjects provided informed consent and the study was approved by the Alberta Cancer Research Ethics Committee of the Alberta Health Services. Following probe labelling, hybridization and scanning, population stratification correction using EIGENSTRAT [23] removed 73 subjects (46 cases and 27 controls) that did not co-cluster with Hapmap II Caucasian subjects, which left 623 Caucasian subjects (302 cases and 321 controls). After that, the dataset was filtered by removing any SNP (1) that had any missing calls, (2) whose genotype frequency deviated from Hardy-Weinberg equilibrium (nominal p-value <0.001 in controls) or (3) whose minor allele frequency were less than 5% (>5% frequency considered as common variants); this left a total number of 506,836 SNPs for analysis. For each SNP, we represented wild type homozygous, heterozygous and variant homozygous by 1, 2, and 3 respectively.

A trivial classifier, which just predicts the majority class (here control), will be 321/623 = 51.52% accurate. The challenge is producing a classifier that uses subject SNP data to produce predictions that are significantly more accurate. In particular, we explored tools that use the given labelled dataset to find the patterns that identify breast cancer (i.e., case versus control). Fortunately, the field of machine learning (ML) provides many such learning algorithms, each of which takes as input a labelled dataset, and returns a classifier. These systems typically work best when there are a relatively small number of features – typically dozens to hundreds– but they tend to work poorly in our situation, with over half-a-million features; here, they will often over-fit [24]: that is, do very well on the training data as they find ways to fit the details of this sample, but in a way that does not work well on the subjects that were not part of the training dataset. Note that our goal is to correctly classify such novel (that is, currently-undiagnosed) subjects. We therefore apply a pre-processing step to first reduce the dimensionality of the data, by autonomously identifying a subset of the most relevant SNPs (features). We then give this reduced dataset to a learning algorithm, which produces a classifier [25]. We later discuss how to evaluate the classifier produced by this "feature selection + learning" system.

2.3.1 Feature Selection

In our analysis, as we expect only a subset of the SNPs to be relevant to our prediction task, we focused on ways to select such a small subset of the features. In general, this involves identifying the features that have the highest score based on some criteria (which we hope corresponds to being most relevant to the classification task). In this study, we used the MeanDiff feature selection method, which first sorts the SNPs based on their respective MeanDiff values, which is the absolute value of the difference between mean values of this SNP over the cases and the controls:

$$\mathrm{MeanDiff}(\mathrm{SNP_i}, \mathrm{D}) = |\mu(\mathrm{i}, \mathrm{C}) - \mu(\mathrm{i}, \mathrm{H})| \tag{1}$$

over the dataset D = C ∪ H where C is the set of subjects known to have cancer (each labelled as case) and H is the remaining healthy subjects (each labelled as control), and using Expr(i,j) as the value of the i'th SNP of subject j, $\mu(i, H) = \frac{1}{|H|}\sum_{j\in H} Expr(i, j)$ is the mean value of the i'th SNP over the subset H (the controls) and $\mu(i, C) = \frac{1}{|C|}\sum_{j\in C} Expr(i, j)$is the mean value of the i'th SNP over the subset C (the cases). Note this MeanDiff($\mathrm{SNP_i}$, D) score will be 0 when $\mathrm{SNP_i}$ is irrelevant and presumably larger for SNPs that are more relevant to our prediction task. Here, we decided to use the m=500 SNPs with the largest MeanDiff values.

2.3.2 Learning

To build a classifier, we use the very simple learning algorithm, K-Nearest Neighbors (KNN), which simply stores the (reduced) profiles for all of the training data [26]. To classify a new subject *p*, this classifier determines *p*'s k nearest neighbors, and then assigns p the majority vote. (So if k=5, and *p*'s 5 closest neighbors include 4 controls and 1 case, then this classifier assigns *p* as control). Of course, we need to define distances to determine the nearest neighbors. As we are representing each patient as a m-tuple of the relevant SNP values, we define the distance between two individuals *p* = [p_1, ..., p_m] and *q* = [q_1, ..., q_m] as the square of the Euclidean distance (aka L2 distance) as shown below.

$$\mathrm{d}(\mathrm{p},\mathrm{q}) = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{p}_\mathrm{i} - \mathrm{q}_\mathrm{i})^2 \qquad (2)$$

2.3.3 Learning Parameter Selection

Notice the KNN learning algorithm requires us to specify how many neighbors to consider – the k mentioned above. Which value should we use – i.e., should we use k=1 (i.e., consider only the single nearest neighbor), or k=3 or k=5 or...? It is tempting to set k by: running 1-NN on the data, then determining the apparent error (using leave-one-out cross validation – see below), then computing the error associated with 3-NN, then 5-NN, and so forth; and finally selecting the value k ∊ {1, 3, 5, 7} that produces the smallest error. Unfortunately, this would mean finding a relevant parameter based on its score on the full set of training data, which corresponds to testing on the training data. That is, the k-value that optimizes that score might not be the one that produces the best performance on novel subjects, as the value determined in this fashion can lead to serious over-fitting.

We therefore need a more elaborate method, BestKNN, to determine the appropriate values for this parameter. Here, BestKNN first divides the training data into r=10 disjoint subsets, $D = D_1 \cup\ldots \cup D_r$, then for each i=1..r, defines $D_{-i}=D - D_i$ as the complement of D_i, and lets C_{i1} be the 1-NN classifier that is trained on D_{-i}. For each i, the C_{i1} classifier uses the m SNPs that have the best MeanDiff(., D_{-i}) scores, based on the D_{-i} dataset. As D_{-i} is different from D_{-j} when i≠j, the m SNPs used by C_{i1} will typically be different from the m SNPs used for $C_{j1.}$ BestKNN then computes the accuracy, acc(C_{i1}, D_i), of this C_{i1} classifier over D_i – ie, over data that it was not trained on. It then computes the average accuracy over all r different folds, $score\ (1, D) = \frac{1}{r}\sum_{i=1}^{r} acc(C_{i1}, D_i)$ which is an estimate of how well 1-NN would work over the complete dataset D. BestKNN similarly computes score (3,D) based on 3-NN, and score(5,D), etc., for k∈{1, 3, 5, 7}, then uses the high-watermark as the appropriate value of k. Here, using r=10 folds, it found k^* =7 worked best for our dataset (note this requires computing the top m SNPs, then running the resulting KNN, for 4×10 different datasets; the only purpose of all of this work is to find this k^* value). BestKNN then defines the final classifier based on the top m SNPs over the entire dataset, using this specific k^* =7 value.

2.3.4 Evaluation

The next challenge is estimating the quality of the classifier, C_{623} = BestKNN(D_{623}) – the classifier produced by running BestKNN (which involves the m best MeanDiff SNPs), on our 623 subject cohort D_{623}. Here we use two strategies to evaluate our classification algorithm: (1) by using Leave-One-Out Cross Validation (LOOCV) strategy and (2) by using an external hold-out (validation) dataset.

First, we use the LOOCV strategy, which first runs the BestKNN algorithm to produce a classifier based on N-1=622 training subjects (of the dataset with N=|D|=623 subjects), which is then tested on the 1 remaining subject. We ran these processes N times, so that every subject is used one time as the test dataset. We estimate the true accuracy of C_{623} as the percentage of correctly classified subjects, over these 623 folds. Producing this estimate means running all of BestKNN 623 more times – which, recall, each involves computing the top m

Table 2.1: Confusion matrix for comparison of actual and predicted labels on 623 breast cancer study subjects. Accuracy = (TP+TN)/(TP+FP+TN+FN)=59.55%; Precision = TP/(TP+FP)= 50.40%; Recall/Sensitivity = TP/(TP+FN)=61.92%; Specificity = TN/(TN+FP)=57.32%.

		Predicted Label	
		Case	**Control**
Actual Label	**Case**	187 (TP)	115 (FP)
	Control	137 (FN)	184 (TN)

SNPs for 40+1 different configurations. Some earlier researchers mistakenly ran their feature-selection process over the entire dataset D, and then committed to these features for all folds of the cross-validation process. Unfortunately, this gives inaccurate (overly optimistic) estimates [27-29]. On our task, we found that this incorrect process suggests that the resulting classifier has an apparent accuracy of over 90% -- which is considerably above its true accuracy of around 60% (see below).

Second, we used an external validation dataset of 2287 subjects (1145 breast cancer cases and 1142 controls) from the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer project [30]. Genotyping profiles for these subjects were generated using Illumina HumanHap550 (I5) array platform (555,352 SNPs on the array).To date, this is the only publicly available dataset related to a genome wide association study of breast cancer, which is on Caucasian population set.

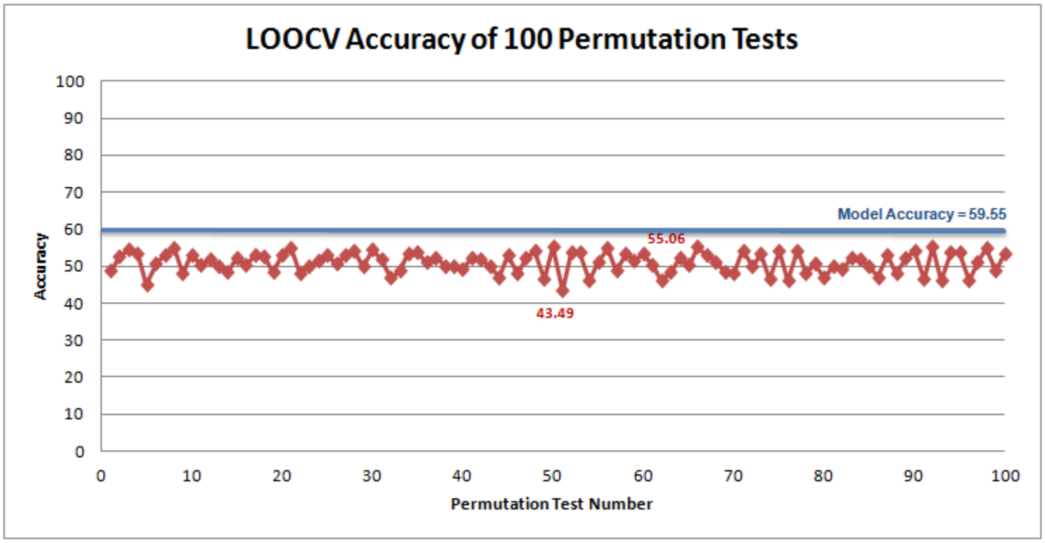


Figure 2.1: Accuracy of a hundred "Permute, Learn, and Evaluate" instances.

The accuracies of 100 random permutation tests. We see that none of these accuracies exceeded the 59.55% accuracy of our model. This means that our result is significantly better than the baseline, with a confidence of more than 99%.

2.4 Results

Table 2.1 provides the confusion matrix of actual versus predicted labels given by the classification model built using BestKNN, over the specified dataset. Our LOOCV estimates the accuracy of this model to be 59.55%; with precision 50.40%, recall/sensitivity 61.92%, and specificity 57.32%. To test if this result is significantly more accurate than the baseline of 51.52%, we applied a permutation test [31]. Here, we permuted the labels in the original dataset randomly, which should destroy any signal relating the SNPs to the cancer/no-cancer phenotype. We then ran the BestKNN to build new classifiers on this new dataset, and ran the LOOCV process to estimate the accuracy of the new model. We repeated this "permute, learn, evaluate" process over 100 permutations. As presented in Figure

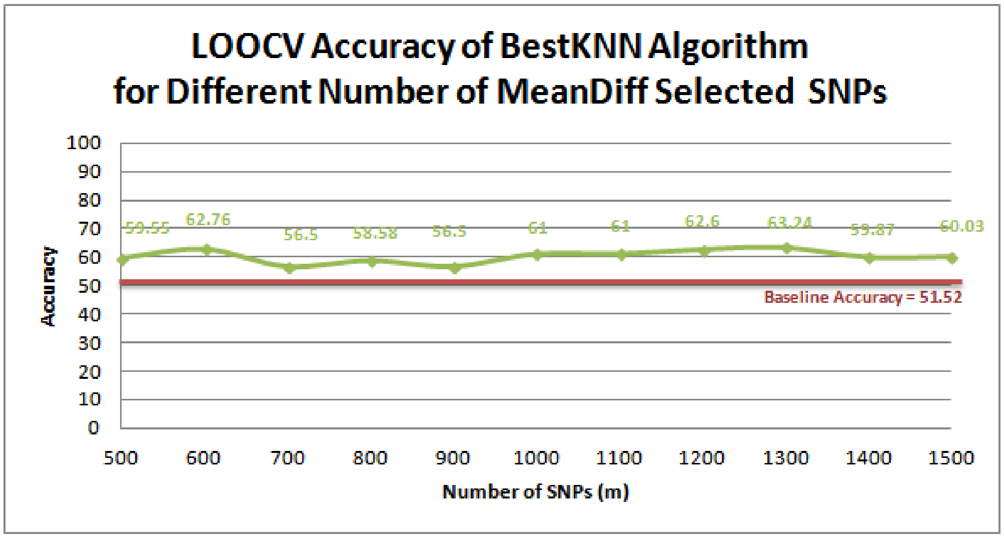


Figure 2.2: Accuracy of the BestKNN algorithm for different numbers of MeanDiff selected SNPs.

Accuracy of the classifiers built using BestKNN on sets of SNPs with the top {500, 600, ..., 1500} MeanDiff scores. This suggests that our model is fairly robust to the number of MeanDiff-selected SNPs, when selecting more than 500 SNPs.

2.1, none of these accuracies (of the 100 models built over randomly permuted labelled datasets) exceeded the 59.55% accuracy of our model. This suggests that our result is significantly better than the baseline, with a confidence of more than 1 – 1/100 = 0.99 – ie, the associated p-value is p<0.01. Figure 2.2, which provides the LOOCV accuracy of the classification model built using BestKNN on sets of SNPs with the top {500, 600, ..., 1500} MeanDiff scores, suggest our model is fairly robust to the number of MeanDiff selected SNPs, when selecting more than 500 SNPs.

To test the effectiveness of our approach, we next explored ways to apply it to other datasets. The standard approach involves running the resulting classifiers on another dataset, whose subjects include values for the same set of features and are

Table 2.2: Confusion matrix for comparison of actual and predicted labels on 2287 CGEMS breast cancer dataset. Accuracy = (TP+TN)/(TP+FP+TN+FN)=60.25%; Precision = TP/(TP+FP)= 60.44%; Recall/Sensitivity = TP/(TP+FN)=59.65%; Specificity = TN/(TN+FP)=60.86%.

		Predicted Label	
		Case	**Control**
Actual Label	**Case**	683 (TP)	462 (FP)
	Control	447 (FN)	695 (TN)

labeled with the same phenotypes. Unfortunately, there are no other public datasets for this phenotype that use the same Affymetrix Human SNP 6.0 array Platform. We did, however, consider applying our C_{623} = BestKNN(D_{623}) classifier on the CGEMS breast cancer dataset that includes 1145 breast cancer cases and 1142 controls genotyped on the Illumina I5 array platform. Unfortunately, due to this difference between the platforms, this dataset includes only 101 SNPs in common with the m=500 SNPs used by C_{623}. As this meant the CGEMS data was missing ~80% of the SNP values used by C_{623}, we obviously could not apply C_{623} directly on this dataset. As this CGEMS breast cancer dataset is the only available genome-wide dataset on Caucasian population, we therefore had to design another experiment to evaluate our approach based on the MeanDiff_{500}+BestKNN learning method. Here, we used the same MeanDiff_{500}+BestKNN algorithm, but trained this method over D_{2287}, the 2287 subjects of CGEMS breast cancer dataset. We again evaluated the performance of this learned model using the LOOCV method. Table 2.2 shows the estimated

Table 2.3: Accuracy of a dozen of different combinations of feature selection and learning methods. 10-fold cross validation accuracies of combination of 4 feature selection methods and 3 learning methods shows that none of these combinations are more accurate than our suggested combination of MeanDiff$_{500}$ feature selection and BestKNN learning (59.55%); indeed, several do not even beat the baseline of 51.52%.

		Feature Selection Methods			
		Information Gain	**MeanDiff**	**mRMR**	**PCA**
Learning Methods	**Decision Tree**	50.88%	52.06%	51.20%	51.69%
	KNN	56.17%	58.71%	57.78%	51.36%
	SVM-RBF	55.37%	57.30%	56.18%	51.84%

accuracy of this learning algorithm on this external validation dataset, BestKNN(D$_{2287}$), is 60.25% (which is significantly better than the baseline of 50.06%), with precision 60.44%, recall/sensitivity 59.65%, and specificity 60.86%. This confirms that our approach and algorithm is reproducible, as this exact system works effectively on a second, very different breast cancer dataset. Notice others have used the same validation approach [32].

Hoping to further improve these results, we explored several techniques – both biologically naïve and informed – for both selecting features and for building the classifier itself. To select features, we considered biologically naïve methods such as information gain [33], minimum redundancy maximum relevance (mRMR) [34] and principal component analysis (PCA) [35]. We also applied other biologically naïve learning algorithms, including decision trees [33], and support vector machines (with RBF kernel) [36]. In all, we tried a dozen of different combinations of the learning and feature selection algorithms (each with its own

Table 2.4: List of breast cancer associated SNPs reported by recent genome wide association studies. 28 SNPs identified by the 8 recent genome wide association studies on breast cancer. The accuracy of the classifier learned over these 28 genotyped SNPs was not better than the baseline of 51.52%.

dbSNP ID	Gene	Reference
rs2981579	FGFR2	Hunter et al., 2007 [30]
rs2420946	FGFR2	Hunter et al., 2007 [30]
rs11200014	FGFR2	Hunter et al., 2007 [30]
rs7696175	TLR1/TLR6	Hunter et al., 2007 [30]
rs17157903	RELN	Hunter et al., 2007 [30]
rs1219648	FGFR2	Hunter et al., 2007 [30]
rs3803662	TNRC9/LOC643714	Easton et al., 2007 [37]
rs889312	MAP3K1	Easton et al., 2007 [37]
rs13281615	8q	Easton et al., 2007 [37]
rs3817198	LSP1	Easton et al., 2007 [37]
rs2981582	FGFR2	Easton et al., 2007 [37]
rs2075555	COL1A1	Murabito et al., 2007 [38]
rs1978503	FLJ45743	Murabito et al., 2007 [38]
rs1926657	ABCC4	Murabito et al., 2007 [38]
rs13387042	2q35	Stacey et al., 2007 [39]
rs3012642	PHKA/HDAC8	Gold et al., 2008 [40]
rs7203563	A2BP1	Gold et al., 2008 [40]
rs6569479	ECHDC1/RNF146	Gold et al., 2008 [40]
rs2180341	ECHDC1/RNF146	Gold et al., 2008 [40]
rs6569480	ECHDC1/RNF146	Gold et al., 2008 [40]
rs4415084	5p12	Stacey et al., 2008 [41]
rs10941679	5p12	Stacey et al., 2008 [41]
rs2067980	MRPS30	Thomas et al., 2008 [42]
rs7716600	MRPS30	Thomas et al., 2008 [42]
rs11249433	1p11.2	Thomas et al., 2008 [42]
rs999737	RAD51L1	Thomas et al., 2008 [42]
rs4973768	SLC4A7	Ahmed et al., 2009 [43]
rs6504950	STXBP4	Ahmed et al., 2009 [43]

range of parameters values) – each of which proved to be computationally intensive (several CPU days). Table 2.3 shows the accuracy of each of these combinations. Here, we see that none of these combinations are more accurate than our suggested combination of MeanDiff_{500} feature selection and BestKNN learning (59.55%); indeed, several do not even beat the baseline of 51.52%. We also used biological information related to cancer to inform feature selection – i.e., use SNPs known to be relevant to breast cancer, rather than our biologically-naïve MeanDiff method: First, we considered the 28 SNPs identified by recent GWASs as being highly associated with breast cancer (see Table 2.4; [30, 37-43]). We trained KNN over the 623 subjects, but using only these 28 SNPs. Unfortunately the LOOCV of this classifier was just baseline, indicating that the SNPs that appear to be the most associated content with breast cancer are not sufficient to produce an effective classifier. Indeed, none of those 28 SNPs appear in the top 500 that MeanDiff selected. While different studies often identify different SNPs as significant, biological pathways seem much more stable, in that certain pathways are identified across multiple studies. This motivated us to try using only the 12,858 SNPs associated with genes of the KEGG's cancer pathways [44] recognized as hallmarks of cancer [1]; unfortunately, the classifier based on these features also did not perform better than baseline. Finally, we built a classifier using only the 1,661 SNPs associated with breast cancer in the F-SNP database [45]; this too had just baseline accuracy. These negative results show that the obvious approach of first using prior biological information to identify SNPs, and then learning a classifier using only those SNPs, does not seem to work here.

2.5 Discussion

Our studies, using MeanDiff within BestKNN, confirm that SNPs do carry information related to breast cancer genetic susceptibility, and that GWPSs are a promising tool for decoding and exploiting this information. While this approach is theoretically applicable for studying other cancer types and diseases, we list below some of the potential limitations that may make it difficult to produce more accurate prediction models, for breast cancer or other diseases:

Small sample size vs. large feature size: As noted earlier, as the number of subjects in this study is significantly less than the number of SNPs (a few hundred instances versus half a million features), we face high-dimensionality problem, which can cause the learning systems to over-fit – i.e., produce models that perform well on the training subjects but relatively poorly on new subjects distinct from those used for training. Two categories of techniques that attempt to tackle high-dimensionality are feature selection and sample integration. This report shows feature selection produces a classifier whose accuracy is significantly above baseline. Sample integration involves increasing the number of subjects in the study by either collecting more instances or by combining the dataset with other existing datasets, perhaps from different laboratories. However, there are still many significant challenges here, including dealing with batch effects [46].

Breast cancer heterogeneity: Breast cancer is biologically heterogeneous. current molecular classifications based on transcriptome-wide analysis, clinical determinations of steroid hormone receptor (like ER) status, human epidermal growth factor receptor 2 (HER2) status, or proliferation rate status (PR), all suggest a minimum of four distinct biological subtypes [47]. Our current dataset ignores the differences by merging these different sub-classes into the single label: case. We might be able to produce a more accurate predictor if we employed more detailed labelling of sub-cases, to produce a classifier that could map each subject to a molecular subtype. However, as our dataset is relatively small, further stratification of cases into subtypes of breast cancer might add to the high-dimensionality problem.

SNPs are only one form of genomic alterations: While this study considered only SNPs, there are also many other heritable genetic factors including mutations, copy number variations (CNVs), and other chromosomal changes. We expect that augmenting the SNP data with additional genetic information, such as insertion/deletion polymorphisms and CNVs, could lead to more accurate breast cancer predictive models. Of course, as this means using yet more features, this could also increase the risk of over-fitting.

Breast cancer is also influenced by non-genetic factors*:* Heritable factors are only part of the issue: while they play a major role in monogenic diseases such as haemophilia, diseases such as tuberculosis and lung cancer have a very high environmental and life style component, meaning genetic component contributes only a small amount to overall risk. Indeed, for many of diseases, the genetic component accounts for only 30-60% of the risk, with the remaining risk due to environmental and life style risk factors. There are many factors that contribute to developing breast cancer, in addition to heritable (DNA based) changes. The major environmental and lifestyle risk factors include age, estrogen exposure (from endogenous and exogenous sources), smoking, radiation exposure, obesity, and lifestyle in general [48]. As the breast cancer predictive model presented here used only germline DNA, it did not incorporate any of these non-genetic variables. We anticipate better results from a comprehensive model that includes both genetic and non-genetic factors.

2.6 Conclusions

We present a genome wide predictive study as a way to understand, and effectively use, data from multiple single nucleotide polymorphisms. We first contrast this approach with the more standard association studies, connecting this predictive approach directly with screening and personalized health care. We also show that it differs from the risk model (such as Gail) as our model can involve a large number of characteristics for each patient (here, hundreds of SNPs).

Our studies confirmed the feasibility of predicting breast cancer susceptibility from genome wide analysis of SNPs, by presenting a learning model that first uses the MeanDiff feature selection technique to identify the best subset of (m=500) SNPs from the over-500K SNPs of the original dataset, then used k-nearest neighbor (with the k learned using an appropriate algorithm) as the classifier over these SNPs. Leave-one-out cross validation estimates the prediction accuracy of this proposed method to be 59.55%. A random permutation test indicated that this result is significantly better than the baseline predictor ($p < 0.01$). Sensitivity analysis of the performance of our classifier showed that our model is robust to the number of MeanDiff-selected SNPs. We externally validated our learning algorithm using 2287 subjects from the CGEMS breast cancer dataset; this produced a classifier whose LOOCV accuracy was again significantly better than the baseline, which shows the reproducibility of our combination of MeanDiff and BestKNN in breast cancer prediction.

To better understand the challenge of this dataset, we systematically explored a large variety of other feature selection and learning algorithms. We found that none of the biologically naïve approaches to feature selection worked as well as our MeanDiff. We also considered many biologically-informed methods to select SNPs – using SNPs reported in the literature to be associated with breast cancer, SNPs associated with genes of KEGG's cancer pathways, and SNPs associated with breast cancer in the F-SNP database. However, those SNPs produced classifiers that were not even better than baseline. These negative findings suggest the challenge of our task, and of the importance of the findings of our study.

We also identified several limitations that may hinder a more accurate predictive model for breast cancer susceptibility. Sporadic breast cancer is a heterogeneous phenotype, which is also heavily influenced by environmental factors. Moreover, while our study does involve 623 samples, this is small relative to the number of features (SNPs) from a whole genome scan; we expect to achieve yet better results given larger sample sizes. Furthermore, we anticipate developing better predictive models by incorporating other information – both other genetic information (such as point mutations, copy number variations, and other structural chromosome changes using next generation sequencing) as well as environmental and lifestyle factors. The fact that our study produced statistically significant results, despite these limitations, demonstrates the potential of this machine learning approach in this context of screening, and of personalized patient care.

2.7 Authors' Contributions

MH designed and implemented the experiments and drafted the manuscript; BD, MHS, and FS helped running preliminary experiments; JRM provided insights from clinical oncology; CEC and SD as investigators on the Canadian Breast Cancer Foundation (CBCF) Tumor Bank in Alberta provided access to clinical data; RG participated in the design of experiments and manuscript edits; SD as the principal investigator of the whole genome breast cancer studies, offered data, provided suggestions during the course of experiments and edited the manuscript. All authors read and approved the final manuscript.

2.8 Acknowledgements

We thank Mr. Bret Hoehn, Ms. Nasimeh Asgarian, Dr. Badan Sehrawat, and Dr. Yadav Sapkota for discussions and help with the data preparation. We sincerely thank Dr. Paula Robson from the Tomorrow Project of the Alberta Health Services, Alberta for discussions and for providing data for our GWAS studies. This study is funded by the Faculty of Medicine & Dentistry/Faculty of Science Interdisciplinary Graduate Studentship Award, University of Alberta (to MH); Natural Sciences and Engineering Research Council and Alberta Innovates Centre for Machine Learning (to RG); Alberta Cancer Foundation, Canadian Breast Cancer Foundation- Prairies/NWT region, Alberta Innovates-Health Solutions funding support for the Alberta Cancer Research Biorepository/CBCF-Tumor Bank (to SD and CEC); Canadian Breast Cancer Foundation-

Prairies/NWT region for financial support through operating grants (to SD and JRM). We acknowledge dbGaP of NCBI for providing us access to the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer dataset.

2.9 References

1. Hanahan D, Weinberg RA: **The hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5)**:**646-674.

2. Buchanan JA, Scherer SW: **Contemplating effects of genomic structural variation**. *Genet Med 2008*, **10**:639-647.

3. Manolio TA: **Genomewide association studies and assessment of the risk of disease**. *N Engl J Med* 2010, **363**:166–76.

4. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: **The NCBI dbGaP database of genotypes and phenotypes.** *Nat Genet* 2007, **39**(10)**:**1181-1186.

5. Baldi P, Brunak S: *Bioinformatics: The Machine Learning Approach, 2nd Edition.* Cambridge, MA: The MIT Press; 2001.

6. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles A: **Machine learning in bioinformatics.** *Briefings in Bioinformatics* 2006, **7**(1)**:**86-112.

7. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S: **Machine learning and its applications to biology**. *PLoS Comput Biol* 2007, **3**(6):e116.

8. Cruz JA, Wishart DS: **Applications of machine learning in cancer prediction and prognosis.** *Cancer Informatics* 2006, **2:**59-78.

9. Mathé C, Sagot M-F, Schiex T, Rouzé P: **Current methods of gene prediction, their strengths and weaknesses.** *Nucleic Acids Res* 2002, **30:**4103-4117.

10. Won K, Prugel-Bennett A, Krogh A: **Training HMM structure with genetic algorithm for biological sequence analysis.** *Bioinformatics* 2004, **20**(18)**:**3613-3619.

11. Yi TM, Lander ES: **Protein secondary structure prediction using nearest-neighbor methods.** *J Mol Biology* 1993, **232:**1117-1129.

12. Pirooznia M, Yang JY, Yang MQ, Deng Y: **A comparative study of different machine learning methods on microarray gene expression data.** *BMC Genomics* 2008, **9**(Suppl 1):S13.

13. Middendorf M, Kundaje A, Wiggins C, Freund Y, Leslie C: **Predicting genetic regulatory response using classification.** *Bioinformatics* 2004, **20** (Suppl 1)**:**I232-I240.

14. Zhou GD, Shen D, Zhang J, Su J, Tan SH: **Recognition of protein/gene names from text using an ensemble of classifiers.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S7.

15. Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, Mackey J, Wishart D, Greiner R, Zanke B: **Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms.** *Clinical Cancer Research* 2004, **10:**2725-2737.

16. Ban HJ, Heo JY, Oh KS, Park KJ: **Identification of type 2 diabetes-associated combination of SNPs using support vector machine**. *BMC Genet*ics 2010, **11**:26.

17. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SFA, Polychronakos C, Hakonarson H: **From disease association to risk assessment: an optimistic view from genome-wide association studies on type-1 diabetes.** *PLoS Genetics* 2009, **5**(10)**:**e1000678.

18. Bondy ML and Newman LA: **Assessing breast cancer risk: evolution of the Gail Model**, *J Natl Cancer Inst* 2006, **98**(17): 1172-1173.

19. Decarli A, Calza S, Masala G, Specchia C, Palli D, Gail MH: **Gail model for prediction of absolute risk of invasive breast cancer: independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition cohort**, *J Natl Cancer Inst* 2006, **98**(23): 1686-1689.

20. Mealiffe ME, Stokowski RP, Rhees BK, Prentice RL, Pettinger M, Hinds DA: **Assessment of Clinical Validity of a Breast Cancer Risk Model Combining Genetic and Clinical Information,** *J Natl Cancer Inst* 2010, **102**(21): 1618-1627.

21. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P, Rosner B, Berg CD, Brinton LA, Lissowska J, Sherman ME, Chlebowski R, Kooperberg C, Jackson RD, Buckman DW, Hui P, Pfeiffer R, Jacobs KB, Thomas GD, Hoover RN, Gail MH, Chanock

SJ, Hunter DJ: **Performance of common genetic variants in breast-cancer risk models,** *New England Journal of Medicine* 2010, **362**: 986-93.

22. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, Mackey J, Greiner R, Damaraju S: **Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility,** *Human Genetics* 2011, **130**(4):529-537.

23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nature Genetics* 2006, **38:**904-909.

24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH; **The WEKA Data Mining Software: An Update**; *SIGKDD Explorations 2009*, **11**(1):10-18.

25. Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.

26. Cover TM, Hart PE: **Nearest neighbor pattern classification.** *IEEE Trans Inform Theory* 1967, **IT-13**:21–27.

27. Boulesteix AL, Strobl C, Augustin T, Daumer M: **Evaluating microarray based classifiers: an overview.** *Cancer Informatics* 2008, **6:**77-97.

28. Van’t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse H: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(31): 530–536.

29. Lee S: **Mistakes in validating the accuracy of a prediction classifier in high-dimensional but small-sample microarray data.** *Stat Methods Med Res* 2008, **17:** 635-642.

30. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer,** *Nature Genetics* 2007, **39**(7):870-874.

31. Good P: *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. 3rd edition. New York: Springer Series in Statistics; 2005.

32. Ahsen ME, Singh NK, Boren T, Vidyasagar M, White MA: **A new feature selection algorithm for two-class classification problems and application to endometrial cancer**, In *Proceedings of the 51st IEEE Conference on Decision and Control: 10-13 December 2012; Maui, Hawaii, USA*.

33. Quinlan JR: **Induction of decision trees.** *Machine Learning* 1986, 1**:** 81-106.

34. Ding C, Peng H: **Minimum redundancy feature selection from microarray gene** expression **data.** *International Conference on Computational Systems Bioinformatics* 2003, 523-528.

35. Jollife IT: *Principal Component Analysis*, Springer-Verlag, New York 1986.

36. Vapnik V: *The Nature of Statistical Learning Theory,* Springer-Verlag, New York 1995.

37. Easton DF, Pharoah PDP, Dunning AM, Pooley K, Cox DR, Ballinger D, Thompson D, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, the Search collaborators2, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Rahman N, Stratton MR, Peto J, Fletcher O, Ponder BAJ: **A genome-wide association study identifies multiple novel breast cancer susceptibility loci,** *Nature* 2007, **447**(7148):1087-93.

38. Murabito JM, Rosenberg CL, Finger D, Kreger BE, Levy D, Splansky GL, Antman K, Hwang S-J: **A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham heart study,** *BMC Medical Genetics* 2007, **8**(Suppl 1):S6.

39. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson, SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le ML, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson L, Kristjansson K, Bergthorsson JT, Kostic J, et al.: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer**, *Nature Genetics* 2007, **39**:865-869.

40. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P: **Genome-wide association study provides**

evidence for a breast cancer risk locus at 6q22. 33, *Proc Natl Acad Sci* **2008, 105(11):**4340-4345**.**

41. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, J*onsson GF, Jako*bsdotti**r** M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, Henderson BE, Kolonel LN, Le ML, Millastre E, Andres R, Saez B, Lambea J, Godino J, Polo E, Tres A, Picelli S, Rantala J, Margolin S, Jonsson T, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, et al.: **Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer**, *Nature Genetics* 2008, **40**:703-706.

42. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, Brinton LA, Sigurdson A, Doody M, Bhatti P, Alexander BH, Buring J, Lee IM, Vatten LJ, Hveem K, Kumle M, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Chanock SJ, Hunter DJ: **A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1)**, *Nature Genetics* 2008, **41**:579–584.

43. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R, Eccles D, Evans DG, Fletcher O, Johnson N, dos Santos Silva I, Peto J, Stratton MR, Rahman N, Jacobs K, Prentice R, Anderson GL, Rajkovic A, Curb JD, Ziegler RG, Berg CD, Buys SS,

McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver WR, Bojesen S, Nordestgaard BG, Flyger H, Dork T, Schurmann P, Hillemanns P, Karstens JH, Bogdanova NV, Antonenkova NN, Zalutsky IV, Bermisheva M, Fedorova S, Khusnutdinova E, Kang D, Yoo KY, Noh DY, Ahn SH, Devilee P, van Asperen CJ, Tollenaar RA, Seynaeve C, Garcia-Closas M, Lissowska J, Brinton L, Peplonska B, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, Hopper JL, Southey MC, Smith L, Spurdle AB, Schmidt MK, Broeks A, van Hien RR, Cornelissen S, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Schmutzler RK, Burwinkel B, Bartram CR, Meindl A, Brauch H, Justenhoven C, Hamann U, Chang-Claude J, Hein R, Wang-Gohrke S, Lindblom A, Margolin S, Mannermaa A, Kosma VM, Kataja V, Olson JE, Wang X, Fredericksen Z, Giles GG, Severi G, Baglietto L, English DR, Hankinson SE, Cox DG, Kraft P, Vatten LJ, Hveem K, Kumle M et al: **Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2**, *Nature Genetics* 2009, **41**:585–590.

44. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res.*2000, **28:** 27-30.

45. Lee PH, Shatkay H: **F-SNP: computationally predicted functional SNPs for disease association studies.** *Nucleic Acids Res.* 2008, **36:** 820-824.

46. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods**. *Biostatistics* 2007, **8**:118-127.

47. Bertucci F, Birnbaum D: **Reasons for breast cancer heterogeneity**. *J Biol* 2008, **7**(2):6.

48. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases**. Nature 2009, **461:**747-753.

Chapter 3

Learning Ancestral Origins Pattern Using Germline Scan of Single Nucleotide Polymorphisms

3.1 Summary

This chapter[2] proposes a novel machine learning method, ETHNOPRED, which uses the genotype and ethnicity data from the HapMap project to learn ensembles of disjoint decision trees, capable of accurately predicting an individual's continental and sub-continental ancestry. Here we provide an alternative technique to address population stratification. Population stratification is a systematic difference in allele frequencies between subpopulations. This can lead to spurious association findings in the case–control genome wide association studies (GWASs) used to identify single nucleotide polymorphisms (SNPs) associated with disease-linked phenotypes. Methods such as self-declared ancestry, ancestry informative markers, genomic control, structured association, and principal component analysis are used to assess and correct population stratification but each has limitations. To predict an individual's continental ancestry, ETHNOPRED produced an ensemble of 3 decision trees involving a total of 10 SNPs, with 10-fold cross validation accuracy of 100% using HapMap

[2] This chapter is prepared based on the following paper: Hajiloo M, Sapkota Y, Mackey JR, Robson P, Greiner R, Damaraju S: **ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction**. *BMC bioinformatics* 2013, **14**(1): 61.

II dataset. We extended this model to involve 29 disjoint decision trees over 149 SNPs, and showed that this ensemble has an accuracy of ≥ 99.9%, even if some of those 149 SNP values were missing. On an independent dataset, predominantly of Caucasian origin, our continental classifier showed 96.8% accuracy and improved genomic control's λ from 1.22 to 1.11. We next used the HapMap III dataset to learn classifiers to distinguish European subpopulations (North-Western vs. Southern), East Asian subpopulations (Chinese vs. Japanese), African subpopulations (Eastern vs. Western), North American subpopulations (European vs. Chinese vs. African vs. Mexican vs. Indian), and Kenyan subpopulations (Luhya vs. Maasai). In these cases, ETHNOPRED produced ensembles of 3, 39, 21, 11, and 25 disjoint decision trees, respectively involving 31, 502, 526, 242 and 271 SNPs, with 10-fold cross validation accuracy of 86.5% ± 2.4%, 95.6% ± 3.9%, 95.6% ± 2.1%, 98.3% ± 2.0%, and 95.9% ± 1.5%. However, ETHNOPRED was unable to produce a classifier that can accurately distinguish Chinese in Beijing vs. Chinese in Denver. ETHNOPRED is a novel technique for producing classifiers that can identify an individual's continental and sub-continental heritage, based on a small number of SNPs. We show that its learned classifiers are simple, cost-efficient, accurate, transparent, flexible, fast, applicable to large scale GWASs, and robust to missing values.

3.2 Background

3.2.1 Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs), as single base substitutions in an individual's DNA, are the most common type of genetic variation in humans. SNPs are evolutionarily conserved and heritable. They give rise to one or more allelic variations at loci and may confer phenotypic variance. Polymorphisms result from the evolutionary processes, and are modified by natural selection. They are common in nature and are related to biodiversity, genetic variation, and adaptation [1]. To date, millions of human SNPs have been identified and recorded in public databases such as dbSNP [2] or Ensembl [3].

3.2.2 Genome wide association studies

A genome wide association study (GWAS) is an examination of a large set of common genetic variants, such as SNPs, over a set of "labeled" individuals, seeking variants that are associated with a phenotype, such as disease susceptibility, disease prognosis or drug response under the "Common Disease-Common Variant" hypothesis [4-5]. A GWAS normally compares the DNA of two groups of participants: subjects who expressed a phenotype (cases) versus subjects who did not (controls). Here, the researcher compares the values of each individual feature (e.g., a specific SNP) in the cases, with the corresponding values for this feature in the controls. If the range of values in these subgroups is significantly different, this feature is said to be associated with the phenotype. In contrast to *candidate* gene polymorphism studies, which test only a few pre-defined genetic regions, GWASs investigate the entire genome [6-7]. The database of genotypes and phenotypes (dbGaP) [8] and the catalogue of published

GWASs [9] archive and distribute the findings from GWASs to the broader scientific community.

3.2.3 Population stratification

Population stratification (aka population structure) is the presence of a systematic difference in allele frequencies between populations or subpopulations, possibly due to different ancestry. We observe population stratification because of the differences in social history, ancestral patterns of geographical migration, mating practices, reproductive expansions and bottlenecks of different human subpopulations [10].

3.2.4 Population stratification in GWASs

While conducting a GWAS, a major concern is the possibility of inducing false positive or false negative associations between a SNP and the phenotype due to population stratification. This has motivated many researchers to consider techniques to address population stratification problem. As a pre-processing step in GWAS, these techniques either exclude some of the study subjects to alleviate the problem or adjust some of the SNPs to correct for population structure [11]. Here we review some of the standard techniques used to deal with population stratification problem in GWASs and discuss their limitations:

3.2.4.1 Self-declared ancestry

Many studies ask subjects to identify their own ethnicity, by reporting their ancestry and country of origin. Then they address the problem of population stratification by including the cases and controls that have the same self-reported ancestry and by excluding other subjects from the GWAS. However this method is sometimes misleading as some people might not know their full lineage information, or simply be mistaken. Furthermore, self-declared ancestry is not always sufficient to control population stratification as nearly all populations are confounded by genetic admixture at some level [12].

3.2.4.2 Ancestry informative markers

Some projects attempt to estimate ancestry using a panel of ancestry informative markers (AIMs) that show the highest absolute value difference in allele frequency between two ancestral populations. A small set (typically tens to hundreds) of well-established AIMs can perfectly distinguish continental differences between individuals [13-16]; however, panels of AIMs, described thus far, are less informative in detecting sub-continental differences in closely related populations such as Europeans [17-25].

3.2.4.3 Genomic control

A widely used approach to evaluate whether a dataset is confounded due to population stratification involves computing the genomic control λ, which is defined as the median χ^2 (1 degree of freedom) association statistic across SNPs, divided by its theoretical median under the null distribution. A value of $\lambda \approx 1$ indicates no stratification, whereas $\lambda > 1$ indicates population stratification or other confounders [26-29]. Despite its widespread application, genomic control method has a fundamental limitation. In the real world, some markers differ in their allele frequencies across ancestral populations more than others while the genomic control corrects for stratification by adjusting association statistics at each marker by a uniform overall inflation factor. This uniform adjustment is not sufficient to deal with both markers that have strong differentiation across ancestral populations and also those with smaller differentiation.

3.2.4.4 Structured association

Structured association techniques are unsupervised learning (clustering) methods such as STRUCTURE [30] which is based on a Bayesian framework, and latent class analysis [31], which is based on maximum-likelihood that assign subjects of a case–control study cohort to discrete subpopulations based on their inter-cluster similarities and intra-cluster dissimilarities [32-33]. Although structured association methods have the advantage of assigning samples into meaningful population groups, they cannot be applied to GWAS datasets because of their intensive computational cost on large datasets provided by recent high-throughput measurements.

3.2.4.5 Principal component analysis

Techniques based on principal component analysis (PCA) [34-36], like EIGENSTRAT [34], are currently the state-of-the-art methods used in GWASs for population stratification correction. The EIGENSTRAT algorithm applies PCA to genotype data to infer continuous axes of genetic variations represented by principal component vectors and then adjusts genotypes and phenotype by amounts attributable to ancestry along each axis. Despite the widespread application of such PCA-based techniques, they have some disadvantages: First, they are not cost-efficient since they require genotyping thousands to millions of markers to be able to calculate principal component vectors. Second, to infer ancestry of subjects they apply PCA, a black-box model, which is not human readable (transparent). Third, as high-throughput measurements produce many missing values, the straightforward PCA does not apply, leading EIGENSTRAT to use missing value imputation. However, such imputation techniques are problematic in population genetics as they ignore inter-individual and inter-ethnic variations, meaning such imputed datasets can lead to spurious association findings [37]. Fourth, the genotyping errors (GEs) that arise in high-throughput SNP measurements are a major issue in association studies [38-44] and substantially affect the efficiency of PCA-based methods like EIGENSTRAT [45].

3.2.5 The purpose of our research study

In this chapter, we introduce a novel method, EHNOPRED, for producing models that can accurately place subjects within continental and sub-continental populations, by applying a supervised learning (classification) technique to datasets from the second and third phases of the international HapMap project [46]. The resulting classifiers can help correct population stratification in association studies, overcoming some of the limitations of the conventional methods listed above. First, self-declared ancestry information is problematic, except possibly for isolated populations with extensive inbreeding. ETHNOPRED does not rely on self-declared ancestry information and analyzes an individual's genome to properly identify his/her ancestry. Second, while small panels of AIMs for continental population identification are designed, panels of AIMs for sub-continental population identification, if designable, either are less informative or use a large set of markers. However, ETHNOPRED produces accurate classifiers not only for continental population detection but also for sub-continental population detection using a small number of markers. Third, ETHNOPRED is not relying on the assumption of the genomic control method that all markers contribute equally to population stratification and instead benefits from the fact that different markers contribute to population differences in different degrees. Fourth, unlike structured association methods, ETHNOPRED classifiers are fast and easily applicable to the large GWAS datasets generated by high-throughput measurement techniques like microarrays and next generation sequencers. Fifth, ETHNOPRED classifiers require genotyping of only tens to hundreds of SNPs for accurate population identification. Hence they are simpler and more cost-efficient in comparison to PCA-based methods, which require genotyping of thousands to millions of SNPs. Sixth, PCA-based methods like EIGENSTRAT are substantially affected by the genotyping errors arisen in high-throughput SNPs measurements [45]. However, low-throughput SNP measurements of tens to hundreds of SNPs required by ETHNOPRED classifiers may be easily validated on independent genotyping platforms to rule out genotyping errors and assess concordance of genotype calls across independent platforms. Once these criteria are established, these selected SNP panels could be used to identify population stratification across projects sharing similar cases and control cohorts in molecular epidemiological studies. Seventh, ETHNOPRED classifiers are a set of easy-to-read rules. Thus unlike PCA-based methods, the decision tree-based classifiers are transparent, and so can provide insight into the population classification problem they are dealing with. Eighth, unlike PCA-based methods, ETHNOPRED classifiers do not require any kind of imputation to handle missing values. ETHNOPRED classifiers are robust to missing values as their ensemble structure allows them the flexibility to deal with missing SNPs by simply removing some decision trees, and still remain able to accurately identify ancestry.

3.3 Methods

3.3.1 Datasets

Our objective is to build predictive tools to determine an individual's continental and sub-continental ancestry based on the values of a small set of his/her SNPs. We develop this tool by applying supervised learners to datasets from the second and third phases of the international HapMap project. The HapMap project is a multi-country effort to identify and catalogue genetic similarities and differences in human beings and to determine the common patterns of DNA sequence variations in the human genome. It is developing a map of these patterns across the genome by determining the genotypes of more than a million sequence variants, their frequencies and the degree of association between them, in DNA samples from subpopulations with ancestry from East and West Africa, East Asia, North and West Europe, and North America.

The HapMap phase II datasets, released in 2007, contained 270 subjects – including 90 Utah residents with ancestry from Northern and Western Europe (CEU), 90 Yorubans from Ibadan, Nigeria (YRI), and a mixture of 45 Japanese in Tokyo and 45 Han Chinese in Beijing (JPT/CHB) – each genotyped on an Affymetrix SNP array 6.0 platform, measuring 906600 SNPs. We utilize the

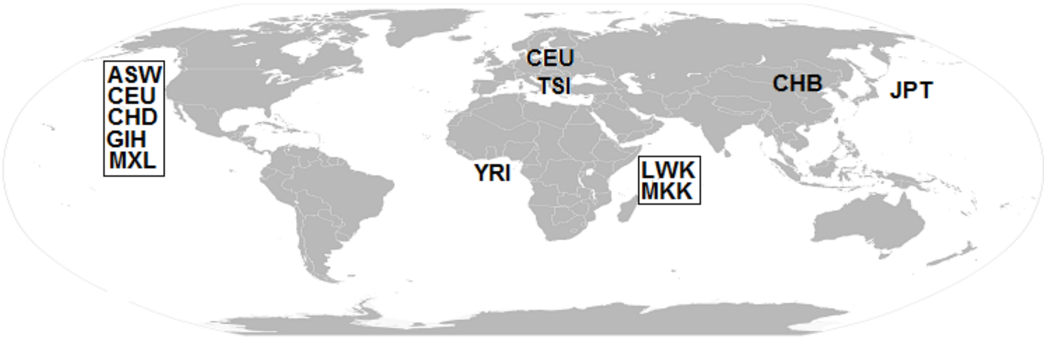


Figure 3.1: Geographic map of the HapMap phase III world populations.

ASW = Southwest USA residents with African ancestry; CEU = Utah residents with Northern and Western European ancestry; CHB = Han Chinese in Beijing, China; CHD = Chinese in Metropolitan Denver, Colorado; GIH = Gujarati Indians in Houston, Texas; JPT = Japanese in Tokyo, Japan; LWK = Luhya in Webuye, Kenya; MKK = Maasai in Kinyawa, Kenya; MXL = Mexicans in Los Angeles, California; TSI = Toscani in Italia; YRI = Yoruba in Ibadan, Nigeria.

HapMap II datasets to build a predictive model for inferring the continental ancestry origins (West Africa vs. East Asia vs. North-West Europe) of an individual. We apply the resulting classifier to a dataset of 696 breast cancer study subjects (348 breast cancer cases and 348 apparently healthy controls) from Alberta, Canada, genotyped on the same Affymetrix SNP array platform. We have self-declared ancestry of these 348 control individuals. These study subjects provided written informed consent and the study was approved by the Alberta Cancer Research Ethics Committee of the Alberta Health Services [47].

The HapMap phase III datasets, released in 2009, contained 1458387 SNPs of 1397 subjects including 87 Southwest USA residents with African ancestry (ASW), 165 Utah residents with ancestry from Northern and Western Europe (CEU), 137 Han Chinese in Beijing, China (CHB), 109 metropolitan Denver,

Table 3.1: Pre-processing statistics of for continental population classification problem based on HapMap Phase II samples.

SNP Groups	Number of SNPs
All SNPs	906600
SNP with Call Rate < 100%	186578
SNPs on Non-autosomal Chromosomes	38306
SNPs Deviated from HWE	184854
Filtered SNPs	295454
Unfiltered SNPs	611146

Colorado residents with Chinese ancestry (CHD), 101 Gujarati Indians in Houston, Texas (GIH), 113 Japanese in Tokyo, Japan (JPT), 110 individuals from Luhya tribe in Webuye, Kenya (LWK), 86 Los Angeles, California residents with Mexican ancestry (MXL), 184 individuals from Maasai tribe in Kinyawa, Kenya (MKK), 102 Toscani Italians (TSI), and 203 Yorubans in Ibadan, Nigeria (YRI). Figure 3.1 shows the geographic map of the HapMap III world populations. We utilize the HapMap III datasets to build predictive models for inferring sub-continental ancestry origins of Africans (LWK vs. MKK vs. YRI), Europeans (CEU vs. TSI), East Asians (CHB vs. JPT), North Americans (ASW vs. CEU vs. CHD vs. GIH vs. MXL), Kenyans (LWK vs. MKK), and Chinese (CHB vs. CHD).

3.3.2 Pre-processing

The allele with the dominant occurrence within a population is called the major allele (A), while the allele occurring less frequently is called the minor allele (B). Together, the alleles from paternal and maternal chromosomal loci can produce three distinct genotypes: When both alleles (*ie*, inherited from both parents) are

Table 3.2: Pre-processing statistics of HapMap phase III datasets and sub-continental population classification problems.

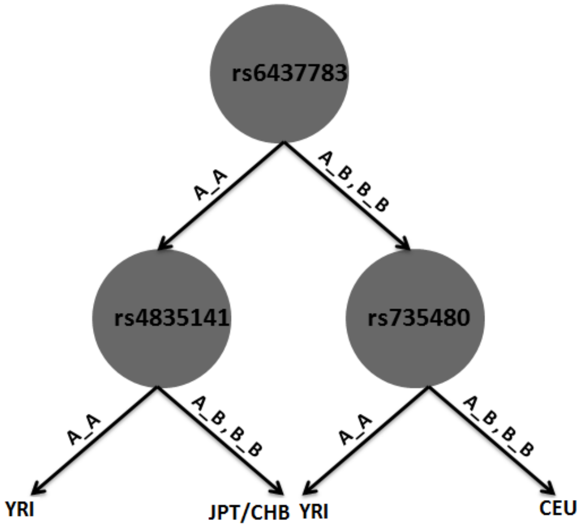
Dataset /Problem	Samples	All SNPs	Call Rate < 100%	Chr X, Y, MT, Unkown	Deviated from HWE	Filtered SNPs	Unfiltered SNPs
ASW	87	1458387	214898	34554	94234	298524	1159863
CEU	165	1458387	376531	34554	81633	427638	1030749
CHB	137	1458387	353208	34554	77028	423270	1035117
CHD	109	1458387	352031	34554	77111	421328	1037059
GIH	101	1458387	234863	34554	85463	314376	1144011
JPT	113	1458387	271105	34554	75502	337033	1121354
LWK	110	1458387	365638	34554	97174	425375	1033012
MKK	184	1458387	411395	34554	105490	471384	987003
MXL	86	1458387	311704	34554	86910	387207	1071180
TSI	102	1458387	268916	34554	81919	326585	1131802
YRI	203	1458387	423100	34554	94449	476513	981874
European	267	1458387	493449	34554	137488	575492	882895
East Asian	250	1458387	475217	34554	129695	565554	892833
African	497	1458387	742671	34554	228268	841790	616597
North American	548	1458387	803678	34554	306572	931993	526394
Kenyan	294	1458387	590202	34554	170547	677326	781061
Chinese	246	1458387	538224	34554	131394	629023	829364

the major alleles (A_A), the genotype is called wild type homozygous; when both the inherited alleles are minor (B_B), the genotype is called variant type homozygous; and when the two alleles are different (A_B), the genotype is called heterozygous.

To build our continental population classifier, we first identified the relevant SNPs from the HapMap II dataset, by removing a SNP if (a) it has a NoCall for *any* of the 270 subjects; (b) it is located on the X, Y, mitochondria (MT), or on an unknown chromosome; or (c) its genotype frequency deviates significantly from Hardy-Weinberg equilibrium (HWE) proportions, tested with Pearson's chi-squared (χ^2) test (nominal p-value < 0.05) [48]. We used criteria (a) to train our model using SNPs without missing values; (b) so the tool would be applicable to anyone, regardless of gender; and (c) by reasoning that observed genotype frequencies that deviate from HWE do not match the expected distributions of alleles, and hence are not reliable. These pre-processing steps removed a total of 295,454 SNPs, leaving 611,146 SNPs amenable for further scrutiny. Table 3.1 summarizes the statistics of the SNPs removed in the pre-processing steps, applied on HapMap II datasets. To build our sub-continental population classifiers, we followed similar filtering criteria on HapMap III dataset. These pre-processing steps respectively removed 841790, 565554, 575492, 931993, 677326, and 629023 SNPs, and left 616597, 892833, 882895, 526394, 781061, and 829364 SNPs amenable for further analysis in African population classification problem, East Asian population classification problem, European population classification problem, North American population classification problem, Kenyan population classification problem, and Chinese population classification problem. Table 3.2 summarizes the statistics of the SNPs removed in the pre-processing steps, applied on HapMap III datasets.

3.3.3 Predictive modeling

Machine learning provides a variety of statistical, probabilistic, and optimization techniques to analyze and interpret data, which allow computers to autonomously learn from past examples by finding patterns to form predictive models – often



Rule 1: IF rs6437783 ϵ {’A_A’} AND rs4835141 ϵ {’A_A’} THEN ethnicity is ‘YRI’
Rule 2: IF rs6437783 ϵ {’A_A’} AND rs4835141 ϵ {’A_B’,’B_B’} THEN ethnicity is ‘JPT/CHB’
Rule 3: IF rs6437783 ϵ {’A_B’,’B_B’} AND rs735480 ϵ {’A_A’} THEN ethnicity is ‘YRI’
Rule 4: IF rs6437783 ϵ {’A_B’,’B_B’} AND rs735480 ϵ {’A_B’,’B_B’} THEN ethnicity is ‘CEU’

Figure 3.2: The first decision tree and associated rule-set of the continental classifier produced by ETHNOPRED algorithm.

The decision tree uses 3 internal nodes (SNPs) acting as decision criterions and 4 leaf nodes (populations) demonstrating decisions. The number of rules in the relevant rule-base is equal to the number of leaf nodes of the decision tree.

finding hard-to-discern patterns, from noisy and complex datasets [49-51]. Machine learning has been applied successfully in many areas: Baldi and Brunak [52], Larranga et al. [53], and Tarca et al. [54] each surveyed various applications of machine learning in biology, medicine, and genetics including gene finding [55], eukaryote promoter recognition [56], protein structure prediction [57], pattern recognition in microarrays [58], gene regulatory response prediction [59], and protein/gene identification in text [60]. Herein, we learn a sequence of CART decision trees for continental and sub-continental population identification [61-62]. While machine learning provides many systems for

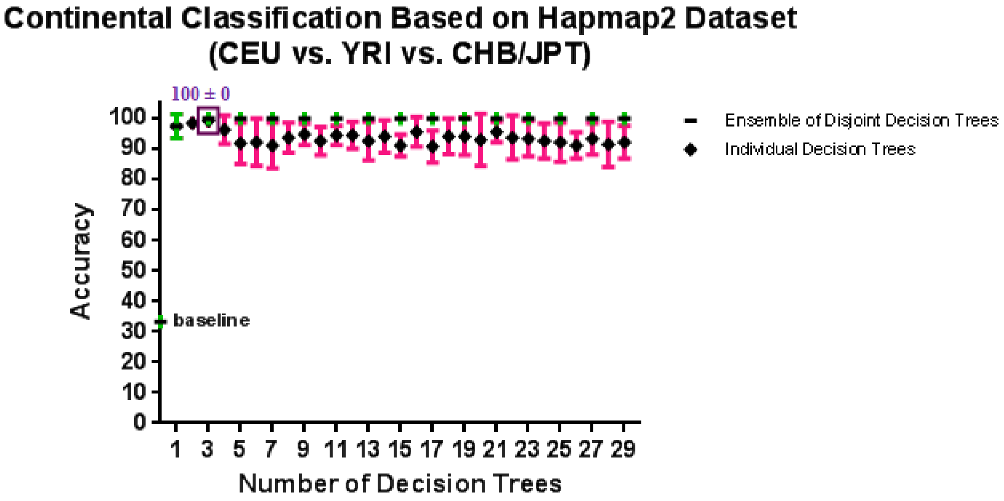


Figure 3.3: A comparison of 10-fold cross validation accuracy of individual decision trees and ensembles of disjoint decision trees of variable size in continental population classification problem using HapMap phase II datasets.

An ensemble of 3 disjoint decision trees involving 10 SNPs has a 10-fold cross validation accuracy of 100% which is significantly better than the baseline accuracy of 33.3%.

learning classifiers, we focus on decision trees as these learners are easy to use (as they do not require the user to provide any input parameters) and relatively fast to train, and the resulting classifiers run quickly and are easy to interpret (which may explain why they are widely applied in biological/medical domains).

“Ensemble learning” refers to a class of machine learning methods that combine the individual decisions of a set of learned “base predictors” to obtain a better predictive performance [63]. In general, an ensemble of predictors will be more accurate than any of its individual members if the constituent predictors are individually accurate and collectively diverse [64]. Ensemble models have been successfully applied on high-dimensional datasets generated by novel “omics” measurements, such as gene expression microarrays [65-66]. Many ensemble

Table 3.3: The average 10-fold cross validation accuracy of ensembles of size m, for m = 1...30 in continental population classification problem.

Number of Models in the Ensemble: m	Number of Ensembles: $\binom{30}{m}$	The Average 10-Fold Cross Validation Accuracy of Ensembles of size m: Acc
1	30	95.38
2	435	91.34
3	4060	98.36
4	27405	97.03
5	142506	99.32
6	593775	98.81
7	2035800	99.67
8	5852926	99.44
9	14300000	99.93
10	30000000	99.92
11	54600000	99.98
12	86500000	99.96
13	120000000	99.99
14	145000000	99.99
15	155000000	99.99
16	145000000	99.99
17	120000000	99.99
18	86500000	99.99
19	54600000	99.99
20	30000000	99.99
21	14300000	99.99
22	5852926	99.99
23	2035800	99.99
24	593775	99.99
25	142506	99.99
26	27405	99.99
27	4060	99.99
28	435	99.99
29	30	99.99
30	1	100

techniques – such as bagging, boosting, AdaBoost, and stacking – rely on manipulation of the input dataset by sampling of subjects or sampling of features,

Table 3.4: The confidence of having m = 9 decision trees without missing SNPs for N = 1...30 decision trees in continental population classification problem.

Number of Decision Trees: N	Confidence of Having m = 9 Decision Trees with No Missing SNPs: C
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0.873
10	4.09
11	10.676
12	20.566
13	32.716
14	45.652
15	58.013
16	68.86
17	77.744
18	84.616
19	89.682
20	93.265
21	95.71
22	97.328
23	98.369
24	99.023
25	99.424
26	99.666
27	99.809
28	99.892
29	99.94
30	99.967

then learning individual base classifiers on these subsets of the input dataset [67]. While the main goal of ensemble predictors is to produce an *accurate* classifier (as the ensemble can sometimes overcome the over-fitting problem reported for

Table 3.5: Comparison of self-declared lineage information and ETHNOPRED's result on 348 controls selected for a breast cancer susceptibility study of Caucasian women of Alberta, Canada.

	ETHNOPRED predicts as CEU	ETHNOPRED predicts as non-CEU
Self-declared lineage information as CEU	330	0
Self-declared lineage information as non-CEU	11	7

decision trees in high-dimensional problems [68]), we used this approach to produce a classifier that is *robust* to missing SNP values. Our system therefore learns a set of *disjoint* trees; we later explain how this allows the classifier to predict the label of a subject, even if that subject is missing many SNP values. Here we explain how ETHNOPRED learns an ensemble of disjoint decision trees, focusing on continental population classifier case. It first applies the CART learning algorithm to the dataset of 270 subjects over the 611146 SNPs mentioned above, to produce the decision tree (Figure 3.2) with 3 internal nodes (each a condition on a specific SNP) and 4 leaf nodes (class labels), corresponding to the 4 rules shown in Figure 3.2. It then removes these 3 SNPs from the list of 611146 SNPs and applies the same CART decision tree learning algorithm to the dataset of 270 subjects and the remaining 611143 SNPs, to produce a second decision tree. We repeat this algorithm, each time removing the SNPs used in the previous trees, to produce the next decision tree. The ETHNOPRED continental population classifier learns N = 29 disjoint decision trees. We explain below that N = 29 guarantees that this system is robust against missing SNP values – that is, based on some simple assumptions, we anticipate that at least 99.9% of the subjects will

Table 3.6: Comparison of self-declared lineage information and EIGENSTRAT's result on 348 controls selected for a breast cancer susceptibility study of Caucasian women of Alberta, Canada.

	EIGENSTRAT predicts as CEU	**EIGENSTRAT predicts as non-CEU**
Self-declared lineage information as CEU	321	9
Self-declared lineage information as non-CEU	0	18

include calls on the SNPs needed to "match" several decision trees; enough trees that the resulting sub-ensemble will be at least 99.9% accurate. This analysis appears below.

Figure 3.3 shows the estimated accuracies of the first k decision tree: the first tree, alone, is 97.41% and the ensemble classifier using the first 3 decision trees is 100%. If accuracy was our only concern, our ensemble classifier would just use these 3 decision trees, involving its 10 SNPs. However, this 3 decision tree system can only classify a subject if that subject includes values for (essentially) all 10 SNPs. Missing genotype data is a common problem in genotyping experiments, due to assay design failures, platform specific differences in the SNPs analyzed or due to hybridization artifacts in these high-throughput array platforms [69]. Here, we show that $N = 29$ decision trees are sufficient, under mild assumptions, to obtain an accuracy (Acc) of $\geq 99.9\%$ with 99.9% confidence (C), even considering missing SNPs: We trained 30 disjoint decision trees and found the average number of SNPs used in these 30 decision trees is $n = 154/30 \approx 5.13$. We then assumed that, for the Affymetrix genome wide SNP array 6.0 platform, NoCall's are independent from one SNP to another, and that the probability that a

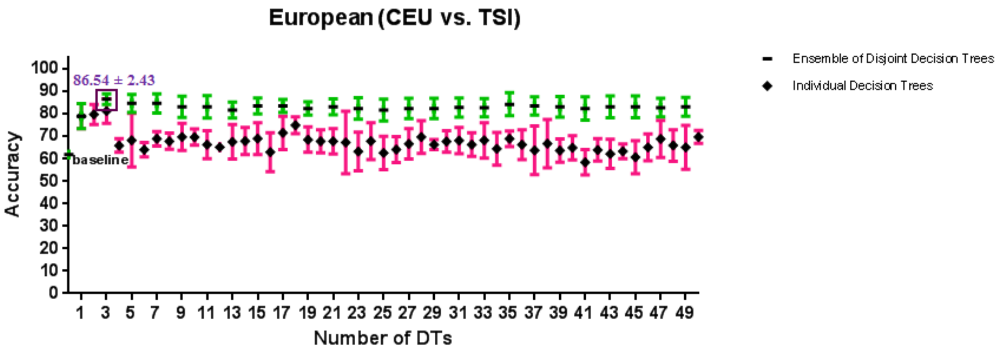


Figure 3.4: A comparison of 10-fold cross validation accuracy of individual decision trees and ensembles of disjoint decision trees of variable size in European population classification problem using HapMap phase III datasets.

An ensemble of 3 disjoint decision trees involving 31 SNPs has a 10-fold cross validation accuracy of 86.5% ± 2.4% which is significantly better than the baseline accuracy of 61.8%.

SNP value will be a NoCall is at worst $u = 0.1$ (based on assessment on the HapMap II dataset). This means that the probability that a subject will include all of the SNPs for a decision tree is $p \leq (1\text{-}u)^n = 0.9^{5.13} = 0.59049$, and so the probability that a subject will not include all of the SNPs of a decision tree is at least $q = 1 - p = 0.40951$. We now ask how many decision trees (m) are needed to insure that the average accuracy (Acc) of any subset of m trees is at least 99.9%. We therefore considered a sampling of ensembles of size 1 (i.e., individual decision trees) and calculated the average 10-fold cross validation accuracy. We next computed the average 10-fold cross validation accuracy over a sample of pairs of decision trees; then over triples, and so forth, for $i = 1..30$ (Table 3.3). We found that $m = 9$ is sufficient to obtain an average 10-fold cross validation accuracy (Acc) of 99.9%.

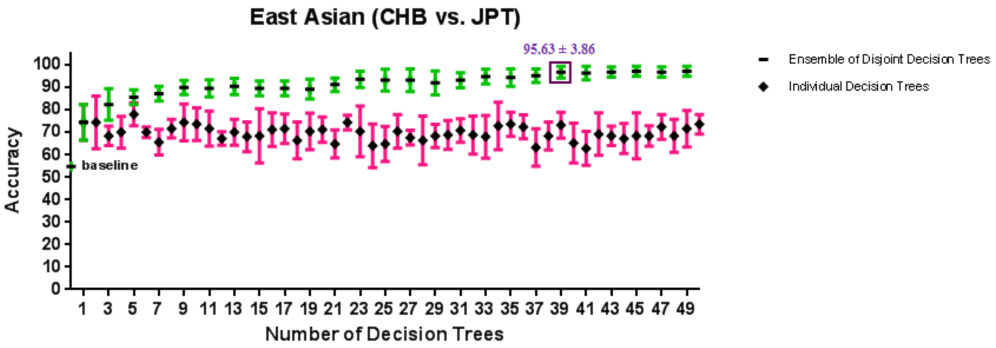


Figure 3.5: A comparison of 10-fold cross validation accuracy of individual decision trees and ensembles of disjoint decision trees of variable size in East Asian population classification problem using HapMap phase III datasets.

An ensemble of 39 disjoint decision trees involving 502 SNPs has a 10-fold cross validation accuracy of 95.6% ± 3.9% which is significantly better than the baseline accuracy of 54.8%.

The next challenge was in determining how many trees (N) are necessary, to be confident that the SNPs for 99.9% of all subjects will include calls on all of the SNPs for at least 9 trees. The probability of having at least m decision trees with no missing SNPs, given N decision trees, with probability p that a decision tree includes only specified SNPs, is:

$$C = \sum_{i=m}^{N}\left[\binom{N}{i} \times p^{i} \times (1-p)^{N-i}\right] \qquad (1)$$

Table 3.4 shows the values for C based on different values for N; here, we see N = 29 decision trees is sufficient to have 99.9% confidence (C) that a subject will include all of the SNPs in at least m = 9 decision trees, which our earlier experiments show is sufficient to produce an accuracy of ≥ 99.9%.

3.3.4 Models' usage for population stratification correction

For each continental and sub-continental ancestry identification problem, the pre-

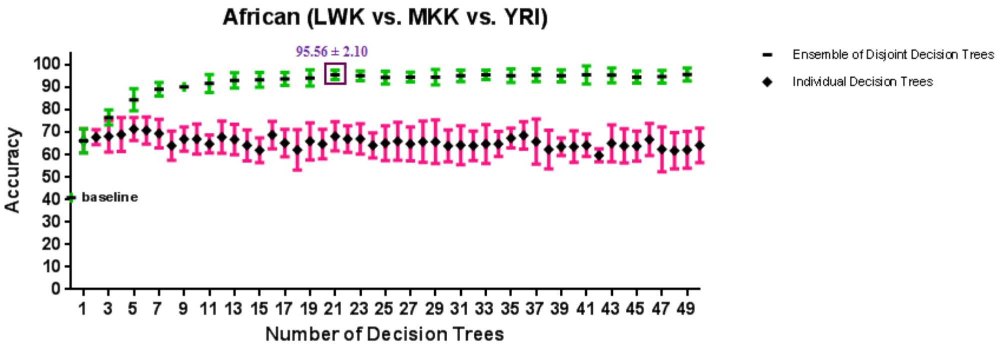


Figure 3.6: A comparison of 10-fold cross validation accuracy of individual decision trees and ensembles of disjoint decision trees of variable size in African population classification problem using HapMap phase III datasets.

An ensemble of 21 disjoint decision trees involving 526 SNPs has a 10-fold cross validation accuracy of 95.6% ± 2.1% which is significantly better than the baseline accuracy of 40.8%.

processing and predictive modeling steps produce a model (i.e., in the case of continental classification problem, the model is an ensemble of 29 decision trees) that can be used to classify novel subjects. For example, in continental population identification, we need to only find the values {A_A, A_B, B_B, NoCall} of the relevant 149 SNPs, then hand this set of 149 values to each of the 29 decision trees. Each tree involves a small number of SNPs (typically 3–7); if they are all specified (that is, none are "NoCall") for a novel subject, this tree will produce a predicted label – one of the three ethnicity groups: CEU, YRI, or CHB/JPT. If not, the tree makes no prediction. This will lead to a set of at most 29 predicted ethnicity values for this subject. As no human population is homogenous, given a novel subject with unknown ancestry, our model can provide a vector of population inclusion probabilities. For example, when classifying a novel person with the initial continental classification, imagine 15 trees vote for CEU, 4 for

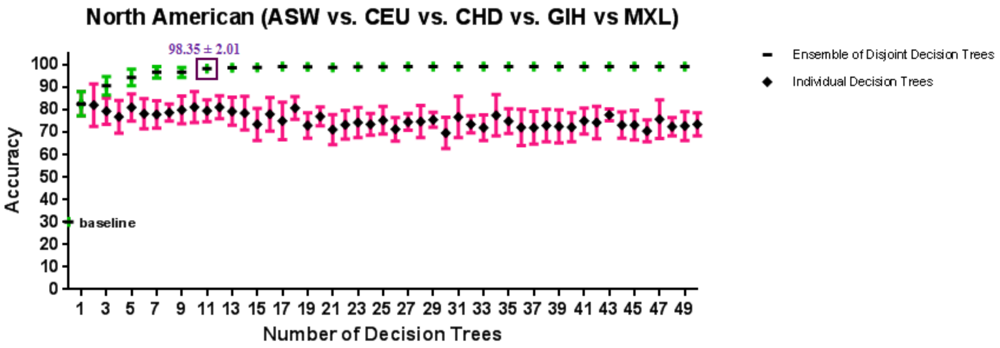


Figure 3.7: A comparison of 10-fold cross validation accuracy of individual decision trees and ensembles of disjoint decision trees of variable size in North American population classification problem using HapMap phase III datasets.

An ensemble of 11 disjoint decision trees involving 242 SNPs has a 10-fold cross validation accuracy of 98.3% ± 2.0% which is significantly better than the baseline accuracy of 30.1%.

YRI, 8 for JPT/CHB, and 2 are silent; this would produce the vector (15/27, 4/27, 8/27). These vector-valued predictions provide flexibility for researchers conducting a GWAS, as they can then, for example, define cut-off criterion for including a subject within a population under study. For each subject, continental classifier then returns, as ethnicity label, the ethnicity with the largest number of trees. In the *Results* section, we explain such panels for resolving the population stratification problem in closely related populations within a continent or a country as well.

3.3.5 Evaluation

We built the ETHNOPRED classifiers using HapMap II and HapMap III datasets as training data. Before using each classifier, we estimated its quality using a 10-fold cross validation (CV) [70]. This meant partitioning the training dataset into 10 disjoint folds. Each time we used nine of these folds (9/10th of data) as training

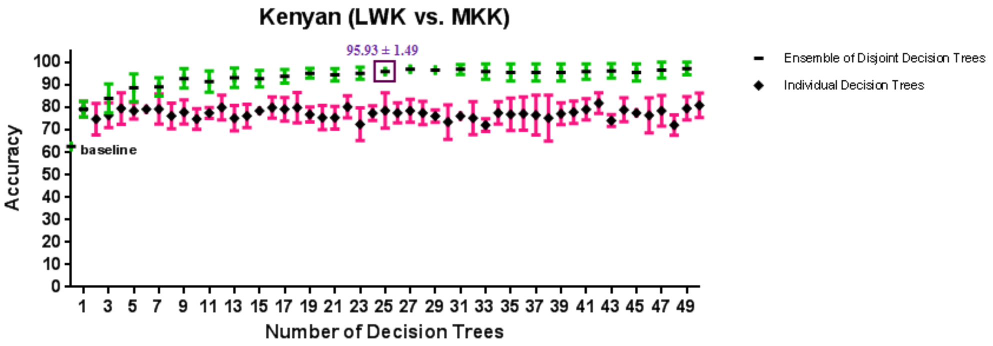


Figure 3.8: A comparison of 10-fold cross validation accuracy of individual decision trees and ensembles of disjoint decision trees of variable size in Kenyan population classification problem using HapMap phase III datasets.

An ensemble of 25 disjoint decision trees involving 271 SNPs has a 10-fold cross validation accuracy of 95.9% ± 1.5% which is significantly better than the baseline accuracy of 62.6%.

set for learning a sequence of decision trees, applying the algorithm explained in the *Predictive Modeling* section. We then used the remaining fold (1/10th of data) as a test set; here to compute, for each subject, class labels (one from each decision tree), and also the majority vote over these models (corresponding to the ensemble classifier). As we knew the true label for these subjects, we then obtained an accuracy score (the percentage of correct predictions over the total number of predictions) for each of the disjoint decision trees and for the final ensemble. We repeated this process 10 times, each time measuring accuracy of the predictors on a different fold. We estimated the final accuracy of the decision trees and ensemble model as an average of these 10 folds, with variance based on the spread of these 10 numbers. We used a similar way to evaluate the quality of the ETHNOPRED(k) classifier, where each such classifier was involved in returning the majority vote over subsequence of k individual decision trees.

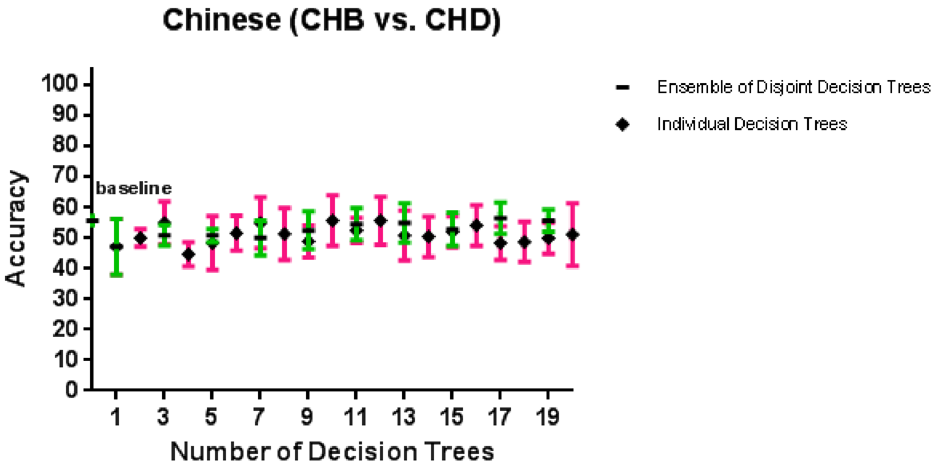


Figure 3.9: A comparison of 10-fold cross validation accuracy of individual decision trees and ensembles of disjoint decision trees of variable size in Chinese population classification problem using HapMap phase III datasets.

We considered several individual decision trees and ensembles of various sizes, but none had 10-fold cross validation accuracy better than the baseline accuracy of 55.7%.

3.4 Results and Discussion

3.4.1 Continental ancestry identification

Table 3.1 summarizes the statistics of the SNPs removed in the pre-processing step, which recall filtered out each SNP with a call rate of less than 100%, or that are located on X, Y, MT, or an unknown chromosome, or deviated from the HWE; this removed 295454 SNPs, leaving 611146 SNPs for further analyses. The final ensemble model, learned from all 270 subjects of the HapMap Phase II datasets, was composed of 29 disjoint decision trees, which each involved between 3 to 7 SNPs and between 4 to 8 leaf nodes/rules. This corresponds to a total of 178 rules involving 149 SNPs in the ensemble model. Figure 3.3 presents the 10-fold cross validation (CV) accuracy of the disjoint decision trees built based on the ETHNOPRED algorithm showing that the mean of the 10-fold CV accuracy of these models was between 90.7% and 99.3%. We see that the ensemble over only the first tree had a mean accuracy of 97.4%; the accuracy decreased (albeit insignificantly) to 95.9% by adding the second tree; the ensemble over 3 (or more) trees was 100% accurate. While adding additional trees to the ensemble did not improve the accuracy, our approach did increase its robustness to missing SNP values, as it means ETHNOPRED can produce a classification label even if the values of some of 149 SNPs for a subject were missing. Recall that ETHNOPRED can classify most subjects with missing SNP values as it can ignore any tree that includes missing SNPs, and returns as label the majority vote of the remaining trees.

To further assess the accuracy of ETHNOPRED, we also used a hold-out set of 696 breast cancer subjects (348 breast cancer cases and 348 controls) genotyped in Alberta, Canada. We had self-declared ethnicity labels for the control subjects. Here, we compared our ETHNOPRED against the commonly-used EIGENSTRAT system, in terms of the prediction accuracy and genomic control inflation factor (λ) improvement. Here, we extracted the values of ETHNOPRED's 149 SNPs for each subject. Note that 17 of these 149 SNPs had NoCalls for at least one subject. For each subject, each of ETHNOPRED's 29 decision trees predicted the subject's ethnicity to be one of "CEU", "YRI", "JPT/CHB", or "Missing". Continental classifier then calculates the covariate probability vector and returns the ethnicity with the majority vote as the predicted label for that subject. Prior knowledge of the subjects' ethnicity labels, when

available, would help assess the predictive accuracies of ETHNOPRED (or EIGENSTRAT) – eg, many previously published studies (including [45]) have used the HapMap subjects' self-declared ethnicity label to evaluate their ethnicity classifiers. We extrapolated this logic to calculate the prediction accuracies of ETHNOPRED over 348 control subjects, based on their self-declared ethnicity. Table 3.5 shows that ETHNOPRED's ethnicity classification matched closely with the subject's self-reported ethnicity (96.8%). Table 3.6 provides similar statistics for EIGENSTRAT (97.4%). The ETHNOPRED classifier labels 677 subjects as "CEU"; we could therefore use only these subjects and exclude the other 19 subjects for which either "YRI" or "CHB/JPT" is the majority ancestry covariate. Then we computed the inflation factor using the Genomic Control method for these subjects. For the entire sample size of 696 unclassified subjects in the association study, the computed inflation factor was 1.22, whereas the inflation factor computed for the 677 subjects classified as "CEU" by ETHNOPRED was 1.11, and the inflation factor for the 623 subjects classified as "CEU" by EIGENSTRAT was 1.10. While ETHNOPRED's learned classifier gives roughly the same improvement to the inflation factor as EIGENSTRAT, it offered the advantage of using a set of only
149 SNPs to achieve the classification of ethnicity label (CEU), which is significantly smaller than the 906,600 SNPs used by EIGENSTRAT.

3.4.2 Sub-continental ancestry identification

Table 3.2 summarizes the statistics of the SNPs filtered in the pre-processing step: those SNPs with a call rate of less than 100%, or located on X, Y, MT, or on an

Table 3.7: Summary of the sub-continental classification problems results.

This table summarizes the result of our studies on various sub-continental classification problems. The “Number of Subjects, Split” column shows the total number of subjects, followed by the list of (ethnic-group; number) pairs, giving the name of each subgroups and its size here. The “Number of SNPs” column gives the number of SNPs used for this study. The “Baseline” column gives the baseline accuracy of just using the majority class. The “DT1 (Number of SNPs), Accuracy” column provides the number of SNPs in the first decision tree, and its estimated 10-fold cross-validation accuracy. The “Minimal Number of DTs (Number of SNPs), Accuracy” column gives the minimal number of disjoint decision trees required to achieve the highest accuracy, and the number of SNPs involved, in these trees. The “Number of Robust DTs (Number of SNPs)” column gives the number of decision trees required to achieve robustness and the number of SNPs involved.

Problem	**# Subjects, Split**	**# SNPs**	**Base Line**	**DT1(# SNPs), Accuracy**	**# Accurate DTs (#SNPs), Accuracy**	**# Robust DTs (# SNPs)**
European	267, CEU: 165 TSI: 102	882895	61.8%	1 (10), 79.0% ± 5.6%	3 (31), 86.6% ± 2.4%	15 (180)
East Asian	250, CHB: 137 JPT: 113	892833	54.8%	1(12), 74.4% ± 7.9%	39 (502), 95.6% ± 3.9%	67 (877)
African	497, LWK:110 MKK: 184 YRI: 203	616597	40.8%	1(23), 66.2% ± 5.3%	21 (526), 95.6% ± 2.1%	157 (4236)
North American	548, ASW: 87 CEU: 165 CHD: 109 GIH: 101 MXL: 86	526394	30.1%	1(19), 82.7% ± 5.4%	11 (242), 98.4% ± 2.0%	70 (1643)
Kenyan	294, LWK: 110 MKK: 184	781061	62.6%	1(11), 79.2% ± 3.5%	25 (271), 95.9% ± 1.5%	31 (341)
Chinese	246, CHB: 137 CHD: 109	829364	55.7%	1(15), 47.2% ± 9.1%	- (-), ≤55.7%	- (-)

unknown chromosome, or deviated from the HWE; starting with 1458387 SNPs in the HapMap III dataset, this filtering removed 493449, 475217, 742671, 803678, 590202, and 538224 SNPs respectively in European, East Asian, African, North American, Kenyan, and Chinese population classification problems, and left 882895, 892833, 616597, 526394, 781061, and 829364 SNPs for further analyses. Table 3.7 summarizes the results of our study on these sub-continental population classification problems respectively for the case of European, East Asian, African, North American, Kenyan, and Chinese population classification problems. Figures 3.4, 3.5, 3.6, 3.7, 3.8, and 3.9 show the 10-fold CV accuracy of the individual disjoint decision trees and ensembles of varying size built over those trees using the ETHNOPRED algorithm. The baseline accuracy calculated by simply classifying every subject to the majority class in each of these sub-continental identification problems is as follows: 61.8%, 54.8%, 40.8%, 30.1%, 62.6%, and 55.7%. In each of these problems, the accuracy of a single decision tree, using 10, 12, 23, 19, 11, and 15 SNPs, is as follows: 79.0% ± 5.6%, 74.4% ± 7.9%, 66.2% ± 5.3%, 82.7% ± 5.4%, 79.2% ± 3.5%, and 47.2% ± 9.1%. These accuracies are significantly better than the baseline accuracy in every case except the Chinese one. Excluding the Chinese case, ensembles of 3, 39, 21, 11, and 25 decision trees using 31, 502, 526, 242, and 271 SNPs have accuracy equal to 86.6% ± 2.4%, 95.6% ± 3.9%, 95.6% ± 2.1%, 98.4% ± 2.0%, and 95.9% ± 1.5%, which are all statistically significantly better than the accuracy of the individual decision trees in other sub-continental classification problems. While adding additional trees to these ensembles does not improve the accuracy, using the arguments described in *Predictive Modelling* section, these additional trees do increase its robustness to missing SNP values; our analysis shows that an ensemble of 15, 67, 157, 70, and 31 decision trees using 180, 877, 4236, 1643, and 341 SNPs guarantees both accuracy and robustness to missing values in these cases. As mentioned above, ETHNOPRED is unable to produce a classifier that can distinguish between Chinese in Beijing and Chinese in Denver. This is not a limitation of our method given the fact that the first Chinese immigrant arrived in U.S. less than 200 years ago and for this reason the genetic differences between these two populations in terms of SNPs should not be considerable.

3.5 Conclusions

This chapter presents a new algorithm called ETHNOPRED that can learn classifiers (each an ensemble of disjoint decision trees) that can identify continental and sub-continental ancestry of a person. While this task is motivated by the challenge of addressing population stratification, it might be useful in-and-of itself, to help determine a person's ancestry. Applying this approach to downstream association tests/analysis may reduce the false positive and false negative findings by (i) removing the confounding subjects or alternatively, (ii) treating population classification probabilities as a covariate. Our results show that our machine learning approach is able to find distinctions between populations when there is a distinction. Unlike AIMS, our method can accurately distinguish genetically close populations such as Europeans, East Asians, Africans, North Americans, and Kenyans. Unlike many structured association methods, ETHNOPRED is fast and easily extendible to large scale GWASs. Furthermore, ETHNOPRED uses decision trees, which are much simpler and easier to understand than models based on principal component analysis, such as

EIGENSTRAT. Note also that decision trees can be easily translated into a set of comprehensible rules, which renders the model completely transparent to the user. While EIGENSTRAT typically uses data from genome wide scans, often involving hundreds of thousands of SNPs, ETHNOPRED uses a small number of SNPs to accurately determine the ancestry of subjects. This means our method is especially useful even in the absence of whole genome (high density) SNP data (*e.g.*, during Stage 2 or Stage 3 of a GWAS). Moreover, as it requires genotypes of only a small number of SNPs, it gets less affected by the genotyping errors compared with methods such as EIGENSTRAT, as there is typically a smaller percentage of genotyping errors when dealing with such small number of probes. ETHNOPRED's ensemble structure makes it robust to missing values, as its multiple trees include enough redundancies that it can return accurate predictions even if it discards some of decision tree while dealing with missing SNPs. We believe that this property of ETHNOPRED makes it beneficial over commonly-used methods that use imputation methods for missing values, as those techniques may introduce bias or imperfect estimations. These points all argue that future GWAS studies should consider using ETHNOPRED to estimate the ethnicity of their subjects, towards addressing possible population stratification. While our ETHNOPRED system is focused on predicting ethnicity, it is within the general machine learning framework, of using training information from a group of subjects to produce a personalized classifier that can provide useful information about subsequent subjects. This chapter shows that this framework can work effectively to solve important problems.

3.6 Authors' Contributions

MH designed the ETHNOPRED method, conducted the experiments and drafted the manuscript; YS prepared the breast cancer dataset, performed genomic control analyses, and offered manuscript edits; JRM offered interface to clinical oncology; PR provided control samples and lineage information for breast cancer study control samples; RG participated in the experimental design and provided manuscript edits; SD conceived the plan to devise ETHNOPRED, offered the breast cancer study data, and offered manuscript edits. All authors read and approved the final manuscript.

3.7 Acknowledgements

We acknowledge financial support through the interfaculty graduate studentship from the Faculty of Medicine and Dentistry and the Faculty of Science, University of Alberta (to MH); the Natural Sciences and Engineering Research Council and the Alberta Innovates for Machine Learning (to RG); and the operating grants to projects from the Alberta Cancer Foundation and Canadian Breast Cancer Foundation- Prairies/NWT Region (to SD). The apparently healthy individuals (controls) recruited in Alberta, Canada was made possible through a provincial program, "The Tomorrow Project" funded by the Alberta Cancer Prevention and Legacy Fund, managed by Alberta Innovates-Health Solutions. We acknowledge the anonymous reviewers for their constructive suggestions and valuable feedback that improved technical quality of this chapter substantially.

3.8 References

1. Jobling MA, Hurles ME, Tyler-Smith C: *Human Evolutionary Genetics: Origins, Peoples and Disease*. New York: Garland Science 2004.

2. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Research,* **29**(1): 308-311.

3. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Goates G, Cunnigham F, Cutts T, Down T, Dyer S C, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Parlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007**. *Nucleic Acids Research* 2007*,* **35**(Database Issue): D610-D617.

4. Lander ES and Schork NJ: **Genetic dissection of complex traits.** *Science* 1994, **265**:2037-2048.

5. Hirschhorn JN and Daly MJ: **Genome-wide association studies for common diseases and complex traits.** *Nature Review Genetics* 2005, **6**: 95-108.

6. Freedman M et al.: **Assessing the impact of population stratification on genetic association studies.** *Nature Genetics* 2004, **36**: 388-393.

7. Marchini J et al.: **The effects of human population structure on large genetic association studies.** *Nature Genetics* 2004, **36:** 512-517.

8. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: **The NCBI dbGaP database of genotypes and phenotypes.** *Nature Genetics* 2007, **39**(10)**:**1181-1186.

9. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proceedings of National Academy of Science* 2009, **106**(23):9362-7.

10. Cardon LR, Palmer LJ: **Population stratification and spurious allelic association**. Lancet 2003, 361:598-604.

11. Wu C, DeWan A, Hoh J, Wang Z: **A Comparison of Association Methods Correcting for Population Stratification in Case-Control Studies**, *Annals of Human Genetics* 2011, **75**(3): 418-427.

12. Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D: **Using ancestry-informative markers to define populations and detect population stratification.** *J Psychopharmacol* 2006, **20**(4 Suppl)**:**19-26.

13. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, *et al.*: **Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America.** *Hum Mutat* 2009, **30**(1)**:**69-78.

14. Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon- Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF: **An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels**. *BMC Genet* 2009, **10**:39.

15. Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, *et al.*: **Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs.** *Forensic Sci Int Genet* 2007, **1**(3–4)**:**273-280.

16. Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T: **A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications.** *Hum Mutat* 2008, **29**(5)**:**648-658.

17. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN: **Demonstrating stratification in a European American population**. *Nature genetics* 2005, **37**(8):868-872.

18. Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK: **European population substructure: clustering of northern and southern populations**. *PLoS genetics* 2006, **2**(9):e143.

19. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K: **An Icelandic example of the impact of population structure on association studies**. *Nature genetics* 2005, **37**(1):90-95.

20. Seldin MF, Price AL: **Application of ancestry informative markers to association studies in European Americans**. *PLoS genetics* 2008, **4**(1):e5.

21. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK et al: **Analysis and application of European genetic substructure using 300 K SNP information**. *PLoS genetics* 2008, **4**(1):e4.

22. Tian C, Kosoy R, Lee A, Ransom M, Belmont JW, Gregersen PK, Seldin MF: **Analysis of East Asia genetic substructure using genome-wide SNP arrays.** *PLoS ONE* 2008, **3**(12)**:**e3862.

23. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, Bustamante CD: **Genomewide patterns of population structure and admixture in West Africans and African Americans**. *PNAS* 2010, **107**:786-791.

24. Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, Pahl MV, Silva G, Belmont JW, Hanson RL, Knowler WC, *et al.*: **A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping.** *Am J Hum Genet* 2007, **80**(6)**:**1014-1023.

25. Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD: **Measuring European population stratification with microarray genotype data.** *Am J Hum Genet* 2007, **80**(5)**:**948-956

26. Devlin B and Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55:** 997-1004.

27. Reich D and Goldstein D: **Detecting association in a case-control study while allowing for population stratification.** *Genetic Epidemiology* 2001, **20:** 4-16.

28. Devlin B et al.: **Genomic control to the extreme,** *Nature Genetics* 2004, **36:**1129-1130.

29. Clayton DG et al.: **Population structure, differential bias and genomic control in a large-scale, case–control association study,** *Nature Genetics* 2005, **37**: 1243-1246.

30. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association mapping in structured populations**. *American Journal of Human Genetics* 2000,**67:** 170-181.

31. Satten G et al.: **Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model,** *American Journal of Human Genetics* 2001, **68**: 466-477.

32. Pritchard JK et al.: **Inference of population structure using multilocus genotype data,** *Genetics* 2000, **155**: 945-959.

33. Rosenberg NA et al.: **Genetic structure of human populations**. *Science* 2002, **298**: 2381-2385.

34. Price AL et al.: **Principal components analysis corrects for stratification in genome-wide association studies,** *Nature Genetics* 2006, **38**: 904-909.

35. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genetics* 2006, **2**: e190.

36. Novembre J and Stephens M: **Interpreting principal component analyses of spatial population genetic variation.** *Nature Genetics* 2008,**40:** 646-649.

37. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nat Rev Genet* 2008, **9:**356-369.

38. Ahn K, Gordon D, Finch SJ: **Increase of rejection rate in case-control studies with the differential genotyping error rates**. *Statistical applications in genetics and molecular biology* 2009, **8**(1):Article25.

39. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE et al: **Population structure, differential bias and genomic control in a large-scale, case-control association study**. *Nature genetics* 2005, **37**(11):1243-1246.

40. Kang SJ, Finch SJ, Haynes C, Gordon D: **Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies**. *Human heredity* 2004, **58**(3-4):139-144.

41. Londono D, Haynes C, De La Vega FM, Finch SJ, Gordon D: **A cost-effective statistical method to correct for differential genotype misclassification when performing case-control genetic association**. *Human heredity* 2010, **70**(2):102-108.

42. Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC: **Effects of differential genotyping error rate on the type I error probability of case-control studies**. *Human heredity* 2006, **61**(1):55-64.

43. Plagnol V, Cooper JD, Todd JA, Clayton DG: **A method to address differential bias in genotyping in large-scale association studies**. *PLoS genetics* 2007, **3**(5):e74.

44. Rice KM, Holmans P: **Allowing for genotyping error in analysis of unmatched case-control studies**. *Annals of human genetics* 2003, **67**(Pt 2):165-174.

45. Rakovski CS, Stram DO: **A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors**. *PloS one* 2009, **4**(6):e5825.

46. The International HapMap Consortium: **The International HapMap Project**. *Nature* 2003, **426**: 89-796.

47. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, Mackey J, Greiner R, Damaraju S: **Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility,** *Human Genetics* 2011, 130(4):529-37.

48. Pearson K: **Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs**. *Philosophical Transactions of the Royal Society of London* 1903*, Ser. A* **200** (321–330): 1–66.

49. Mitchell T: *Machine Learning*. New York: McGraw Hill 1997.

50. Duda RO, Hart PE, Stork DG: *Pattern classification.* 2nd edition. New York: Wiley 2001.

51. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edition. New York: Springer 2009.

52. Baldi P, Brunak S: *Bioinformatics: The Machine Learning Approach, 2nd edition.* Cambridge, Massachusetts: The MIT Press 2001.

53. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles A: **Machine learning in bioinformatics.** *Briefings in Bioinformatics* 2006, 7(1)**:**86-112.

54. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S: **Machine learning and its applications to biology**. *PLoS Computational Biology* 2007, **3**(6):e116.

55. Mathé C, Sagot M-F, Schiex T, Rouzé P: **Current methods of gene prediction, their strengths and weaknesses.** *Nucleic Acids Res* 2002, **30:**4103-4117.

56. Won K, Prugel-Bennett A, Krogh A: **Training HMM structure with genetic algorithm for biological sequence analysis.** *Bioinformatics* 2004, **20**(18)**:**3613-3619.

57. Yi TM, Lander ES: **Protein secondary structure prediction using nearest-neighbor methods.** *J Mol Biology* 1993, **232:**1117-1129.

58. Pirooznia M, Yang JY, Yang MQ, Deng Y: **A comparative study of different machine learning methods on microarray gene expression data.** *BMC Genomics* 2008, **9**(Suppl 1):S13.

59. Middendorf M, Kundaje A, Wiggins C, Freund Y, Leslie C: **Predicting genetic regulatory response using classification.** *Bioinformatics* 2004, **20 Suppl 1:**I232-I240.

60. Zhou GD, Shen D, Zhang J, Su J, Tan SH: **Recognition of protein/gene names from text using an ensemble of classifiers.** *BMC Bioinformatics* 2005, **6**(Suppl 1).

61. Quinlan JR: **Induction of decision trees**. *Machine Learning* 1986*,* **1***:*81-106.

62. Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and Regression Trees.* New York: Chapman &Hall (Wadsworth, Inc.) 1984.

63. Dietterich TG: **Ensemble methods in machine learning**. *Lecture Notes in Computer Science* 2000.**1857**:1-15.

64. Kuncheva LI, Whitaker CJ: **Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy**. *Journal of Machine Learning* 2003, **51**(2): 181-207.

65. Tan AC, Gilbert D: **Ensemble machine learning on gene expression data for cancer classification**, *Applied Bioinformatics* 2003. **2**:S75–S83.

66. Peng Y: **A novel ensemble machine learning for robust microarray data classification**, *Computers in Biology and Medicine* 2006, **36**(6): 553-573.

67. Polikar R: **Ensemble based systems in decision making**, *IEEE Circuits and Systems Magazine* 2006.**6**(3): 21-45.

68. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *Journal of the American Statistical Association* 2002, **97**(457): 77-87.

69. Lin DY, Hu Y, Huang BE: **Simple and Efficient Analysis of Disease Association with Missing Genotype Data,** *American journal of human genetics* 2008, **82**(2): 444-452.

70. Boulesteix AL, Strobl C, Augustin T, Daumer M: **Evaluating microarray-based classifiers: an overview**, Technical Report Number 005, Department of Statistics, University of Munich, 2007.

Chapter 4

Assessing the Feasibility of PAC Learning Phenotypes Using High-Throughput Omics Profiles via Sample Complexity Bounds

4.1 Summary

This chapter[3] applies the computational learning theory framework to elucidate the differences between learning breast cancer and ancestral origins concepts from high-throughput SNP profiles, to help clarify which learning problems are easier to solve, in terms of sample complexity. While most of these published predictive studies present the empirical error of a model used to learn a specific biomedical phenotype given a group of subjects profiled by a recent omics measurement technology, very few explain why learning is feasible in some cases and infeasible in others. Our experiments on high-throughput germline SNP profiling showed that some tasks (eg, predicting continental and subcontinental ancestral origins of individuals) are quite easy, while others (such as predicting the susceptibility to breast cancer) are extremely difficult. Our analysis suggests that the ancestral origin prediction problem is a case of *realizable learning* (from a

[3] This chapter is prepared based on the following papers:
Hajiloo M: **Learning disease patterns from novel high-throughput genomics profiles: why is it so challenging?**, *Lecture Notes in Computer Science* 2013, **7884**: 328-333.

Hajiloo M, Greiner R: **Assessing the feasibility of learning biomedical phenotypes via large scale omics profiles**, NIPS Workshop on Machine Learning in Computational Biology (NIPS MLCB), Lake Tahao, USA, December 2013.

simple hypothesis class such as 3-node decision trees), albeit from many irrelevant features. This suggests the sample complexity upper-bound of $O\left(\frac{1}{\varepsilon}\left(\ln|H_{AO}| + \ln\left(\frac{1}{\delta}\right)\right)\right) = O\left(\frac{1}{\varepsilon}\left(\mathrm{r} \times \ln(p) + \ln\left(\frac{1}{\delta}\right)\right)\right)$ where r is the number of relevant features, p is the number of features in a high-throughput scan of SNPs, H_{AO} is the hypothesis class of 3-node decision tree out of p features, ε is the estimation error, and δ is the confidence parameter. On the other hand, the breast cancer prediction problem is a case of *unrealizable learning* (unrealizable even from a class of very expressive hypotheses, such as 10-term 20-feature DNFs) with *relevant hidden features* and *hidden subclasses*. This suggests the sample complexity lower-bound of $\Omega\left(\frac{L_{H_{BC}}}{\varepsilon^2}\left(d + \ln\frac{1}{\delta}\right)\right)$, where H_{BC} is the hypothesis class of 10-term 20-feature DNFs, $L_{H_{BC}}$ is the optimal Bayes error of learning given H_{BC}, d is the VC dimension of H_{BC}, ε is the estimation error, and δ is the confidence parameter. These findings can help future omics researchers interested in predictive studies by suggesting how they can estimate the necessary and sufficient number of training examples required for their predictive studies.

4.2 Background

In addition to the widespread interest of the biomedical community in conducting association and risk assessment studies on datasets generated by omics profiling, many biomedical researchers are now performing predictive studies. However, most of these predictive studies experimentally mix-and-match different algorithms for pre-processing, feature selection, and learning, then conclude by presenting the empirical error of their model using an evaluation strategy such as cross validation or evaluation on a hold-out dataset. The omics field would benefit from analytical assessments that mathematically explain why learning is feasible in some cases but infeasible in others. This chapter begins to fill this gap by introducing the computational learning theory framework to the biomedical community and using this framework to elucidate the differences between two recently published predictive studies, predicting breast cancer and ancestral origins, each using the omics profiles generated by genotyping the germline SNPs of samples on Affymetrix Human SNP 6.0 array [1-2]. These two studies had similar sample and feature sizes, but the prediction performance of the resulting predictive models were very different. As these studies lie at the extremes of the spectrum of predictability of biomedical phenotypes, analyzing them using the computational learning theory lens may help us to understand the limits of learnability of biomedical phenotypes.

In the breast cancer prediction task [1], we utilized 696 samples (348 breast cancer cases and 348 apparently healthy controls) to learn a model that predicts whether a new subject will develop breast cancer, based on her SNP profile. Despite trying a wide range of biologically-aware and biologically-naïve (statistical) supervised learning approaches, we could never achieve the generalization error better than 0.4. In the ancestral origin prediction task [2], we utilized the international HapMap project Phase II and III datasets [3] to learn models that can predict an individual's continental and subcontinental ancestral origins. While the breast cancer prediction had only marginal success, it was very easy to achieve generalization errors of less than 0.1 in predicting continental and subcontinental ancestral origins. For example, in the continental ancestral origin prediction problem, using 270 samples (1/3 in each continent), a single CART decision tree [4] with 3 internal nodes (SNPs), had the generalization error rate of 0.03, and an ensemble of 3 disjoint decision trees with 3-4 internal nodes (SNPs) each, achieved the generalization empirical error rate of 0.

4.3 Methods: Ancestral Origin Learning Case

It is well-known that an individual's SNP profile is sufficient to identify his/her ancestral origins [5]. Our recent analysis on learning continental and subcontinental ancestral origins confirms that a small set of SNPs provides the information needed to identify one's ancestral origins [2]. In each of these continental and subcontinental classification problems, we identified many equally good concepts/patterns, in form of disjoint small decision trees (ie, whose features were disjoint); as these patterns were accurate and diverse, we were able to increase the model accuracy by making an ensemble over these disjoint decision trees [6]. Our empirical study suggests that using a small sample size in order of hundreds of samples suffices for learning ancestral origins from high-throughput SNP profiles using models as simple as small decision trees.

From the computational learning theory viewpoint, the ancestral origin learning problem is a case of realizable learning (from a simple hypothesis class such as decision trees) in presence of many irrelevant features. Here, we present the definition of realizable learning and learning in presence of many irrelevant features and their associated PAC learning bounds. Section 4.4.1 provides an estimate of the sample complexity upper-bound for PAC learning the target concept of ancestral origins.

4.3.1 Realizable Learning

The optimal Bayes error (approximation error) of learning the target concept c using the hypothesis class H (L_H), is the error of the hypothesis h∈ H, with the minimum disagreement with c on the input data ($L_H = \inf_{h \epsilon H}$ {error(h) = $\Pr_{X \epsilon D}[c(x) \neq h(x)]$}). A hypothesis class H over the input space X for learning the target concept c is realizable if the optimal Bayes error of H equals zero and is unrealizable if it is greater than zero. A target concept is realizable given a hypothesis class if it is fully expressible using at least one of the hypotheses of that class and if the training labels provided for the learner are not noisy.

Theorem 1: The sample complexity upper-bound for finding a hypothesis h, in the finite hypothesis class H, which is (ε, δ)-close to the target concept c in the realizable learning case is $m_U = \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$ [7].

This means that regardless of the complexity of the target concept and the complexity of the distribution on the input data, having a training dataset of size greater than or equal to m_U is *sufficient* to guarantee that we can find a hypothesis h that is ε-close to the target concept c, with probability of greater than or equal 1-δ.

Theorem 2: The sample complexity lower-bound for finding a hypothesis h, in the hypothesis class H with VC dimension d, which is (ε, δ)-close to the target concept c in the realizable learning case is $m_L = \frac{1}{\varepsilon}\left(\max\left(\frac{d-1}{4}, \frac{\ln\frac{1}{\delta}}{8}\right)\right) = \Omega\left(\frac{1}{\varepsilon}\left(d + \ln\frac{1}{\delta}\right)\right)$ [8-10].

This means that, regardless of the complexity of the target concept and the complexity of the distribution on the input data, having a training dataset of size of greater than or equal than m_L is *necessary* to guarantee that we can find a hypothesis h that is ε-close to the target concept c, with probability of greater than or equal 1-δ.

4.3.2 Learning with Many Irrelevant Features

Learning with many irrelevant features means the training dataset offered to the learner contains many features that are irrelevant to the target concept.

Theorem 3: Let X be the input space of p binary features and $H_{p,r}$ be the hypothesis class of all Boolean functions over r out of those p features. The sample complexity upper-bound for finding a hypothesis h, in the finite hypothesis class $H_{p,r}$, which is (ε, δ)-close to the target concept c in the realizable learning case, given r relevant features, p-r irrelevant features is $m_U = O\left(\frac{1}{\varepsilon}\left(\ln|H_{P,r}| + \ln\frac{1}{\delta}\right)\right) = O\left(\frac{1}{\varepsilon}\left(2^r + r \times \ln p + \ln\frac{1}{\delta}\right)\right)$ [11].

Theorem 4: Let X be the input space of p binary features and $H_{p,r}$ be the hypothesis class of all Boolean functions over r out of those p features with VC dimension d. The sample complexity lower-bound for finding a hypothesis h, in the finite hypothesis class $H_{p,r}$, that is (ε, δ)-close to the target concept c in the realizable learning case, given r relevant features, p-r irrelevant features, is

$m_L = \Omega\left(\frac{1}{\varepsilon}\left(d + \ln\frac{1}{\delta}\right)\right) = \Omega\left(\frac{1}{\varepsilon}\left(2^r + r \times \ln p + \ln\frac{1}{\delta}\right)\right)$ [11].

These bounds suggest that both sample complexity upper-bound and lower-bound have logarithmic dependence on the number of irrelevant features. This implies that the presence of many irrelevant features does not make the learning task substantially more difficult, at least in terms of the number of examples needed for learning. However, depending on the algorithm used, the computational complexity might be an issue while dealing with many irrelevant features [11]. Although we provided the sample complexity bounds for the general hypothesis class of all Boolean functions over r out of those p features, it is easy to show that this logarithmic dependence holds for specific hypothesis classes of Boolean functions, such as r-conjunctions and s-term r-DNFs, as well.

4.4 Methods: Breast Cancer Learning Case

Like most cancers, breast cancer occurs because of an interaction among many environmental, lifestyle, and genetic factors. The major environmental and lifestyle risk factors include age, lack of childbearing or lack of breastfeeding, obesity, estrogen exposure (from endogenous and exogenous sources), radiation exposure, certain chemicals exposure, smoking, alcohol intake, and physical inactivity [12]. Moreover, heritable genetic factors include point mutations, SNPs, CNVs, and structural chromosome variations [13]. Among these many different genetic, environmental, and lifestyle factors, our learner was only given SNPs in our breast cancer learning problem [1]. Furthermore, breast cancer is biologically heterogeneous disease, with a high degree of diversity between and within tumors as well as among cancer-bearing individuals [14]. However, these distinctions are ignored in our dataset, as these subclasses are merged into the single "breast cancer" label.

From the computational learning theory viewpoint, because of the above-mentioned reasons, the breast cancer learning problem is a case of unrealizable learning with relevant hidden features and hidden subclasses. Here, we present the unrealizable learning and learning with hidden subclasses and their associated PAC learning bounds. Section 4.4.2 provides an estimate of the sample complexity lower-bound for PAC learning the target concept of breast cancer.

4.4.1 Unrealizable Learning

The sample complexity upper-bound and lower-bound for unrealizable learning where the optimal Bayes error is not equal to zero ($L_H > 0$) are as follows:

Theorem 5: The sample complexity upper-bound for finding a hypothesis h, in the finite hypothesis class H, which is (ε, δ)-close to the optimal concept in H, in the unrealizable learning case is $m_U = \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right) = O\left(\frac{1}{\varepsilon^2}\left(\ln|H| + \ln\frac{1}{\delta}\right)\right)$ [15].

Theorem 6: The sample complexity lower-bound for finding a hypothesis h, in the hypothesis class H with VC dimension d and optimal Bayes error L_H, which is (ε, δ)-close to the target concept c in the realizable learning case is $m_L = \frac{L_H}{4\varepsilon^2}\left(\max\left(\frac{d-1}{8}, \ln\frac{1}{4\delta}\right)\right) = \Omega\left(\frac{L_H}{\varepsilon^2}\left(d + \ln\frac{1}{\delta}\right)\right)$ [9-10].

4.4.2 Learning with Hidden Subclasses

In many real-world learning problems, such as our breast cancer learning problem, there are hidden subclasses in the labels provided to the learner. These implicit subclasses increase the complicatedness of the target concept c. We use two learning problem examples to give an idea on how the sample complexity of learning increases when there are hidden subclasses:

Learning Problem Example 1: Let the target concept in a learning problem be a specific conjunction of r features ($c_1 = x_1 \wedge x_2 \wedge \dots \wedge x_r$) out of p features from the concept class C_1 of conjunctions of r out of p features. Now, let us employ the hypothesis class $H_1 = C_1$ to PAC learn the target concept. What are the sample complexity upper-bound (m_{u1}) and lower-bound (m_{L1}) for PAC learning this problem?

Learning Problem Example 2: Let the target concept in a learning problem be a specific s-term r-DNF function out of p features ($c_2 = T_1 \vee T_2 \vee \dots \vee T_s$ in which each $T_i = x_{i1} \wedge x_{i2} \wedge \dots \wedge x_{ir}$) from the concept class C_2 of s-term r-DNF functions. Now, let us employ the hypothesis class $H_2 = C_2$ to PAC learn the target concept. What are the sample complexity upper-bound (m_{u2}) and lower-bound (m_{L2}) for PAC learning this problem?

As both of these problems are realizable learning problems with optimal Bayes error of 0, we can apply Theorems 1 and 2 to calculate and then compare their sample complexity upper-bound and lower-bound. Therefore:

$$m_{u1} = O\left(\frac{1}{\varepsilon}\left(\ln\left(\binom{\mathrm{p}}{\mathrm{r}} \times 2^{\mathrm{r}}\right) + \ln\frac{1}{\delta}\right)\right) \leq O\left(\frac{1}{\varepsilon}\left(\ln(\mathrm{p}^{\mathrm{r}} \times 2^{\mathrm{r}}) + \ln\frac{1}{\delta}\right)\right)$$

$$m_{u1} = O\left(\frac{1}{\varepsilon}\left(\mathrm{r} \times \ln(\mathrm{p}) \ + \ln\frac{1}{\delta}\right)\right)$$

$$m_{u2} = O\left(\frac{1}{\varepsilon}\left(\ln\left(\binom{\binom{\mathrm{p}}{\mathrm{r}} \times 2^{\mathrm{r}}}{\mathrm{s}}\right) + \ln\frac{1}{\delta}\right)\right) \leq O\left(\frac{1}{\varepsilon}\left(\ln\left(\left(\binom{\mathrm{p}}{\mathrm{r}} \times 2^{\mathrm{r}}\right)^{s}\right) + \ln\frac{1}{\delta}\right)\right) \leq$$

$$O\left(\frac{1}{\varepsilon}\left(\ln((\mathrm{p}^{\mathrm{r}} \times 2^{\mathrm{r}})^{\mathrm{s}}) \ + \ln\frac{1}{\delta}\right)\right)$$

$$m_{u2} = \ O\left(\frac{1}{\varepsilon}\left(\mathrm{s} \times \mathrm{r} \times \ln(\mathrm{p}) \ + \ln\frac{1}{\delta}\right)\right)$$

Theorem 7: For 1 ≤ r ≤ p and 1 ≤ s ≤ $\binom{p}{r}$, let C be the class of functions expressible as s-term r-DNF formulas and let q be any integer, r ≤ q ≤ p such that $\binom{q}{r}$ ≥ s. Then a lower-bound for the VC dimension of C is $s \times r \ \times \ \left\lfloor \log_2(\frac{p}{q}) \right\rfloor$ [16].

Given Theorems 2 and 7 and considering the fact that conjunctions of r out of p features can be basically considered as 1-term r-DNFs, we find

$$m_{L1} = \Omega\left(\frac{1}{\varepsilon}\left(r \times \log_2(p) + \ln\frac{1}{\delta}\right)\right)$$

$$m_{L2} = \Omega\left(\frac{1}{\varepsilon}\left(s \times r \times \log_2(p) + \ln\frac{1}{\delta}\right)\right).$$

For the analysis of the breast cancer learning problem, we use a target concept such as c_2 as each of its s terms can model one of the subclasses of the heterogeneous phenotype of breast cancer. The above-mentioned results show that both of sample complexity bounds grow linearly in s if we consider the target concept to be in form of a s-term r-DNF formula.

4.5 Results and Discussion

4.5.1 Ancestral Origin Learning Case

Here, we calculate the sample complexity upper-bound for learning the target concept of continental ancestral origins (African vs. Asian vs. European) using the hypothesis class H_{AO} of 3-node decision trees out of p features. The same sort of analysis can explain the story of the subcontinental population identification problems as well. Theorem 1 provides the sample complexity upper-bound in the realizable learning case. Given this theorem, we first calculate $|H_{AO}|$ and then compute the *sufficient* number of samples to find a hypothesis h which is (ε, δ)-close to the target concept.

The number of hypotheses in H_{AO} ($|H_{AO}|$) is equal to the number of ways to select 3 out of p features ($\binom{p}{3}$), times the number of different binary trees with 3 internal nodes (5), times the number of ways to branch on each internal node (2^3), times the number of different assignments of 3 labels to 4 external nodes (3^4).

Therefore:

$$|H_{AO}| = 3240 \times \binom{p}{3} \leq 3240 \times \frac{P^3}{6}$$

The sample complexity upper-bound to find a hypothesis h in H_{AO} which is (0.05, 0.01)-close to the target concept, given p = 611146 features (SNPs) would be 1018, as:

$m_U = \frac{1}{\varepsilon}\left(\ln|H_{AO}| + \ln\frac{1}{\delta}\right) \leq \frac{1}{0.05}\left(\ln\left(\frac{3240}{6}\right) + 3 \times \ln(611146) + \ln\frac{1}{0.01}\right)$ = 1018.

This result confirms that the presence of many irrelevant features does not make learning substantially difficult due to the logarithmic dependence of this bound on the number of irrelevant features. Furthermore, our result implies that even in the worst possible choice for the data distribution, having 1018 samples is sufficient for finding a 3-node decision tree that is (0.05, 0.01)-close to the target concept. Note that this is only a factor of ~4 times more samples than what we had.

4.5.2 Breast Cancer Learning Case

Here, we calculate the sample complexity lower-bound for learning the target concept of breast using the hypothesis class H_{BC} of 10-term 20-DNF formulas out of p features. Theorem 6 provides the sample complexity lower-bound in the unrealizable learning case. Given this theorem, we first calculate the VC dimension d of H_{BC} and then compute the *necessary* number of samples to find a hypothesis h which is (ε, δ)-close to the target concept.

Given Theorem 7, the VC dimension of H_{BC} (d) equals:

d = $s \times r \times \left\lfloor \log_2\left(\frac{p}{q}\right)\right\rfloor$

d = $10 \times 20 \times \left\lfloor \log_2(\frac{p}{21})\right\rfloor$

Note that q=21 is the smallest number that satisfies both conditions in Theorem 7 (r ≤ q ≤ p and $\binom{q}{r}$ ≥ s).

The sample complexity lower-bound to find a hypothesis h in H_{BC} which is (0.05, 0.01)-close to the optimal concept in this hypothesis class, given the fact that none of the hypotheses in H_{BC} have an error rate of zero due to the existence of relevant hidden features ($L_{H_{BC}} \approx 0.3$ as our experimental analysis in Chapter 2 suggests), given p = 506836 features (SNPs) would be 10915, as:

$$m_L = \frac{L_H}{4\varepsilon^2}\left(\max\left(\frac{d-1}{8}, \ln\frac{1}{4\delta}\right)\right)$$

$$\geq \frac{0.3}{4\times(0.05)^2}\left(\max\left(\frac{\left(10\times20\times\left\lfloor\log_2\left(\frac{506836}{21}\right)\right\rfloor\right)-1}{8}, \ln\left(\frac{1}{4\times0.01}\right)\right)\right) = 10915.$$

This result implies that in the worst case, 10915 samples are necessary for finding a 10-term 20-DNF formula that is (0.05, 0.01)-close to the optimal concept which itself is $L_{H_{BC}}$-close (0.3-close) to the target concept. Note that our result is given with the assumption that breast cancer has 10 different subtypes each expressible by a conjunction of 20 variables. However, the actual concept class of the target might be even much more complex than this. Furthermore, note that this is a factor of ~18 times more samples than what we had. Of course, the analysis given here is for the worst possible choice of the target concept and data distribution and our breast cancer task might not necessarily be such a pessimistic case.

4.6 References

1. Hajiloo M, Damavandi B, Hooshsadat M, Sangi F, Cass CE, Mackey JR, Greiner R, Damaraju S: **Using genome wide single nucleotide polymorphism data to learn a model for breast cancer prediction**, *BMC Bioinformatics* 2013, **14**(S13): S3.

2. Hajiloo M, Sapkota Y, Mackey JR, Robson P, Greiner R, Damaraju S: **ETHNOPRED: a novel machine learning method for accurate continental

and sub-continental ancestry identification and population stratification correction, *BMC Bioinformatics* 2013, **14**(1): 61.

3. The International HapMap Consortium: **The International HapMap project**. *Nature* 2003, **426**: 89-796.

4. Quinlan JR: **Induction of decision trees**. *Machine Learning* 1986*,* **1***:*81-106.

5. Allocco DJ, Song Q, Gibbons GH, Ramoni MF, Kohane IS: **Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms**. *BMC genomics* 2007, **8**(1): 68.

6. Kuncheva LI, Whitaker CJ: **Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy.** *Machine Learning* 2003, **51**(2)**:**181-207.

7. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK: **Occam's razor**. *Information processing letters* 1987, **24**(6): 377-380.

8. Ehrenfeucht A, Haussler D, Kearns M, Valiant L: **A general lower bound on the number of examples needed for learning**. *Information and Computation* 1989, **82**(3): 247-261.

9. Devroye L, Lugosi G: **Lower bounds in pattern recognition and learning**. *Pattern recognition* 1995, **28**(7): 1011-1018.

10. Devroye L, Györfi L, Lugosi G: *A probabilistic theory of pattern recognition.* vol. 31. Springer 1996.

11. Almuallim H, Dietterich TG: **Learning Boolean concepts in the presence of many irrelevant features**. *Artificial Intelligence* 1994, **69**(1): 279-305.

12. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461:**747-753.

13. Cho WC: *An Omics Perspective on Cancer Research.* New York, NY: Springer 2009.

14. Bertucci F, Birnbaum D: **Reasons for breast cancer heterogeneity**. *Journal of Biology* 2008, **7**(2):6.

15. Haussler D: **Decision theoretic generalizations of the PAC model for neural net and other learning applications**. *Information and computation* 1992, **100**(1): 78-150.

16. Littlestone N: **Learning quickly when irrelevant attributes abound**. *Machine Learning* 1988, **2**(4): 285-318.

Chapter 5

Conclusion

We applied experimental studies based on the supervised learning framework and analytical studies based on the computational learning theory framework to address our research questions. Our results direct us to the following thesis statements:

I. Developing an *accurate* predictive model for breast cancer susceptibility is *not feasible* from a labeled training dataset that contains only hundreds of samples given the hundreds of thousands of features in the form of germline genome-wide scan of single nucleotide polymorphisms (SNPs). The highest LOOCV accuracy of the predictive models built for this task is 59.95%, which is only slightly over the baseline of 51.52%. Our investigation suggests that this breast cancer prediction task is a case of unrealizable learning (unrealizable even from a class of very expressive hypotheses, such as 10-term 20-feature DNFs) with relevant hidden features and hidden subclasses. If so, then a simple sample complexity analysis suggests that we will need many more training examples that were available (by a factor over 18).

II. Developing *accurate* predictive models for continental and subcontinental ancestral origins is *feasible* given a labeled training dataset with hundreds of samples and the hundreds of thousands of features in the form of germline genome-wide scan of single nucleotide polymorphisms (SNPs). The 10-fold CV accuracy of the predictive models built using our ETHNOPRED learning method are 100% ± 0%, 86.5% ± 2.4%, 95.6% ± 3.9%, 95.6% ± 2.1%, 98.3% ± 2.0%, and 95.9% ± 1.5% respectively for the continental, European, East Asian, African, North American, and Kenyan population identification problems. These results suggest that the ancestral origin prediction task is a case of realizable learning (from the class of simple 3-node decision trees), albeit from many irrelevant features. This suggests that a small sample size is sufficient, which is why our learning was successful.

Although this dissertation addressed predictive study of breast cancer susceptibility and ancestral origins given SNP profiles, our results are applicable for analyzing other biomedical phenotypes given other omics profiles as well. We conclude by presenting a summary of the contributions of this research study and summarizing some possible future directions of this research.

5.1 Summary of the Contributions

There are several published results describing models learned to predict biomedical phenotypes from some omics profile. However, the field somehow lacks a precise presentation of the limits of applicability of predictive studies. This dissertation aimed to fill this gap to some extent by introducing predictive studies and contrasting them with association and risk assessment studies in Chapters 1, 2, and 3. Furthermore, we identified a number of common pitfalls in the design and evaluation of the performance of published predictive studies. These pitfalls lead to an overestimation of the accuracy of the predictive models. These pitfalls, as explained in Chapter 2, are 1) evaluating the model's performance without using a hold-out set based on cross validation only, 2) feature selection based on a dataset and testing the resulting model (that used those features) on the same dataset, 3) learning parameter value selection without using cross validation, and 4) comparison of different models without application of any hypothesis testing method such as t-test and permutation test. As we were aware of these common pitfalls, we designed our studies to avoid these mistakes.

Chapter 2 introduced the first genome wide predictive study of breast cancer susceptibility using high-throughput SNP profiles. Although the predictive models designed for breast cancer susceptibility in this chapter have limited accuracy and are not yet clinically relevant, we demonstrated that there is still a signal in high-throughput SNP profiles related to breast cancer susceptibility and that the supervised learning framework is capable of finding this signal. We believe we could boost the accuracy for predicting breast cancer susceptibility if we could overcome the limitations of this study (as presented in section 2.4). We expect taking the following steps would address these limitations: 1) integrating additional environmental, lifestyle, and omics profiles of the study participants, 2) recruiting more study participants, and 3) including of the breast cancer subtypes in the labeling of the study participants. Note that this is consistent with our analytic evaluation, presented in Chapter 4.

Chapter 3 proposed a novel supervised learning algorithm that learns an ensemble of disjoint decision trees, ETHNOPRED, for geographical categorization of individuals based on their genography. This algorithm generates models that have some nice properties at the performance time: 1) high accuracy, 2) robustness to missing SNP values, 3) cost-efficiency due to the small number of features

(SNPs) used in the models, and 4) high interpretability due to the visible structure of the model. We believe our algorithm or modified versions of it can provide these excellent properties for other predictive studies as well.

Chapter 4 applied the computational learning theory framework to reveal the differences between learning breast cancer susceptibility pattern and ancestral origins pattern from high-throughput SNP profiles from the sample complexity viewpoint. It also provided the PAC learning framework that can be used to assess the feasibility of conducting predictive studies on biomedical phenotypes. This framework can help researchers to design better predictive models, at least by telling them how large a training dataset is needed.

5.2 Future Directions

Possible future directions for this research are to explore the following research questions:

I. Chapters 2 and 4 suggest that the predictive study of susceptibility to a *multifactorial* disease, such as breast cancer, via profiling study subjects using a single omics profiling technique, such as SNP profiling, leads to a model with limited accuracy. How will the accuracy of the learned classifier change if we provide other omics (genomics, epigenomics, transcriptomics, metabolomics, etc), environmental, and lifestyle profiles of the subjects as well?

II. Chapters 2 and 4 suggest that the predictive study of susceptibility to a *heterogeneous* disease such as breast cancer via providing high-level organ-based labeling of the study subjects leads to a model with limited accuracy. Does a more accurate labeling of training examples, in terms of clarification of diseases subtypes, help to build more accurate predictive models for susceptibility to a multifactorial disease?