05-2015

Epistasis, Polygenic Effects, and the Missing Heritability Problem: A Review of Machine Learning as Applied to Genetic Association Studies

Peter Jenson

Epistasis, Polygenic Effects, and the Missing Heritability Problem:

A Review of Machine Learning as Applied to Genetic Association Studies

By

Peter Jenson, B.S.

Capstone Project

Submitted in partial fulfillment of the requirements for the degree of

Master of Biomedical Informatics

Oregon Health and Science University

May 2015

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

PETER J JENSON

"*Epistasis, Polygenic Effects, and the Missing Heritability Problem: A Review of Machine Learning as Applied to Genetic Association Studies*"

Has been approved

Michael Mooney, PhD

Capstone Advisor

Epistasis, Polygenic Effects, and the Missing Heritability Problem: A Review of Machine Learning as Applied to Genetic Association Studies

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

TABLES

FIGURES

ACKNOWLEDGEMENTS

I'd like to thank Michael Mooney for serving as my advisor and guiding my efforts throughout this project. I'd also like to thank all of the other professors I've had throughout my time at OHSU, including Shannon McWeeney, Kemal Sonmez, Eilis Boudreau, and Beth Wilmot. I'd also like to thank Diane Doctor for assisting me with the necessary leave-of-absence periods that I had to take over the last year.

Finally, I'd like to give a special thanks to my wife-to-be, Carrie Hartley, for being there for me always and providing me with the encouragement and support I needed to complete this project. The past year or so has been very….. eventful. And I unfortunately don't mean that in a good way. Quite frankly, I'm thankful simply to have survived. Having someone by my side at the hospital and throughout my recovery has been absolutely invaluable and I don't know that I'd still be alive to write this if it were not for her taking me to the ER initially and being there for me consistently throughout the entire ordeal. Nonetheless, the future looks very bright even though I can now very much attest to the fascinating unpredictability of life. All in all, I'm just thankful to be alive, and I remain enthusiastic as ever about what the future holds in store.

ABSTRACT

In recent years, there has been a significant academic drive to determine the cause of the missing heritability problem which is commonly associated with the genetics of complex traits such as complex diseases. In this paper, I present an in-depth literature review by introducing the nature of this problem and discussing the current state of genetic association studies which commonly involves multi-locus models as predictors (as opposed to the conventional single-locus association studies) that became prominent with the advent of genome-wide sequencing technology. Such genotype-phenotype models that involve sets of genes as predictors is wrought with analytical challenges, such as the curse of dimensionality, heterogeneity, and small main effect sizes. Machine learning as a general method to aid in the search for the missing heritability shows promise as exemplified by numerous studies by other authors, presented in this paper, demonstrating the utility of machine learning methods for addressing the specific challenges that exist in modern genetic association studies. Within the context of genetic association studies, I introduce machine learning as a general concept, present a basic overview for several different machine learning methods, and go on to present an expansive discussion on neural networks and multifactor dimensionality reduction.

INTRODUCTION

With the advent of genome-wide sequencing technology, genome-wide association studies (GWAS) have become a reality and a popular method of discovering genetic variants associated with phenotypes of interest such as complex diseases. Although such studies have identified over 1,200 genetic loci significantly correlated with over 165 common human diseases and traits [1][2][3], these identified loci of interest explain only a minor proportion of the heritability of such phenotypes. Such has been dubbed the "missing heritability" problem wherein the genetic variants that we are capable of identifying as significant correlates of common human traits fail to explain most of the heritability that we can calculate from correlations of relatives using family data. It has been recently hypothesized that this missing heritability is a result of failure to take into account epistatic interactions among causative genetic loci. Some authors have argued, for example, that we may have already identified the majority of the genetic loci necessary to account for this missing heritability, but our failure to accurately model genetic interaction among these genetic variants has resulted in the missing heritability phenomenon[4]. Zuk et. al proposed a model of epistatic interaction for Crohn's disease which takes into account 71 known risk-associated loci [4]. With the usual assumption that the loci contribute to Crohn's disease purely through additive effects, only 21.5% of the estimated heritability is accounted for. However, when the authors' proposed model that takes into account epistatic interaction is used, they find that 80% of the missing heritability for Crohn's disease can be explained by the 71 known risk-associated genetic loci. Others have suggested that the sets of loci known to be associated with common traits are incomplete and that the phenomenon of missing heritability likely arises from our failure to detect more of those genetic loci that are truly causative factors [5]. Such authors argue that the missing heritability

problem is not simply a result of failure to accurately model genetic interaction among known causative genetic variants, but instead a failure to have identified a significant number of genuinely causative variants as being statistically significant which then leads to incomplete genotype-phenotype models for traits under investigation [5].

The missing heritability problem has been a significant issue in the study of complex traits such as complex diseases [6]. Complex diseases are the result of many different factors – factors which are often both unknown and, rather than being independent contributors, statistically dependent on one another. Many common diseases have been studied and found to be definitively complex in nature, such as various kidney diseases, asthma, Alzheimer's, Parkinson's, osteoporosis, multiple sclerosis, diabetes, cancer, autism, alcoholism, drug addiction, and many others [7]. Such diseases are the collective result of genetic variants at many different loci (such as in cases of epistatic interaction among genes, where genes at different loci are not independent from one another in their effect but instead interact together in a non-additive manner to produce a given phenotype) and environmental factors [6]. As such, they do not display standard Mendelian patterns of inheritance. An individual having inherited known genetic risk factors for a disease does not necessarily mean that that they will actually develop that disease, as the effects of the known genetic risk factors may in fact be dependent on other genetic variants as well as environmental factors. Because there are so many variables responsible for complex diseases, many of them undergoing statistical interaction with one another, detection of the underlying genetic causes for such diseases has historically been a difficult endeavor that is still very much an evolving process.

In this review, I focus on the application of machine learning methods to the discovery of epistasis and polygenic effects in complex diseases. The definition of epistasis can itself be a controversial issue. In Fisher's working definition which tends to be the most commonly accepted definition today [8], genetic epistasis is a non-additive interaction between alleles at different loci that serves as a causative factor in determining phenotype [9]. For our purposes here, the term "epistasis" will be analogous to Fisher's definition in which alleles at different genetic loci interact in a non-additive manner such that they enhance, dampen, or modify entirely each other's effects with regard to their contribution to phenotype. I also focus on non-epistatic polygenic effects (e.g. additive effects) which is relevant to the purpose of this paper given that genetic loci can contribute additively to a phenotype but with such a minute effect on an individual level that we are often incapable of detecting such loci as causative genetic factors [10][11]. Additionally, epistatic interactions themselves still can actually directly lead to these same kinds of additive effects that are so minute that we are often incapable of detecting them [5].

Machine learning, explicitly defined later in this paper, is a process that takes advantage of available data to develop and refine a working statistical model. In the area of complex diseases, we take advantage of machine learning and genetic data in the attempt to discover and quantify the underlying genetic risk factors and disease contributors. In the early stages of using machine learning methods to study the genetics of diseases, such methods were employed for the purpose of analyzing the independent effect of a single genetic loci with regard to a given disease of interest. These methods are insufficient to adequately analyze the genetic contributors involved in complex diseases, as the statistical effect of genetic loci that are genuinely causative may appear to be negligible using such analytical methods which are incapable of detecting

interactions among loci [12] [13]. Multi-locus methods have been developed to analyze the effect that sets of loci have on complex disease development. Such methods are capable of quantifying the behavior of loci within the more biologically-accurate context of a set of genes wherein the loci interact with one another to produce the phenotype. In complex disease development, many loci may act collectively with and dependent upon other loci to exert the causative phenotypic effect of interest even when the effect of an individual contributing loci may appear to be statistically insignificant when viewed as an independent effector [11][14].

In this paper, I will present an in-depth literature review of the topics introduced here. I focus on two specific multi-locus machine learning methods, neural networks and multifactor dimensionality reduction, and use the information regarding the theory and application of such methods to construct an overarching theme of the utility of machine learning methods in general for genetic association studies, particularly with regard to complex traits and GWAS. I will begin by presenting an overview of several different machine learning methods and then go on to thoroughly discuss neural networks and multifactor dimensionality reduction specifically.

METHODS

This literature review was performed in order to identify challenges in genetic association studies and demonstrate the ability of machine learning techniques to address these problems and act as valuable tools for performing genetic association studies. A comprehensive literature review of the latest research papers and articles surrounding genetic association and machine learning applications in this field was performed. This literature review utilized articles with the following attributes:

- Journal articles that were published in peer reviewed journals.
- Articles that referred to genetic association studies and machine learning techniques that are explicitly used in genetic association studies.
- Compiled and presented on Pubmed
- Priority was given to articles published most recently

During the in-depth literature review that was performed for this paper, many journal articles were read in their entirety and categorized. In addition to the papers on the subject of genetic association studies, numerous articles were read that corresponded to various machine learning techniques. The following bullet points identify primary attributes utilized for selecting significant articles to be used in the literature review:

- Category of article:

- Primary source article describing a technique/concept and used to introduce and provide the background for a subject.
- Analysis of a machine learning technique and comparison of various machine learning techniques in genetic association studies.
    - These papers were especially important in identifying variations to commonly used machine learning techniques in the field of genetic association studies.
- Genetic association studies that utilized machine learning techniques to achieve some outcome of interest, such as the identification of epistatic interactions or causative genes for phenotypes of interest
    - These studies were utilized to provide the reader with the context of the current use of machine learning techniques that were discussed in this paper in the field of genetic association studies.
    - These articles were also useful in identifying current uses of machine learning techniques in genetic association studies and providing examples of recent discoveries that were found using machine learning techniques.

- Abbreviated list of terminology and words of interest used in searching for articles (the full list of all search terms and queries used can be found in the Appendices):
    - Articles that contained the terms“genetic association”, “genome wide association”, “GWAS”, were discovered and read to gain a background understanding of this area.
    - Articles that contained the term “missing heritability” were discovered and read in order to gain an understanding of this problem

- Articles that used the term "machine learning" in conjunction with each of the aforementioned terms such as "genetic association" were also found to identify the appropriate machine learning techniques and gain a general understanding of the field
- Articles were also queried using the specific name of a machine learning technique (e.g. Neural Networking or MDR) and the aforementioned terms such as "genetic association" in order to gain an understanding of how such machine learning techniques have been and are currently being utilized in the field of genetic association studies
- Articles that used the terms "epistasis", or "interaction" in conjunction with "gene", "genome", or "genetic", were discovered and read in order to gain an understanding of epistasis and gene-gene interactions

A total of 116 articles were found to be relevant for inclusion in this literature review. Abstracts were read in order to determine whether or not the paper would be possibly relevant for inclusion, and those deemed possibly relevant were read in their entirety and filtered to the final number of 116 papers selected for inclusion in this literature review. The journal articles that were discovered and found to be significant were used in this paper to provide the proper background and framework for presenting machine learning techniques that are utilized in genetic association studies. In addition, these papers served to address the attributes of specific machine learning techniques including their strengths, weaknesses, and potential alterations or variations. Finally, the articles used in this literature review also provided context for the current use of identified machine learning techniques in modern genetic association studies and were

used to provide the reader with specific examples of these machine learning techniques being utilized in genetic association studies.

RESULTS AND DISCUSSION

Machine Learning: A Necessary Approach

In the ongoing effort to determine the reason behind the missing heritability problem, taking into account epistatic interactions in both disease models and association studies is a promising direction to take. However, statistical models that take into account gene-gene interactions are wrought with a plethora of challenges that must be recognized and accounted for (Table 1). In addition, several other significant challenges exist in genetic association analyses that are independent of gene-gene interactions such as genetic heterogeneity [15][16] (Table 1). Genetic heterogeneity arises when a single phenotype is independently associated with different loci or different sets of loci (the term also describes the phenomenon in which a single phenotype is independently associated with different alleles at the same locus). If heterogeneity is not accounted for by the analytical method, statistical power to detect the associated genetic variants is limited [5]. Consider, for example, some hypothetical disease *X* and several hypothetical genetic variants g_1, g_2, ..., g_n that all independently lead to the *X* phenotype. If a case-control study involves a case group that consists of individuals with many of the different genetic causes for the disease, and the individuals with any particular variant do not possess any of the other variants, the presence of these genetic subgroups together in the same case group will confound any analysis that attempts to find a single genetic variant significantly statistically associated with the entirety of the case group. Methods that account for heterogeneity often involve cluster analysis, wherein the analysis of genetic markers is used to build groups of individuals with similar genetic profiles so that case-control comparisons can be made for each of the different genetic profiles as case groups [17].

As touched upon earlier in this paper, complex diseases also involve both additive and nonadditive interactive effects among associated genes. These effects can be associated with an increase in the probability of making a type II statistical error, wherein truly associated genetic markers are falsely determined to not be associated with the phenotype of interest. For example, the main effect of individual genetic markers may be so small as to be statistically insignificant even when each of those markers actually contributes to the phenotype of interest. The significance of such genes may only be observable when they are viewed as a *set* of independent variables in a multidimensional predictive model which demonstrates the additive effect [10][11][18]. Epistatic interactions among genetic loci can themselves, in fact, directly lead to such additive effects that are so minute that typical analytical methods fail to detect the effect at all [5]. As well, when a loci's primary contribution to the phenotype of interest occurs as epistatic interaction(s) with other genes, the main effect of that gene may be statistically insignificant (or even nonexistent) and therefore remain unidentified as a loci of interest for the phenotype being examined [12]. Thus, multidimensional methods that account and search for the existence of both additive effects and interaction among genetic markers are very useful tools in the study of complex disease genetics.

The 'curse of dimensionality' and the multiple testing problem –two very commonly encountered problems in association studies—must also be accounted for by the analytical methods. The curse of dimensionality arises in genetic studies that model gene-gene interactions as a consequence of the fact that, as the number of genes being studied increases, the number of potential interactions occurring between those genes increases in an exponential manner [13]. Machine learning methods exist that can account for this through feature selection processes that filter the predictor combinations under investigation to include only those that are determined to

be likely to be significant for the purpose of the study. The curse of dimensionality also results in circumstances where the involved sample size is not adequately large enough to accurately accommodate the statistical model under investigation that involves multiple genes under study [19]. In fact, in some circumstances, there may not exist *any* samples in the dataset for which specific predictive multi-dimensional models of interest can be evaluated. For example, suppose we are studying some particular phenotype and we're interested in determining if a significant interaction exists between smoking and some genetic marker *x*. If the sample size is inadequate, there may exist no individuals in the dataset for which both binary variables are positive –that is, the sample did not include any individuals who are both a smoker and a carrier of genetic marker *x*—and thus the interaction of interest cannot be studied with the given dataset and machine learning methods cannot account for this. The multiple testing problem refers to the increase in the probability of making a type I statistical error (the probability of rejecting the null hypothesis when it is, in fact, true) as multiple different independent variables or sets of independent variables are tested in the same data. As more independent variables are included in the model, the likelihood of an association being detected as significant, when it is merely a result of random chance (e.g. the independent variable is not actually a causative factor for the outcome), increases. Although methods exist to account for this problem, employing them results in a decrease in the overall statistical power of the analysis [20]. A summary of these major challenges encountered in modern genetic association studies is provided in Table 1.

Fortunately, many different machine learning methods exist that provide significant aid in addressing these challenges encountered in modern genetic association studies. In the 1983 first-volume edition of *Machine Learning*, Herbert A. Simon wrote "Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task (or tasks

drawn from a population of similar tasks) more effectively the next time". Borrowing from Simon's definition, machine learning can be defined here as the branch of artificial intelligence (AI) which concerns the study and development of AI systems which are adaptive in response to the input of data. Machine learning methods are essential to the study of genetic data, both as a result of the size of the data (e.g. number of genomic loci for which variations are possible, number of variants at each locus, etc.) as well as the complexity of the genotype-phenotype relationship (e.g. epistasis, polygenic effects in general, and heterogeneity). Even discounting the latter consideration of complexity, the sheer size of the data involved in genomic study necessitates computer-driven processes to observe and compare across all the individual data points at all the different genetic loci under consideration, particularly with regard to genome-wide association studies (GWAS) which compare case and control groups or analyze quantitative trait status across millions of known single nucleotide polymorphisms (SNPs) [21].

Table 1: Challenges in Finding the Missing Heritability	**Description**
Heterogeneity	Arises when the trait of interest in the dataset has different causative factors that each independently lead to the trait. For example, a disease may be caused by multiple genotypes that each independently lead to the same disease status (genetic heterogeneity.

Small Main Effect	Arises when the main effect of a genuinely causative gene is so small as to be statistically insignificant. Additionally, the joint effect of a set of genes may be significant even though the main effect of each gene is statistically insignificant and therefore undetectable when modeled as an independent predictor. As discussed, epistatic interactions can also directly lead to main effect sizes that are so small as to be statistically insignificant and undetectable [5][12]. This issue is related to the issue of inadequate sample size given that small sample sizes result in diminished statistical power to detect smaller effect sizes.
Curse of Dimensionality: Combinatorial Complexity	As the number of predictors in the model increases, the number of potential interactions increases exponentially. This makes studying gene-gene interactions difficult due to the extraordinary combinatorial complexity involved in doing so when the model uses a large number of genes and/or gene variants as predictors.
Curse of Dimensionality: Inadequate Sample Size	The sample size can be inadequate to test for certain interactions of interest, such as in those circumstances where there exists few to no cases in the sample for which certain predictive models can be evaluated. This issue is related to the issue of small main effects given that small main effects are less likely to be detected the smaller the sample size is.
Multiple Testing Problem	This follows directly from computational complexity. As the number of predictors in the model increases, the chance of making a type I statistical error increases solely as a result of evaluating multiple predictors sequentially.

Table 1 - Summary of Major Challenges in Modern Genetic Association Studies

---

The branch of artificial intelligence that is machine learning encompasses many different methods and techniques. There exists several different classifications of machine learning methods, such as supervised or unsupervised learning. In supervised machine learning methods, algorithms are used to analyze data objects to determine which pre-defined class the object belongs to. An example of supervised learning would be the analysis of images of individual handwritten numbers with the purpose of the algorithm being to correctly assign each data point to the actual number it was meant to represent. In unsupervised machine learning methods (AKA "cluster analysis"), data points are analyzed for the purpose of establishing clusters of data points

that are similar in their characteristics where pre-defined classes for the data points do not yet exist. An example of unsupervised learning would be the analysis of genetic data in order to group individual data points together into clusters based on some measure of genetic similarity where no specific classes of genetic similarity for the data have yet been defined. As touched upon earlier, this grouping method is commonly used to account for heterogeneity when performing genetic studies [17]. Machine learning methods can also be classified based on whether or not they depend on parametric statistical modeling [22]. There are strengths and weaknesses generally associated exclusively with either class. In general, parametric methods can be more informative from a quantitative standpoint while carrying greater statistical power (ability to detect differences where differences exist) if assumptions about the data are correct, whereas non-parametric methods are capable of handling a much larger number of predictor variables accurately than parametric methods while also being less dependent on assumptions about the data [23]. The choice of which class of methods to use to analyze the genetic data associated with complex diseases will depend largely on the question(s) posed by the investigator, as both of the classes of methods are uniquely capable when compared to the other.

In this review, I will expand in-depth on two prominent machine learning methods used today in genetic association studies: neural networks and multifactor dimensionality reduction. However, in addition to neural networks, multifactor dimensionality reduction, and the many variants associated with both of these machine learning methods, many other major machine learning methods also exist such as support vector machines, random forests, and k-nearest neighbors. Each of the machine learning methods selected for inclusion here in this paper have

been employed in recent years specifically to address challenges in modern genetic association analysis. Table 2 contains an overview of each of these selected machine learning methods, each of which has been used extensively in genetic association studies while having variants that have been developed to add functionality and predictive power for the specific challenges faced in genetic association studies.

Table 2: Selected Machine Learning Methods	**Brief Description**
Neural Networks	Neural networks are one of the most common methods used in genetic association studies [24]-[31]. The algorithm was built in emulation of a biological nervous system, wherein an interconnected network of neurons or "nodes" exists and a threshold value must be reached by a node before it fires its signal to an adjacent node. These typically consist of an input layer wherein data is fed into the neural network, a hidden layer that acts as a "black box" wherein the connected nodes perform computation, and an output layer wherein an output of interest such as classification is delivered. These are capable of approximating any function and are thus universal approximators. Many variants such as genetic programming neural networks [25] and grammatical evolution neural networks [26] have been developed that add functionality that is particularly useful for genetic association studies.
Multifactor Dimensionality Reduction	MDR divides the data into numerous partitions to perform k-fold cross validation. The data is partitioned into both a training data set and a testing data set. Next, every possible multifactor class has a case/control ratio calculated and each multifactor class is screened based on a pre-set case/control ratio. For example, with three loci each with two different variants, there would be a total of 8 possible genotype combinations each defining a separate multifactor class. This facilitates the detection of non-additive interactions between predictors, such as in cases of epistatic interaction among genes. The multifactor classes are separated into two groups, a low risk group and a high risk group, based on their case/control ratio. This process is then repeated for each possible cross-validation interval with the error being assessed each time using the data partition reserved for testing. The model with the highest prediction accuracy among all those produced by cross-validation is then selected for use. Given that MDR is an exhaustive search, the method has been criticized for a failure to scale well to very high-dimensional data such as GWAS with a large number of predictors. As such, variants such as generalized MDR and 3 way split MDR have been utilized to reduce the computational power needed to perform

	an MDR analysis [32]-[35]. MDR also possesses other variations developed to address other challenges in genetic association studies such as heterogeneity or unbalanced datasets[36]-[42].
Support Vector Machines	The support vector machine method is a non-probabilistic binary linear classifier that creates a linear separator to divide data points into two separate clusters. In addition to linear separators, kernel functions can be used in SVM to create non-linear separators so that classes that are not linearly separable can be accurately distinguished. Training data is used to train the SVM model through the assignment of examples to one of two possible categories. This training phase iterates until an optimum margin is created using the calculated separator between the classification of the two categories. The SVM method has been extensively used recently in genetic association studies [43]-[48], and variations of SVM have been created, such as the multiclass SVM, to allow for the analysis of data with more than two possible outcomes[43][46][49]-[52].
Random Forest	Random forest is based on creating a “forest” of binary decision trees where in the tree base learner is typically grown using the CART (Classification and Regression Tree) methodology. It utilizes multiple decision trees created from random subsets of predictor variables. These trees are created from a training set that usually consists of a bootstrap sampling equal to ⅔ of the data being analyzed. The unused data is called “out of bag” data and is used with the training data to estimate the error of each tree. The multiple trees that were created are screened and the best decision at each node is selected based on aggregate voting across all trees in the forest. This method is known to be less prone to overfitting than other machine learning methods. However, the generation of multiple trees and the computations involved are extensive and generally have a higher computational time than other methods. To account for this, variations such as random jungle have been developed to utilize modern computing processes that decrease the computational time needed to perform an analysis while retaining the same predictive power as traditional random forest methods [53]-[57]. The random forest technique is a relatively new method in the field of machine learning, and several authors have conducted analyses showing that random forests can outperform many other methods with regard to predictive power [53]-[56].
K Nearest Neighbors	A method that utilizes training data to assign a class membership to clusters of data points. The value “K” is user defined and the algorithm uses this value to determine the individual classes based on the training data. For a given query point (an input data point), the algorithm assigns the query point to the majority class among its K nearest neighbors as calculated by some distance metric (e.g. Euclidean distance). This method iterates until a class membership model is created that is below an acceptable maximum threshold in training data. K nearest-neighbor is another machine learning technique that has seen use in genetic association studies in recent years [58]-[62]. Like the other machine learning methods listed here in this table, variations of K nearest-neighbor such as Relief-F and ReliefSeq have been developed specifically to address the challenges faced in modern genetic association studies [58][63].

Table 2 - A Fundamental Overview of Selected Machine Learning Algorithms

Neural Networks

Artificial neural networks (which will hereafter be referred to simply as "neural networks") were developed in emulation of the biological neural networks that compose a living nervous system. In the simplest description, this involves a biological network composed of individual neurons which, upon receiving a stimulating input that achieves or exceeds some threshold value, transmits a signal to adjacent neurons via axonal connections. Neural networks can be thought of as composed of three separate layers: (1) an "input layer", wherein data as initially fed into the network, (2) a "hidden layer", wherein input data is processed, and (3) an "output layer", wherein the processed data is compiled, analyzed, and the output of interest (e.g. "True" or "False") is computed. Neural networks are a non-parametric statistical technique that makes no underlying assumptions about the data [64]. Given that neural networks do not depend on the construction of a model built on assumptions, they are particularly effective choices for the analysis of high-dimensional datasets such as in genome-wide association studies [64]. Neural networks can also be used to approximate any function and are thus considered to be universal approximators. Additionally, neural networks have been shown to perform very well with respect to computation time even when dealing with massive high-dimensional datasets [65].

Basic Theory and Methodology

The single-layer perceptron model, a very simple and basic feedforward neural network algorithm, is a linear separator used for the purpose of binary classification. The single-layer perceptron model is limited in that it can only accurately identify classes that are linearly

separable from one another. It consists of a simple two-layer model with a simple output function, though many different input nodes are typical and acceptable. Figure 1 shows a typical perceptron neural network. The leftmost layer is the input layer and it consists of nodes labeled from i_1 to i_n, with each input node taking as input a different data point from an input dataset. Each of these input data points are multiplied by the respective weight values w_1 through w_n and the resulting product is fed forward into the next node, labeled with a "∑" symbol. This node sums up the products fed into it from all the input nodes according to equation 1.

$$x = \sum (i_1 * w_1) + (i_2 * w_2) + \ldots + (i_n * w_n) \quad (1)$$

The resulting *sum* is then fed forward into the threshold output node. The output node typically consists of a threshold sigmoid function similar to that shown in Equation 2 for the purpose of delivering a binary output variable that can be either a 1 or a 0. As data in the network flows only in one direction, from the previous layer to the following layer with no loops occurring, this is an example of a feedforward neural network.

$$\mathrm{F(x)} = \{1, if\ \frac{1}{1+e^{-x}} \geq 0.5, 0, if\ \frac{1}{1+e^{-x}} < 0.5 \quad (2)$$

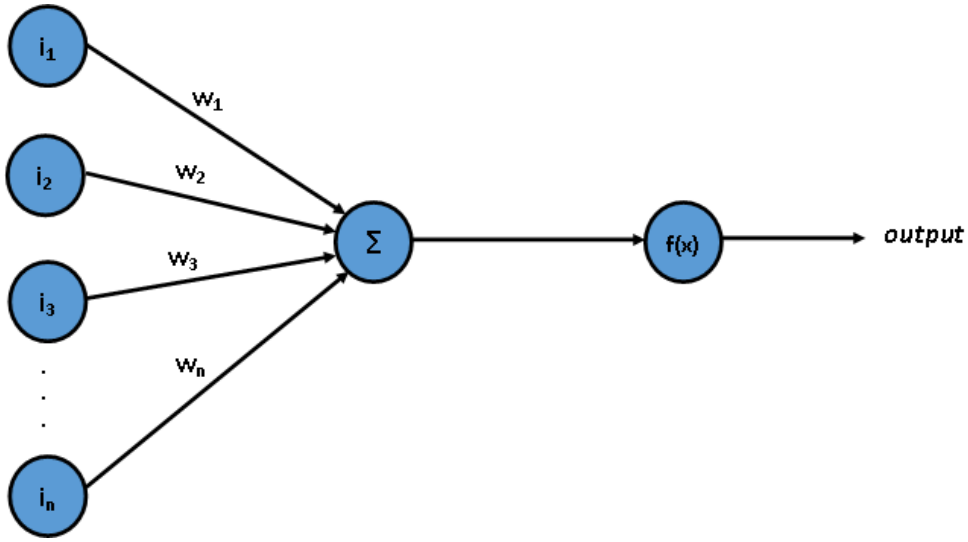

Figure 1 - Visual Representation of a Single-Layer Perceptron Neural Network

Training such a neural network to be an effective classifier involves systematically modifying the weight values. The neural network employs a backpropagation algorithm to train itself that results in the network weights being adjusted until some threshold level of accuracy is achieved. The backpropagation algorithm requires a differentiable error function as in Equation 3 that is used to update the network weights and train the model after each data case is fed through the network.

$$E = \frac{1}{2}(output - target\ value)^2 \quad (3)$$

Adjusting the network weights in response to error is the key to training the neural network to output accurate classification. The values for these weights are updated after each run using the

backpropagation algorithm. This algorithm adjusts the weights in the network in the direction necessary to lessen the error on the previous run. This continues in a stepwise fashion until some threshold level of accuracy is achieved, at which point the neural network is trained and ready for classification.

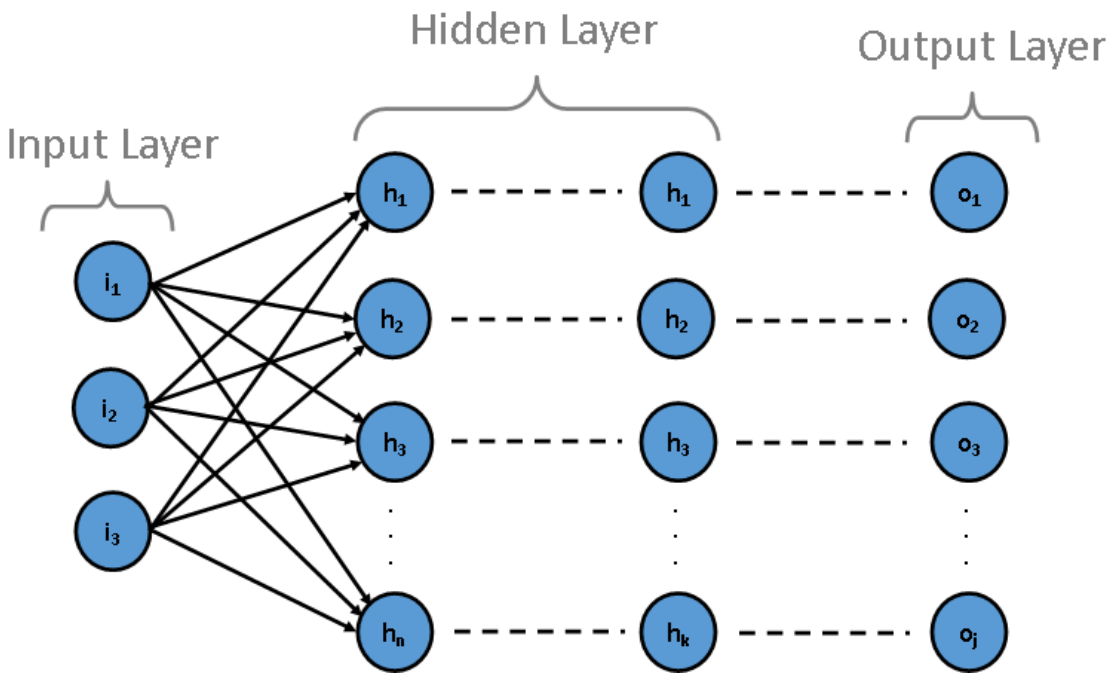

Figure 2 - Visual Representation of a Multilayer Perceptron Neural Network

This basic perceptron model for neural networks can be extended to include three or more layers which allows the model to be capable of accurately classifying data regardless of whether

or not the individual classes are linearly separable. Such models consist of (1) an input layer; (2) a hidden layer that can contain an arbitrary number of layers within itself, each with an arbitrary number of nodes; and (3) an output layer that can consist of as many nodes as is necessary for the problem. Figure 2 shows an example of such a model with three input nodes (though this can be as many as desired), an arbitrary number of hidden layers and nodes within each layer, and an arbitrary number of output nodes. It is standard for every node in any particular layer to be connected to every node in the subsequent layer, and Figure 2 represents this concept using a series of horizontal dashed lines. Multi-layer perceptron models are much more powerful than single-layer perceptron models and are, in fact, capable of approximating any kind of functional relationship regardless of the complexity of the function [66]. The hidden layer in such networks is often considered a kind of "black box" wherein the explicit details of the connections and weights are not necessarily known by the user as the network can become very complex. Nonetheless, the functionality is what is important, and the neural network is designed to train and optimize itself in response to input data on its own without direct human oversight of the hidden layers being either necessary or desirable.

Neural networks have proven to be an extremely effective and useful tool for the analysis of genetic data, such as in identifying genetic markers associated with a particular disease phenotype. When using a neural network for such a task, genetic data typically serves as the input with the output of interest being different phenotypes. The encoding of that input and output data, however, can vary depending on the type of analysis being conducted. For example, the input data can be used to indicate whether or not certain genetic markers, such as SNPs, are present, with a "0" value indicating that the marker is not present while a "1" value indicates that the marker is present. The output can then be either a "1" to indicate that a phenotype of interest,

such as disease status, is predicted to be present in the subject, or a "0" to indicate that the phenotype is not predicted to be present in the subject.

Modern Variants of Neural Networks

Many of the more modern variants to neural networks focus on architecture optimization and feature selection. The first neural network algorithms involved a neural network architecture that was largely a product of design based on estimation [24]. As discussed, neural networks are very much a kind of digital black box in which data can be input in order to, fascinatingly, achieve output data that accurately classifies the input cases. That hidden layer that acts as black box, however, must be specifically defined in terms of the number of hidden layers and number of neurons in each hidden layer. It is the weights between nodes in the hidden layer that are trained and automatically altered by the algorithm; the number of hidden layers and number of neurons within the hidden layers, however, are predefined and held static using the standard backpropagation algorithm. Newer neural network algorithms have been developed, however, that train and modify not just the weights in the hidden layer, but also the entirety of the neural network architecture itself including the number of input nodes (predictors), the number of hidden layers, and the number of nodes within each hidden layer.

Ritchie et al. used a genetic programming algorithm to build the architecture of a neural network and demonstrated that the neural network achieved through genetic programming had a lower prediction error and higher power as compared to the standard backpropagation neural network constructed without architecture optimization through genetic programming [67]. The application of genetic programming to the construction of neural networks was done to improve

upon the standard process of constructing a neural network wherein the general architecture is largely a product of arbitrary decision. Referred to as GPNNs for "genetic programming neural networks", these neural networks are constructed through a standard genetic algorithm process that involves (1) random generation of neural network architectures, including random selection of predictors and number of nodes in the network, (2) training each of the individual neural networks through backpropagation, (3) evaluating and ranking the performance of each individual neural network, (4) selecting the highest-performing network architectures and randomly combining their features (representing genetic crossover) and adding in random modifications (representing genetic mutation) to create a new set of network architectures, and (5) repeating the entire process on the new generation of network architectures until some endpoint is achieved (e.g. preset number of generations or a threshold level of accuracy) at which point the single best-performing neural network from the latest generation is chosen for use.

The methods used to optimize neural network architecture can vary widely, with many recent developments in neural network algorithms that focus on optimizing architecture through feature selection showing promise as effective tools for detecting epistatic interactions in genetic data that would otherwise be missed [24][25]. Feature selection is particularly important in GWAS that aim to detect gene-gene interactions given that, as the number of predictors under consideration increases, the number of potential interactions between them increases exponentially. This combinatorial explosion can make certain analyses computationally infeasible, and feature selection algorithms are often used to address the challenge that results from this by filtering the set of predictors to include only those that are determined to be statistically likely to be of importance. Some examples of these feature selection algorithms would be the already-mentioned genetic programming neural networks (GPNN) [67],

grammatical evolution neural networks (GENN) [26], and Bayesian neural networks [68]. Selected variants of neural network algorithms with accompanying descriptions are given in Table 3. Yet, in spite of the many evolutions of neural network algorithms that have occurred since they became a popular tool, the standard multilayer perceptron model, a basic feedforward neural network that uses backpropagation to achieve its trained state, remains popular and in wide use due to its proven effectiveness as a classification and pattern recognition tool to aid in biomarker identification and classification [27]-[31].

Table 3: Neural Network Variants	**Description**
Backpropagation Neural Network [64]	A standard neural network model that involves using a backpropagation algorithm for the purpose of training the network weights. The backpropagation algorithm uses a partial derivative to determine how the overall error of the network changes with respect to an individual weight. The value of each weight is shifted by a pre-determined quantity in the direction necessary to lessen the error associated with the output of the network and this procedure is repeated for each data point used in the training dataset. This method does not inherently involve any feature selection or architecture optimization.
Genetic Programming Neural Networks (GPNN) [67]	Utilizes a genetic programming algorithm to optimize the architecture of the neural network. This algorithm works by randomly generating neural network features (such as the predictors to use as input) to serve as the parent generation, training each network, evaluating the performance of each network, creating an offspring generation of neural networks by randomly combining the features of the highest-performing parents, and repeating the process until some pre-defined threshold is achieved.
Grammatical Evolution Neural Networks (GENN) [26]	This method is a "cousin" to GPNN in that it utilizes a genetic algorithm to optimize the architecture of the neural network. It differs in that it uses a context-free grammar (CFG) to define the individual networks used in the algorithm. Grammatical evolution uses a population of linear genomes that are translated by the grammar into a phenotype which, in the case of GENN, is a neural network. Using a grammar to map a neural network results in a separation between

	genotype and phenotype, and the result is a greater genetic diversity in the population of neural networks as compared to GPNN.
Bayesian Neural Networks (BNN) [68]	Similar to a standard backpropagation neural network, but these differ in that the architecture is a product of a Bayesian prior distribution rather than being a product of somewhat arbitrary user design. Training is done by sampling from a Bayesian joint posterior of the network structure and weights. Instead of considering a single answer to question of interest, Bayesian methods allow for the consideration of a distribution of possible outcomes. Additionally, given that every parameter of a BNN is given a prior distribution, variable importance can be quantified and used for the purpose of feature selection resulting in a feature-trimmed model that is less computationally expensive than it would otherwise be.
Fuzzy Neural Network [69]	Neural networks constructed with the use of fuzzy logic to map the input space to the output space. Fuzzy logic operates on the basis of approximations and degrees of truthfulness (or belonging in the case of classification) as opposed to definite values and boolean logic. A useful feature of fuzzy neural networks is that the input space can be mapped to the output space by a series of IF-THEN rules (as opposed to mathematical summation of signals as is typical in non-fuzzy neural networks). This can lead to an enhanced classification accuracy when analyzing imprecise or uncertain data such as heterogeneous traits in genetic association studies.

Table 3 - Selected Neural Network Variants with Accompanying Descriptions

Neural Networks and Gene-Gene Interactions

Neural networks have been used extensively in the literature to detect gene-gene interactions. By their very nature, they are well-suited for this task given that they can accurately model arbitrarily complex relationships among the input nodes even when they interact in an intricate manner (such as when more than two SNPs are involved in an interaction that affects trait status). Although this methodology is well-suited for the detection of interactions among predictors as well as for accurate outcome prediction when the predictors interact significantly, it is not possible to determine the specific type of interaction that the predictors undergo (e.g.

additive, multiplicative, or epistatic) from the output of a neural network. Nonetheless, the utility of neural networks for detecting the presence of gene-gene interactions has been demonstrated extensively in the literature, even if the method is not inherently capable of determining which kinds of interaction are at play in a particular genotype-phenotype relationship.

Gunther et al. investigated the ability of standard feedforward multilayer perceptron neural networks with backpropagation to model epistatic interactions in six separate two-locus disease models [70]. The authors compared the performance of neural networks to several different logistic regression models using simulated data (Table 4). The six different epistatic models used are described in detail by the penetrance matrices and accompanying description given in Figure 3. To generate the data, a population of one million individuals with two biallelic loci, *A* and *B*, was generated for each epistatic model for both a high-risk and low-risk scenario (presented in Table 5) for a total of twelve populations. The genetic information was drawn randomly for both loci with a minor allele frequency of 0.3. It was assumed that both loci are in linkage equilibrium and that Hardy-Weinberg equilibrium holds. The case-control status was drawn according to the probability given by the epistatic penetrance matrices using the numbers in Figure 3. For each of the twelve populations, parameters were selected to achieve a disease prevalence equal to 0.01. In standard fashion, the genotype at each loci served as the input data for the neural network while the output node represented the penetrance (or probability of disease status) for each genotype used as input. The methods were each evaluated on their ability to estimate the theoretical penetrance matrix for each disease model.

ADD: $f_{ij} = P(Y = 1 \mid G_A = i, G_B = j) = a_i + b_j$

HET: $f_{ij} = a_i + b_j - a_i \cdot b_j$

MULT: $f_{ij} = a_i \cdot b_j$

EPI RR: $f = [P(Y = 1 \mid G_A = i, G_B = j)]_{i,j} = \begin{bmatrix} c & c & c \\ c & c & c \\ c & c & r \cdot c \end{bmatrix},$

EPI DD: $f = \begin{bmatrix} c & c & c \\ c & r_1 \cdot c & r_1 \cdot c \\ c & r_1 \cdot c & r_2 \cdot c \end{bmatrix}$

EPI RD: $f = \begin{bmatrix} c & c & c \\ c & c & c \\ r_1 \cdot c & r_1 \cdot c & r_2 \cdot c \end{bmatrix}$

Figure 3 - Penetrance Matrices for Disease Models [70]

In this figure, *i* and *j* represent the genotypes at the two loci G_A and G_B, respectively; *c* denotes the baseline risk and *r* denotes the risk increase or decrease. By choosing the parameters *r*, r_1, r_2 and the ratios a_1/a_o, a_2/a_0, b_1/b_0, and b_2/b_0, respectively, different risk scenarios can be generated. The first model (ADD) is a standard additivity model for the two loci. The second model (HET) assumes that there is both an additive and multiplicative interaction between the two loci. The third model (MULT) assumes that the interaction between the two loci is entirely multiplicative. The fourth model (EPI RR) assumes that both

genes have a recessive effect on the disease where both loci must each carry the mutated allele in order for an increased or decreased risk to be observed. In the fifth model (EPI DD) both loci are assumed to be dominant, and the occurrence of an increased or decreased risk occurs only if both loci carry at least one mutated allele. The sixth model (EPI RD) is a mixed epistatic model in which one gene has a dominant effect on the disease while the other gene has a recessive effect. [This figure is reused in accordance with the permissions set forth by the original publisher of this content, BioMed Central]

For both the ADD and HET model in [70], and for both the low-risk and high-risk scenarios in which each model was studied, the neural network outperformed logistic regression as measured by the difference between the predicted and actual penetrance for the simulated data. For the MULT model, and for both the low-risk and high-risk scenarios, logistic regression outperformed the neural network algorithm by a significant margin. For the EPI RR and EPI RD models, and for both the low-risk and high-risk scenarios, neural networks substantially outperformed logistic regression. For the EPI DD model, logistic regression slightly outperformed the neural network algorithm in the high risk scenario, while the neural network algorithm slightly outperformed logistic regression in the low risk scenario. The quantitative results are summarized in Table 6, and the authors chose to omit the numerical data for the HET model while explaining that it was "virtually identical" to the data for the ADD model. The authors posit that the reason why logistic regression outperformed the neural network algorithm for the MULT model is due to the MULT model corresponding exactly with the model that logistic regression inherently uses for analysis. The authors concluded that neural networks are a "promising" tool for genetic analysis and are capable of successfully modeling several different

kinds of gene-gene interactions, but that further research is necessary to take full advantage of their utility.

Nonetheless, this study is significant for several reasons. The first is that it demonstrated the ability of even the standard multilayer perceptron neural network algorithm with backpropagation, one of the most basic neural network forms, to model gene-gene interactions. The second is that it demonstrated the ability of neural networks to detect multiple kinds of gene-gene interactions, such as additive, multiplicative, epistatic, and mixed. The third is that it demonstrated the ability of neural networks to predict the *probability* that a given genotype is a disease case (the penetrance for each genotype being examined) when disease status is dependent on epistatic effects. The fourth is that it showed that even the most basic neural network form offers comparable if not superior performance as compared to logistic regression for detecting gene-gene interactions, and this is a notable feature given that neural networks are known to scale well to high-dimensional datasets whereas logistic regression methods do not [70]. In spite of the significance of this study, is not without limitations. The authors did not explore the feature selection capabilities of advanced neural network algorithms such as GPNN or GENN, and this feature is critically important in the analysis of GWAS data and one of the unique benefits associated with the application of neural networks to GWAS. Additionally, the neural networks that the authors used had between two and five neurons, and it is possible that the neural networks would have performed even better had the authors used a more complex architecture (e.g. more hidden layers and neurons).

Number of parameters.

	Neural network		
0 hidden neurons	3		
1 hidden neuron	5		
2 hidden neurons	9		
3 hidden neurons	13		
4 hidden neurons	17		
5 hidden neurons	21		
	Logistic regression	**Logistic regression (DV)**	
Null model (NM)	1	1	
One main effect (SiA/SiB)	2	3	
Both main effects (ME)	3	5	
Full model (FM)	4	9	

Table 4 - Number of parameters and basic model information for each neural network and logistic regression model used for comparison [70]

[This figure is reused in accordance with the permissions set forth by the original publisher of this content, BioMed Central]

Risk scenarios.

Two-locus disease model	Low risk scenario	High risk scenario
ADD, HET, MULT	$a_1 = 2 \cdot a_0$	$a_1 = 5 \cdot a_0$
	$a_2 = 4 \cdot a_0$	$a_2 = 10 \cdot a_0$
	$b_1 = 5 \cdot b_0$	$b_1 = 5 \cdot b_0$
	$b_2 = 10 \cdot b_0$	$b_2 = 10 \cdot b_0$
EPI RR	$r = 5$	$r = 10$
EPI DD, EPI RD	$r_1 = 2$	$r_1 = 5$
	$r_2 = 4$	$r_2 = 10$

Table 5 - Risk Scenarios Investigated for Each of the Disease Models in Figure 3 [70]

[This figure is reused in accordance with the permissions set forth by the original publisher of this content, BioMed Central]

Table 6	Mean Absolute Difference			
	Low Risk		High Risk	
	Neural Network	Logistic Regression	Neural Network	Logistic Regression
Additive Model (ADD)	0.2313	0.253	0.2059	0.2544
Multiplicative Model (MULT)	0.2428	0.1637	0.2178	0.1833
Epistatic Recessive (EPI RR)	0.2071	0.3503	0.141	0.2755
Epistatic Dominant (EPI DD)	0.3095	0.3071	0.2524	0.2648
Epistatic Mixed (EPI RD)	0.2239	0.2799	0.1563	0.2329

Table 6 - Summary of Results from [70]

The values for mean absolute difference quantify the difference between the actual penetrance for the simulated dataset and the predicted penetrance achieved using the respective models. The values for logistic regression here are taken in each instance from the single best performing model the authors employed.

More recently, Beam et al. proposed the use of a Bayesian neural network as a useful alternative to several other machine learning methods when it comes to performing genome-wide association studies (GWAS) that account for epistasis [68]. The authors proposed the use of a Bayesian neural network as an answer to some of the challenges in computational complexity associated with GWAS that take into account epistasis. Bayesian neural networks are similar to

the standard feedforward backpropagation neural networks described earlier in this paper, but they differ in that (1) the architecture of the hidden layer is a product of a Bayesian prior distribution, and (2) the neural network is trained by sampling from a Bayesian joint posterior of the network structure and weights. As modern GWAS typically consider several million single-nucleotide polymorphisms, the number of potential interactions to consider from such a study is on the order of trillions. Given that, in Bayesian neural networks, every parameter of the neural network is given a prior distribution, model features such as variable importance can be quantified and used to build the architecture of the network in a deliberate and statistically-motivated manner. This is a very useful aspect of Bayesian neural networks within the context of GWAS given that the set of predictors under consideration can be filtered to include only those with a posterior distribution centered around larger values, effectively reducing the dimensionality of the analysis and therefore reducing the computational complexity involved.

Additive risk model

Genotype	AA	Aa	aa
BB	η	$\eta(1+\theta)$	$\eta(1+2\theta)$
Bb	$\eta(1+\theta)$	$\eta(1+2\theta)$	$\eta(1+3\theta)$
bb	$\eta(1+2\theta)$	$\eta(1+3\theta)$	$\eta(1+4\theta)$

Threshold risk model

Genotype	AA	Aa	aa
BB	η	η	η
Bb	η	$\eta(1+\theta)$	$\eta(1+\theta)$
bb	η	$\eta(1+\theta)$	$\eta(1+\theta)$

Epistatic risk model

Genotype	AA	Aa	aa
BB	η	η	$\eta(1+4\theta)$
Bb	η	$\eta(1+2\theta)$	η
bb	$\eta(1+4\theta)$	η	η

Table 7 - Risk Models Evaluated [68]

The symbols η and θ in the tables represent the baseline risk and effect size, respectively. The authors simulated genotypes for the disease SNPs for a range of minor allele frequencies (MAFs) and simulated the disease status for 1,000 cases and 1,000 controls using the risks given in this table. They embedded the causal SNPs in a background of 998 non-causal SNPs, for a total of 1,000 SNPs to be considered. For each combination of effect size, $\theta \in \{0.5, 1.0, 1.5, 2.0\}$, $MAF \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and model type (Additive, Threshold and Epistasis) they generated 100 datasets. This yielded a total of 6,000 datasets for evaluation. [This figure is reused in accordance with the Creative Commons Attribution (CC BY) license granted by BMC Bioinformatics that is applicable to this content]

The authors in [68] compared the statistical power of Bayesian neural networks to several other methods in two separate phases. The authors defined statistical power as the fraction of datasets for which each method was able to correctly identify both disease-causing SNPs. In the first phase, the authors used each method to analyze 6,000 simulated case and control datasets that each involved 1000 SNPs. The risk models evaluated in the first phase are provided in Table 7. The authors compared Bayesian neural networks to the chi-squared method and BEAM across four different effect sizes (0.5, 1.0, 1.5, and 2.0) five different minor allele frequencies (0.1, 0.2, 0.3, 0.4, and 0.5). For the additive and threshold model, the superiority of Bayesian neural networks to detect causative SNPs with small effect sizes and small minor allele frequencies was significantly greater than either of the other methods tested. For instance, in the additive model with an effect size of 0.5 across all MAFs tested, Bayesian neural networks had an average statistical power of ~0.85, whereas the second best performing method, the chi-squared test, had an average statistical power of ~0.35. Additionally, in the threshold model for the effect sizes of 0.5 and 1.0 across all MAFs tested, Bayesian neural networks had an average statistical power of ~0.39 (high of 0.90, low of 0.0), whereas the second best performing method, again the chi-squared test, had an average statistical power of ~0.22 (high of 0.76, low of 0.0). However, for the epistatic model, the results are somewhat mixed. At the effect size of 0.5, BEAM had an average statistical power of ~0.43 across all minor allele frequencies tested, whereas Bayesian neural networks (the second best performing method here) had an average statistical power of ~0.34. At the effect size of 1.0, Bayesian neural networks had an average statistical power of ~0.72 across all minor allele frequencies tested, whereas BEAM (the second best performing method here) had an average statistical power of ~0.66. These results demonstrate that Bayesian neural networks offer similar and in some cases superior performance

when compared to the chi-squared method and BEAM when it comes to genetic analysis that involves epistatic disease models. This is particularly interesting given that Bayesian neural networks can be used in GWAS to reduce the computational complexity involved in the analysis while still performing very well as compared to other methods that might be substantially more computationally complex.

In the second phase [68], the authors analyzed Bayesian neural networks against several other methods, including multifactor dimensionality reduction (MDR), using a simulated dataset in which the causative loci influence trait status purely through epistatic interaction. The comparison took place over a purely epistatic model with 5% heritability, and another purely epistatic model with 10% heritability. In this model, there are no main effects and, thus, the authors argue that causative SNPS should be more difficult to detect than in any of the models used in the first phase. For the model with 5% heritability, Bayesian neural networks outperformed all other methods for the MAF of 0.05 with regard to statistical power (1.00 for Bayesian neural networks compared to ~0.85 for the second highest performer that was MDR), but MDR outperformed all other methods for all other MAFs (average statistical power of ~0.95 for MDR compared to 0.61 for the second highest performer that was Bayesian neural networks). For the model with 10% heritability, MDR outperformed Bayesian neural networks by a non-trivial margin (average statistical power of ~0.8 for Bayesian neural networks compared to 1.0 for MDR). Although multifactor dimensionality reduction (described in the following section of this paper) demonstrated superior performance over Bayesian neural networks as measured by power to detect causative SNPs in these genetic models that involve significant epistasis with no main effects, the authors that argue that Bayesian neural networks are still a very useful method for genetic association studies, especially genome-wide association studies, given that

multifactor dimensionality reduction does not scale well to high-dimensional datasets. Overall, this study is notable in that it proposed a new variant of neural networks based on Bayesian probability that demonstrates statistical power comparable to top-performing methods such as MDR while also being much less computationally demanding. Additionally, the Bayesian variant of neural networks proposed in this study offers the direct ability to quantify variable importance which in and of itself is a primary goal of association studies [68]. Finally, given that the analysis used 1000 SNPs as independent variables, this study also illustrates the utility of machine learning methods for detecting gene-gene interactions in high-dimensional genetic data.

The same authors in [68] also carried out another investigation into the utility of Bayesian neural networks for detecting causal genetic loci in the presence of significant epistatic interaction [71]. What makes this particular study unique and provoking, the authors argue, is that they used simulated datasets to model a variety of outcomes whereas most other studies have investigated the significance of epistasis solely in case-control studies or single-value quantitative trait loci. Patient drug response, either to different drugs or different doses of the same drug, can vary widely with many different types of responses occurring across different individuals with regard to the same drug at the same dose. As such, the multiresponse nature of data produced by studies that have examined patient drug response has a greater dimensional complexity than case-control studies. Epistasis, the authors argue, may have an important causative role to play in the reason for this wide variety of observable responses to different drugs and/or different doses of the same drug. The researchers compared the performance of a Bayesian neural network to the standard multivariate analysis of variance technique using simulated data for three different two-loci genetic models: (1) a simple additive model of genetic causation involving no epistasis, (2) an additive model that includes a single interaction term to

represent epistasis among the two causative loci, and (3) a model in which the involved genetic loci affect drug response purely through interaction with each other.

For the additive model [71], both methods tested were insignificantly different and had very similar power to detect the causative SNPs. For the additive model with interactions, Bayesian neural networks outperformed MANOVA throughout most parameter combinations of effect size and MAF, and this difference is particularly notable for the larger values of MAF tested. For the purely interactive model, the methods were again insignificantly different with regard to statistical power, but both methods performed well and even achieved a statistical power of 100% for MAFs of 0.3, 0.4, and 0.5. The authors argue, however, that Bayesian neural networks may be the preferable method for many applications due to making less assumptions about the data and the statistical model as well as being much more capable of dealing with new constraints or different data types. To accommodate a new data type, for example, one only needs to incorporate the likelihood function into the output layer of the network. Nothing else about the basic model needs to be modified in order to analyze a new type of data. It is also worth noting that the third genetic model above, in which genetic loci affect the phenotype solely through epistatic interaction without any individual main effects, is of particular academic interest given that previous authors have called attention to the incapability of current algorithms in use to detect such epistatic effects altogether [72], although strong evidence to suggest that such a genetic model is realistic has yet to be produced [72]. This study further demonstrates the utility of neural networks for detecting gene-gene interactions among high-dimensional data (1,000 SNPs used here), and it also demonstrates the ability of neural networks to effectively accomplish investigations of genetic association that involve both epistatic interaction as well as multiple outcomes (as opposed to simply a binary case-control response variable).

Tong et al. used a standard multilayer perceptron neural work with the backpropagation algorithm to analyze epistatic interactions among 96 genes [28]. The genes examined by the authors in [28] were chosen due to their ability as a set to serve as marker genes whose expression data can be used to distinguish between four different types of blue cell tumors that can otherwise be indistinguishable when present in children. This study is unique and noteworthy in that the authors used genetic expression data to predict the expression of other genes. To perform the analysis, the authors represented all but one gene (the "output gene") as input nodes in the neural network, with the output gene represented as an output node. For each individual gene in the dataset of 96 genes, the authors constructed, trained, and analyzed a separate neural network in which that particular gene was represented by the output node. The individual steps of the process the authors used can be viewed in detail in Figure 13, with the specific type of expression data that was used to train, test, and validate the neural network being microarray expression data. The central hypothesis of the study was that the expression level of a gene represented by the output node in a neural network could be predicted based on the expression of the set of genes used as the input nodes, if the set of genes represented by the input and output nodes were able to explain a particular categorical outcome such as disease status or type of blue cell tumor. If the neural networks could be trained to accurately predict the expression level of the output gene then there is a statistically significant correlation between the set of input genes and the output gene. Given that the set of genes can be used as a biomarker to explain a categorical outcome, this statistically significant correlation is suggestive of an epistatic interaction given that the categorical outcome is significantly correlated with a predictable statistical relationship between the output and input genes.

As such, the genes examined by the authors in [28] were specifically chosen due to their ability as a set to serve as marker genes whose expression data can be used to predict the categorical outcome that, in this case, is the type of blue cell tumor. The strength of the weights in each trained and validated neural network represents the strength of the statistical correlation and therefore the strength of the epistatic interaction between genes. A simplified interaction map for the associations analyzed by the authors is presented in Figure 12. The interaction maps identified by the authors serve as an interesting area for further study to determine the phenotypical effect that such genes in each epistatic cluster may have. Investigating the phenotypical significance of each epistatic cluster may provide further insight into the nature of each blue cell tumor which may ultimately lead to more effective treatments. Furthermore, *in vivo* or *in vitro* investigation of the functions and involved pathways of the genes in each epistatic cluster could help to validate the utility of the method the authors used to identify such epistatic interactions. Finally, follow-up work to determine whether or not such correlations in gene expression are truly a result of epistatic interaction affecting trait status, such as investigations involving gene silencing or gene knockout techniques, would also certainly be of relevant scientific interest.

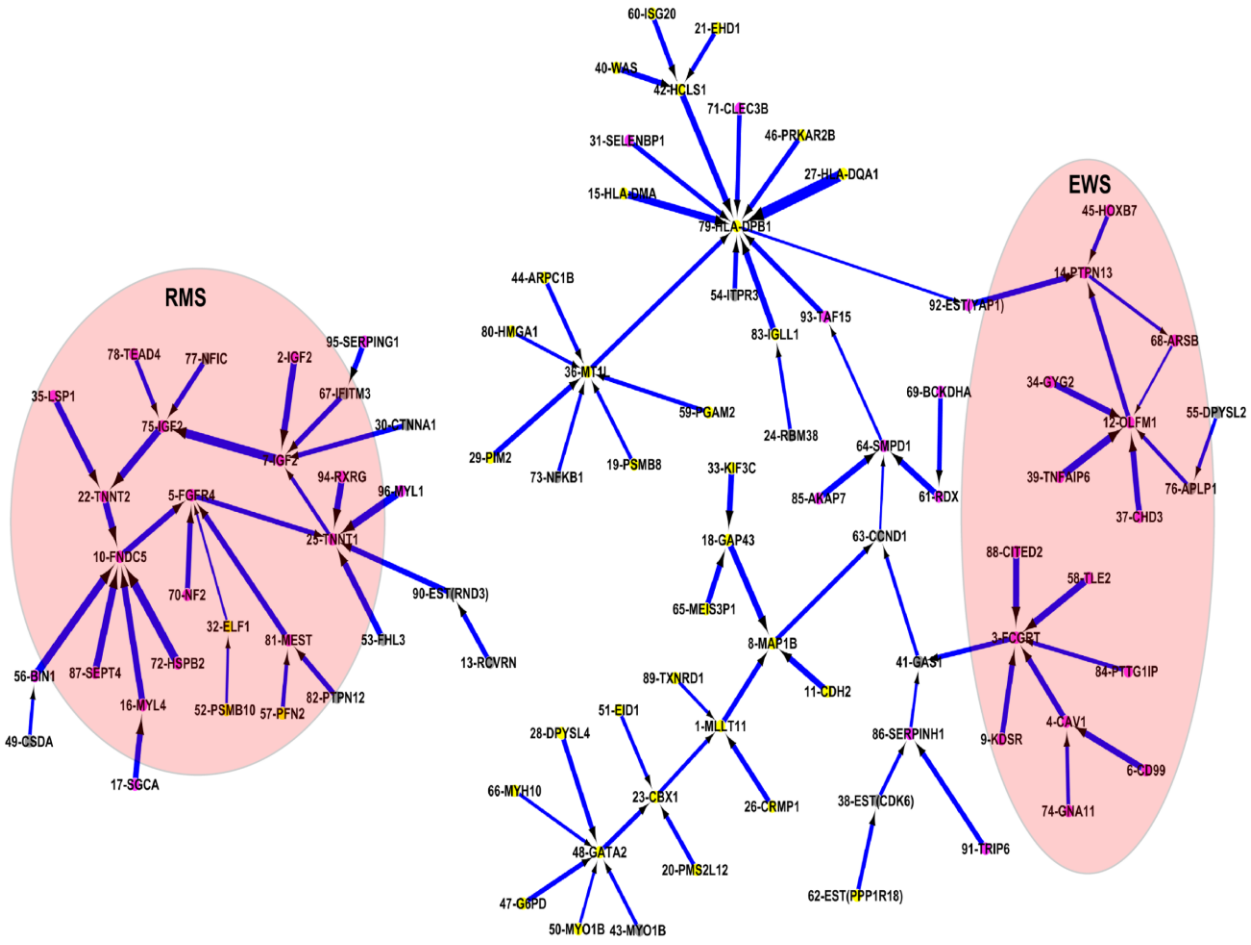

Figure 4 - A simplified interactome network map for the 96 selected genes by ANN network inference algorithm [28]

This interactome network map displays the strongest single association for each of the 96 genes analyzed. The full interactome network map displays the strength of (96*(96-1)) = 9,120 interactions and is thus very difficult to obtain any useful information from. The red nodes are the genes with high expression values in either of the sarcoma cancers (RMS vs EWS). The gray node is the gene with high expression values in more than one cancer groups in which one of these groups is sarcoma cancer. Yellow nodes are the genes with low expression values in both sarcoma cancers. [This figure is reused in accordance with the Creative Commons Attribution (CC BY) license granted by PLOS One that is applicable to this content]

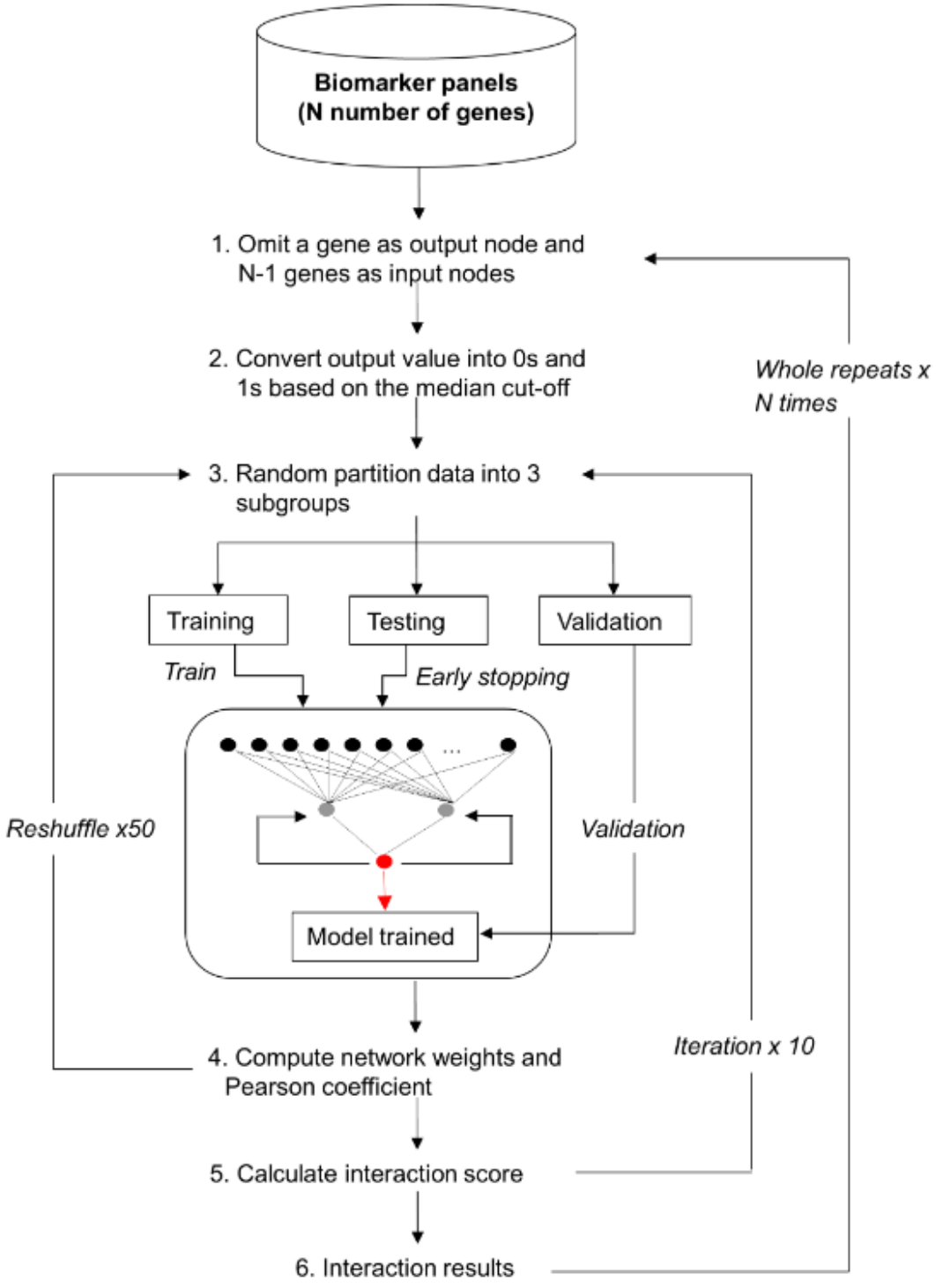
Biomarker panels
(N number of genes)
1. Omit a gene as output node and N-1 genes as input nodes
2. Convert output value into 0s and 1s based on the median cut-off
3. Random partition data into 3 subgroups
Training
Testing
Validation
Train
Early stopping
Model trained
Reshuffle x50
Validation
4. Compute network weights and Pearson coefficient
5. Calculate interaction score
6. Interaction results
Whole repeats x N times
Iteration x 10

Figure 5 - Overview of the interaction algorithm [28]

[This figure is reused in accordance with the Creative Commons Attribution (CC BY) license granted by PLOS One that is applicable to this content]

Neural Networks and Other Challenges in Genetic Association Studies

Neural networks have been used to address many of the challenges that exist in modern genetic association studies such as those cardinal problems that were summarized in Table 1. Heterogeneity has been directly addressed by methods that involve neural network algorithms in several different manners. Cluster analysis, a commonly used unsupervised machine learning algorithm briefly described earlier in this paper, is a commonly employed method of addressing the analytical problems associated with heterogeneity [17][73]. In genetic association studies, the goal of cluster analysis is to create groups of individuals with similar genetic profiles. This is done with the intent of creating case-control groups of individuals wherein the genetic cause of the cases are significantly associated with the same underlying genotype within each respective group. Before using neural networks to perform genetic analyses on a dataset, cluster analysis is often a first step used to create separate groups from the dataset [73]. Afterwards, each group is then analyzed separately with a neural network algorithm used to perform some predictive function with the data such as identification of genetic variants significantly associated with the outcome of interest. This method has been used extensively in the study of various biological interests, from studying genetic association of heterogeneous complex diseases [74]-[77] to successfully diagnosing complex diseases such as Alzheimer's from heterogeneous data such as

readings of regional cerebral blood flow [78]. Neural networks have even been used for the purpose of analyzing the clusters created by cluster analysis, such as for the purpose of identifying those biomarkers that are most important for determining which cluster an individual data point belongs to [79].

Furthermore, variants of neural network algorithms have also been applied to achieve high statistical power even when dealing with markedly heterogeneous data. D'Alessandro et al. used genetic programming to build the architecture of a probabilistic neural network and managed to achieve 100% statistical power to predict the onset of a seizure within 10 minutes from highly heterogeneous electroencephalogram readings [80]. Although [80] is not specifically an example of a genetic association investigation, it nonetheless is relevant for inclusion here given that it serves to demonstrate the ability of neural networks to account for heterogeneity which is, as discussed, a significant problem in genetic association studies. Liang and Kelemen used a hierarchical Bayesian neural network to achieve an accuracy of 88% to correctly classify yeast cell cycles based on highly heterogeneous gene expression data [81]. Ando et al. used a fuzzy neural network system based on majoritarian decision using multiple noninferior models to achieve a statistical power of 91% for predicting patient outcome after chemotherapy based on transcriptional profiling data in spite of the massive heterogeneity associated with the genetics of cancer and drug response [82]. These same authors in [82] later used a fuzzy neural network with microarray data as input to predict the prognosis of patients with diffuse large B-cell lymphoma and achieved an accuracy of 93%, again in spite of the genetic heterogeneity associated with diffuse large B-cell lymphomas [83].

Neural networks have also been used directly to deal with the problem of genuinely causative genes having main effects that are so small as to be statistically insignificant. As discussed, this can happen when a gene's contribution to a phenotype of interest occurs primarily as epistatic interaction with other genes [12], or when the effect of that gene altogether is very small on an individual level (although it may nonetheless actually contribute to the phenotype in an additive manner along with many other genes that also have very small main effects on an individual level [10][11][18]). Many different authors have reported that the use of neural network algorithms allowed them to detect epistatic and/or additive interactions among genes that contribute to a phenotype even when the individual main effects of the interacting genes on that phenotype were statistically insignificant [26][71][84][85][86]. Neural networks have thus proven themselves to be a robust machine learning method that offers greater statistical power to deal with heterogeneity, small effect sizes, and small sample sizes than can be achieved using non-machine learning methods such as logistic regression, particularly when analyzing a large number of predictors with significant interaction [24]-[31][73]-[86].

Neural networks have also been used to address the issues in computational complexity that arise when studying gene-gene interactions (the "curse of dimensionality"), primarily through feature selection processes as is done in genetic programming neural networks [67], grammatical evolution neural networks [87], and probabilistic neural networks such as Bayesian neural networks [68]. Such variants reduce the computational complexity involved by filtering the predictors to include only those that are most critical (as determined by some threshold value) for classification purposes [88][89]. Analyses performed on such methods have demonstrated that equivalent classification accuracy can be achieved through these techniques in spite of the reduction in computational complexity that arises from the resulting feature-trimmed

models [39][67][87][88][89]. Such feature-selection methods are desirable when the aim is to achieve an efficient classification or prediction algorithm when dealing with a large dataset and/or starting with a large number of predictors. This situation is common when studying gene-gene interactions which results in the aforementioned combinatorial explosion of dimensionality. Feature selection can actually improve classification ability by mitigating overfitting problems as well as by allowing a realistic and computationally feasible investigation of interactions among predictors [39][67][88]-[92].

Multifactor Dimensionality Reduction

Multifactor Dimensionality Reduction (MDR) is a technique that was developed to identify significant genes that are affected by non-additive interactions with other genes [93][94]. This is especially important for identifying significant loci for diseases that have missing heritability and/or expression that is significantly influenced by gene-gene interactions [93][94]. MDR, being a non-parametric method like neural networks, has the benefit of not estimating specific parameters and not possessing potentially false assumptions about the interaction between genes [93][94]. The MDR method separates genetic data into a binary outcome variable (high risk or low risk) of all possible multi factor combinations in order to reduce the dimension of predictors from n-dimensions to one-dimension [93][94]. However, the main advantage of MDR analysis is the ability of the newly created multilocus model to identify non-additive interactions between genes [93][94][95].

Although multifactor dimensionality reduction as an analytical technique is not a recent development, it is still fairly common and is still being used extensively to study epistasis

[32][36][93][94][95]. MDR is frequently used in genetic association studies due to its high predictive power in detecting genotype combinations that are significantly associated with a disease of interest [32][36][93][94][95]. However, MDR has the following weaknesses: inability to quantify the effect of genotype combinations on the disease of interest, problems arising from unbalanced data and genetic heterogeneity, and lower predictive power when examining a high number of genotype combinations at numerous loci [36][37][93][94][95]. To help mitigate and account for these weaknesses, various alterations to the MDR method have been developed and proposed. For example, in order to quantify the effect of specific gene-gene interactions on a disease, techniques involving the use of odds ratios and weighted risks have been developed [95][37]. Mahachie et al analyzed a technique called model-based MDR that was developed to account for issues with unbalanced data and genetic heterogeneity in MDR analysis, but concluded that accounting for genetic heterogeneity still remains a significant challenge for MDR and genetic association studies in general [36]. Finally, MDR analysis can be computationally expensive and require a large amount of time due to the multiple iterations that must be performed when the model is being trained and during cross validation. Therefore, techniques such as the three way split method described by Winham et al have been developed that reduce the time required to perform a MDR analysis while not significantly reducing the predictive power of the developed model [33]. A complete summarization of the most recent and highly utilized alterations of MDR is provided in the section titled “Modern Variations to Multifactor Dimensionality Reduction”.

Basic Theory and Methodology

Multifactor dimensionality reduction is a particularly effective method for detecting gene-gene interactions based on the inherent methodology involved [93]. Multifactor Dimensionality Reduction first reduces the dimensionality of multifactor classes from n dimensionality to one dimensionality so that each combination of genotype values are compared [93]. For example, 4 loci that each have 2 possible variants each would give a total of 16 different genotype combinations [93][94]. Cross validation is then performed in order to estimate the prediction error of the multilocus model composed of the optimal genotype combination. Cross validation consists of dividing the data into n number of partitions, with 1 partition acting as the testing data and n-1 partitions acting as the training data, and repeating this process for every possible partitioning of the data [93][94]. Each of the genotype combinations are analyzed for case/control ratios using the training partition of the dataset and the genotype combinations are then separated into a high-risk group if their case/control ratio is higher than a pre-set value, usually 1, or low-risk group of the case/control ratio is below the value [32][93][94][95]. This is performed in each partition cycle and the prediction error of the multilocus model is calculated for each partition cycle using the testing partition of the dataset. The multilocus model, which consists of genotype combinations and their prediction of disease state, is selected based on which multilocus model from all of the partition cycles has the lowest prediction error [93][94]. There are many variations of MDR analysis, which will be discussed below. Figure 14 shows a typical Multifactor Dimensionality Reduction analysis.

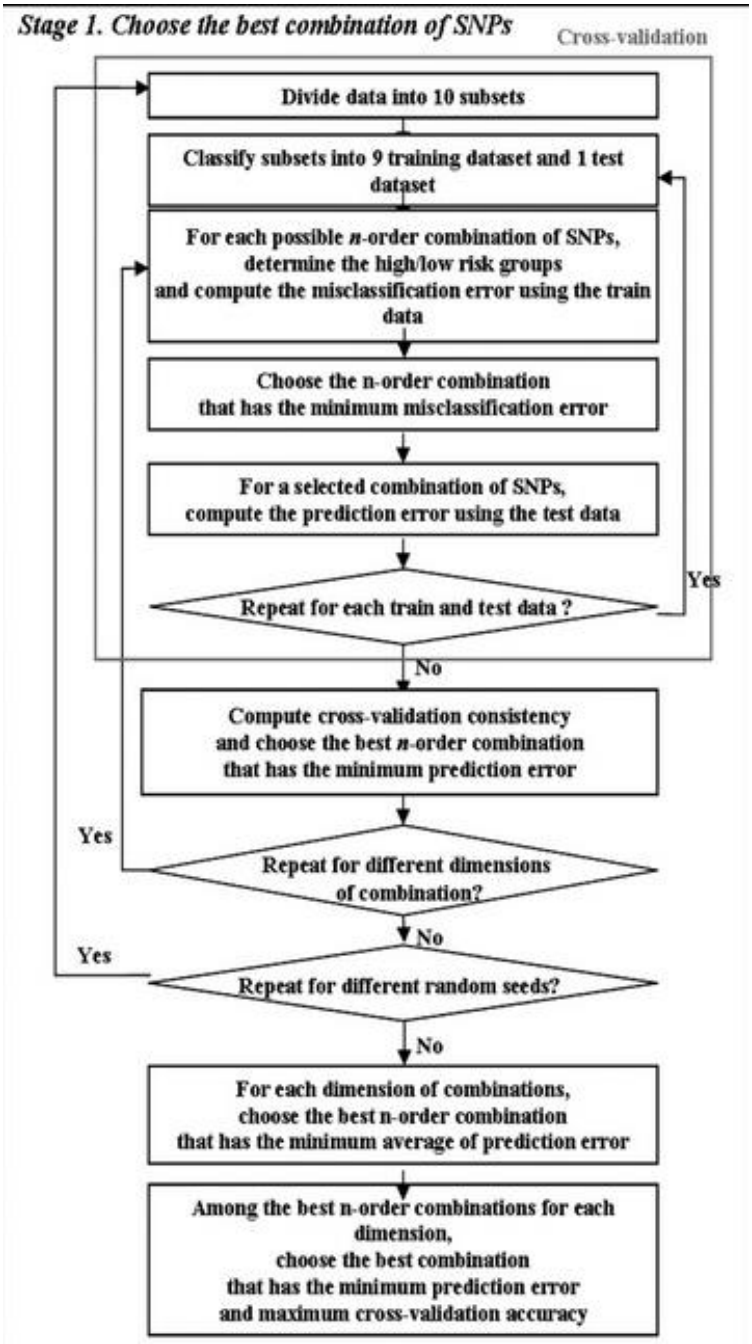

Figure 6 - Multifactor Dimensionality Reduction Workflow [95]

[Permission to reuse this figure was obtained from Copyright Clearance Center's RightsLink service. Oxford University Press has granted me permission to reuse this figure via License # 3623410665013]

Modern Variations to Multifactor Dimensionality Reduction

One major drawback to using Multifactor Dimensionality Reduction in genetic association studies that seek to identify gene-gene interactions between more than 2 loci is the high false positive and false negative error rates that can exist due to the method's simple comparison of case and control ratios in those instances where the case and control ratio for a genotype is close to the ratio for the entire dataset [93][94][95]. In addition, Multifactor Dimensionality Reduction provides an assessment of high risk or low risk to specific genotypes, but does not provide specific quantitative data for diseases risk [93][94][95]. To address these concerns, Chung et al developed a modified form of MDR called Odds Ratio Based Multifactor Dimensionality Reduction (OR MDR) [95]. In OR MDR, odds ratios are calculated for each of the genotype combinations [95]. When the odds ratios are analyzed, it is possible to identify genotypes that have a positive association with the disease (odds ratio >1) and genotypes that have a negative association with the disease (odds ratio <1) [95]. In addition, confidence intervals for the odds ratios can also be calculated by analyzing the distribution of the odds ratios for the genotypes. To get this confidence interval, a bootstrapping technique is used that resamples a random genotype combination's case and control and repeating this process 100,000 times [95]. This distribution is used to obtain the confidence interval for the odds ratio, which in turn will allow for a comparison of the significance of various genotype combinations that were found in the analysis [95]. This is an advantage when compared to regular MDR methods as the use of odds ratios allows for quantitative comparison of the ratios for various genotype

combinations on disease risk [93][94][95]. A diagram of the workflow for the odds ratio MDR analysis can be found in Figure 15.

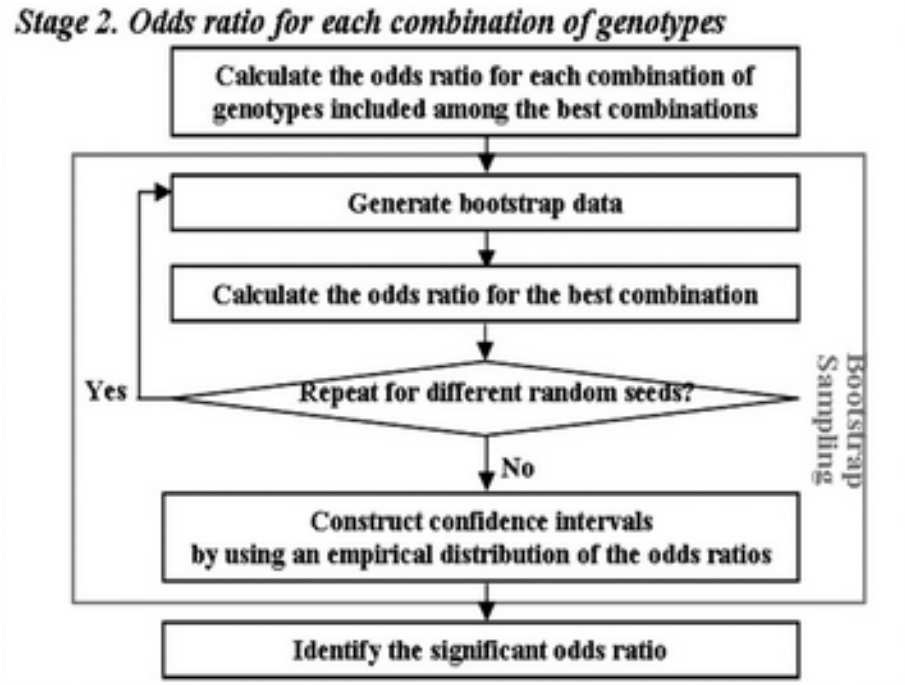

Figure 7 - Odds Ratio MDR Workflow [95]

[Permission to reuse this figure was obtained from Copyright Clearance Center's RightsLink service. Oxford University Press has granted me permission to reuse this figure via License # 3623410665013]

In addition to the OR MDR method, other methods have been developed to quantitatively measure the effect of gene-gene interactions on disease risk [95][37]. A method described by Luo et al called weighted risk-score multifactor dimensionality reduction (WRSMDR) has the advantage of negating the effects of imbalanced datasets on the method's accuracy and providing

a quantitative assessment of the effect of gene-gene interaction on the disease of interest [37]. The WRSMDR method calculates the Bayesian posterior probability of different SNP combinations and uses this as a quantitative measure to analyze the various genotype combination's effect on the disease of interest [37]. Luo et al performed an analysis using MDR and WRSMDR on a simulated dataset and found that there was no difference in predictive power between MDR and WRSMDR for two-locus gene interactions [37]. However, the positive identification rates for three- and four-locus gene interactions was significantly higher using the WRSMDR method than with the MDR method [37]. The ability to utilize unbalance datasets, higher identification rates for gene interactions involving three or more locus, and the ability to quantitatively assess genotype combinations with respect to a disease of interest are significant advantages to the WRSMDR method [37]. The main drawback to utilizing WRSMDR to MDR is the slightly higher amount of time required to perform an WRSMDR analysis due to the additional calculations performed to calculate weighted risk scores [37]. However, the WRSMDR method is significantly more effective than a regular MDR analysis for multilocus models utilizing three or more loci [37]. The time and computational power required for an WRSMDR analysis can be reduced by filtering the dataset to remove non-significant variants prior to analysis [37]. A comparison of the power of MDR and WRSMDR in two, three, and four locus models for balanced and unbalanced data can be found in Tables 8 and 9.

Evaluation Indicator	Two-Locus		Three-Locus		Four-Locus	
	WRSMDR	MDR	WRSMDR	MDR	WRSMDR	MDR
Specific Detection Rate	0.87	0.83	0.74	0.83	0.92	0.46
Detection Rate	1	1	1	1	0.97	0.56
Error Rate	0	0	0	0	0.01	0.44
No Detection Rate	0	0	0	0	0.02	0

Table 8 - Power Comparison of MDR and WRSMDR in Balanced Datasets [37]

Specific detection rate refers to the proportion of simulated datasets where the true model was detected as the best model. The detection rate is the proportion of simulated datasets where the overall model included the true model as part of its multi-locus model. The error rate is the proportion of simulated datasets where the true models were not included in the best overall model [37]. [This figure is reused in accordance with the open access license granted by MDPI that is applicable to this content]

Evaluation Indicator	Two-Locus		Three-Locus		Four-Locus	
	WRSMDR	MDR	WRSMDR	MDR	WRSMDR	MDR
Specific Detection Rate	0.96	0.61	0.57	0.66	0.94	0.68
Detection Rate	1	0.81	0.85	0.85	0.98	0.79
Error Rate	0	0.19	0.03	0.15	0.01	0.21
No Detection Rate	0	0	0.12	0	0.01	0

Table 9 - Power Comparison of MDR and WRSMDR in Unbalanced Datasets [37]

Specific detection rate refers to the proportion of simulated datasets where the true model was detected as the best model. The detection rate is the proportion of simulated datasets where the overall model included the true model as part of its multi-locus model. The error rate is the proportion of simulated

datasets where the true models were not included in the best overall model [37]. [This figure is reused in accordance with the open access license granted by MDPI that is applicable to this content]

Another concern when using Multifactor Dimensionality Reduction is that the inclusion of non-statistically significant genotype combinations in the analysis can decrease the success rate of the analysis [32]. Gui et al have suggested a modified extension of MDR called Robust Multifactor Dimensionality Reduction (RMDR) [32]. In this approach, a Fisher's Exact Test is carried out in order to determine if genotype combinations should be included in the analysis [32]. The rationale behind this extension is that the inclusion of genotype combinations with a case to control ratio similar or identical to the data's overall case to control ratio will unnecessarily decrease the success rate of the analysis and an objective statistical analysis should be performed in order to screen out genotype combinations [32]. Gui et al performed an analysis on two simulated studies and found that analysis with the RMDR method had a significantly higher success rate than a simple MDR analysis [32]. The main drawback for RMDR is that calculation of the Fisher's Exact Tests results in the method being substantially more computationally expensive as compared to regular MDR [32]. However, the authors found that this time could be significantly reduced with the use of a stored "lookup table" for common calculations and that the use of these lookup tables significantly lowered the time it took to run a RMDR analysis [32]. The use of RMDR was a significant development as it allowed for an increase in the predictive power of the developed model with a tradeoff of a higher calculation time. A comparison of the difference in success rate between MDR and RMDR can be found in Table 10. The authors calculated the success rate as the proportion of trials for which MDR or

RMDR correctly identified the two functional SNPs out of each set of 500 balanced and imbalanced datasets.

Heritability (H)	*T=1, n=400*			*T=0.5, n=600*			*T=0.33, n=800*		
	MDR	**RMDR-0.05**	**RMDR-0.1**	**MDR**	**RMDR-0.05**	**RMDR-0.1**	**MDR**	**RMDR-0.05**	**RMDR-0.1**
0.01	0.034	0.076	0.064	0.062	0.084	0.09	0.092	0.186	0.184
0.025	0.1	0.242	0.224	0.236	0.468	0.448	0.342	0.646	0.648
0.05	0.35	0.642	0.64	0.678	0.946	0.95	0.824	0.986	0.99
0.1	0.984	1	1	0.998	1	1	1	1	1

Table 10 - Comparison of Success Rates Between MDR and RMDR Using Two-Locus Model Simulation Studies [32]

The left hand column represents heritability and the individual cells are the success rates. The columns RMDR-0.05 and RMDR-0.1 represent analysis performed with significance levels of 0.05 and 0.1 respectively. For the main column labels, T represents the case/control ratio and n represents the sample size [32]. [Permission to reuse this figure was obtained from Copyright Clearance Center's RightsLink service. John Wiley and Sons, the publisher, has granted me permission to reuse this figure via License # 3623420011246]

One drawback of multifactor dimensionality reduction is the amount of time required to perform the analysis, particularly the cross validation [33]. In an effort to lower the calculation time required for an MDR analysis, Winham et al performed a study on simulated data using the three way split method in place of cross validation [33]. The three way split method splits up the

data into 3 different sections (Figure 16). The first section is a training section to develop the model, the second section is a testing section to refine the model and the third section is a validation section to quantify the predictive power of the model [33]. In addition, to decrease the time of the analysis and the number of false positives, a logistic regression pruning was performed using a pre selected p-value (usually 0.05) in order to remove insignificant variables from the model [33]. From this analysis the authors found that a MDR analysis performed using the three way split instead of cross validation had comparable predictive power when pruning techniques were used [33]. However, the authors found that the predictive power of the analysis using the three way split without pruning suffered as compared to a normal MDR analysis with cross validation [33]. This paper was significant as it provided a new method for decreasing the computational time to perform a MDR analysis while not significantly lowering the predictive power of the model [33]. This paper also identified the significance of using a pruning process and the impact it can have on the predictive power of a model and it's false positive identification rate. A summary of selected variants of the MDR method, including all those discussed here in this section, is presented in Table 11.

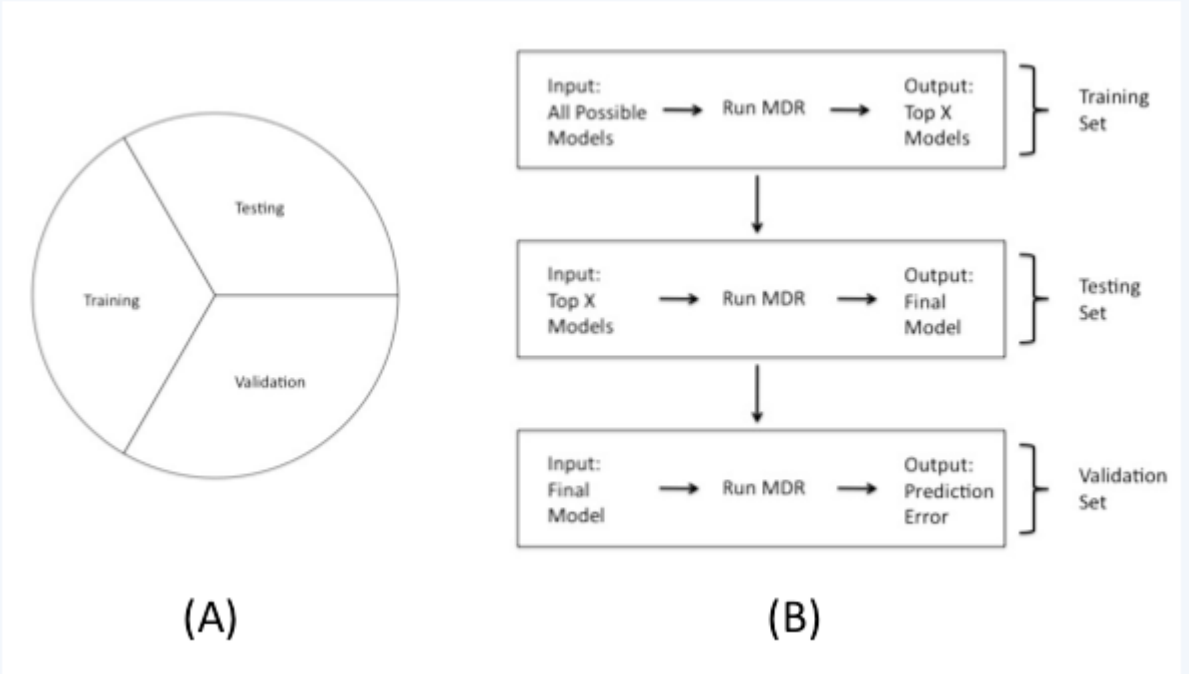

Figure 8 - Visual Representation of the 3 Way Split MDR Method [33]

A) Three-way split of the full sample data. B) - Explanation of how the three-way split is incorporated into the MDR method. First the sample data is randomly split into 3 intervals with representative numbers of cases and controls in each interval. MDR is performed on each of the three splits with all possible models considered in the training set, the top x models considered in the testing set, and the final model considered in the validation set for each k [33]. [This figure is reused in accordance with the Creative Commons Attribution (CC BY) license granted by BMC Bioinformatics that is applicable to this content]

Table 11	Attributes
Odds Ratio MDR (OR MDR) [95]	Provides a quantitative assessment of the association of genotype combinations with a disease of interest through the comparison of odds ratios with confidence intervals. This allows for the classification of genotype combinations as either having a positive or negative association with the disease.
Weighted Risk-Score MDR (WRSMDR) [37]	Compensates for the effect of imbalanced datasets on the model's accuracy and provides another form of quantitative assessment of the gene-gene interactions identified for with respect to the disease of interest. In addition, this method has a higher positive identification rates for analysis of gene interactions between 3 or more loci than regular MDR analysis. However, this method has the disadvantage of a larger computational time than regular MDR analysis.
Robust MDR (RMDR) [32]	Allows for the screening of datasets through a Fisher's exact test to remove non-statistically significant genotype combinations in order to increase the success rate of the analysis.
Three Way Split MDR (3WSMDR) [33]	Significantly lowers the amount of time necessary to perform an MDR analysis. However, it has the drawback of reducing the predictive power of the developed model. Therefore, it is commonly used with a pruning technique in order to achieve a higher predictive power.
Generalized MDR (GMDR) [96]	Utilizes pruning techniques to significantly increase the predictive power of the analysis. Pruning techniques can be included with other MDR methods such as the three way split method.
Model-Based MDR (MB-MDR) [38]	Compensates for issues resulting from unbalanced data and genetic heterogeneity in MDR analysis. In addition, this method has also been found to have a slightly higher predictive power than regular MDR.

Table 11 - Selected Multifactor Dimensionality Reduction Variants with Accompanying Descriptions

Multifactor Dimensionality Reduction and Gene-Gene Interactions

There have been many recent studies that have used Multifactor Dimensionality Reduction in genetic association studies. One recent genetic association analysis performed by Liu et al used MDR to identify genotype combinations that significantly influence the development of obesity in Chinese children [96]. Their analysis utilized a generalized multifactor dimensionality reduction (GMDR) analysis with the addition of logistic regression pruning in order to increase the predictive power of the analysis [96]. This study selected gene candidates that were in the lipid metabolism pathway as there was prior literature that suggested the potential influence of this pathway and in particular specific genes in the pathway [96]. The study utilized 15 different variants from 5 different loci. This study was the first to perform an in depth analysis using MDR on these genes in order to identify what possible gene-gene interactions existed that affected the development of obesity. Their study found that the interaction of genetic variants in the INSIG-SCAP-SREBP pathway were associated with obesity [96]. Table 12 provides an overview of the 9 high-risk genotype combinations that were found in the study. This was significant as prior literature had identified some of these genes individually as being associated with obesity, such as INSIG1 [96]. However, unlike previous studies, the use of MDR in the analysis of gene-gene interactions found an entire set of genetic variants that were interacting and associated with the onset of obesity [96].

rs9308672/rs12487736/rs1883205	Obese (n = 693)		Nonobese (n = 1311)		Percentage ratio (obese/nonobese)
	Number	Percentage	Number	Percentage	
CC/GG/CC	15	2.16	23	1.75	1.23
CC/AG/CT	44	6.35	63	4.81	1.32
CC/GG/TT	8	1.15	10	0.76	1.51
CT/AA/CC	39	5.63	42	3.20	1.76
CT/AG/CC	55	7.94	90	6.86	1.16
CT/GG/CC	37	5.34	45	3.43	1.56
CT/AG/TT	37	5.34	55	4.20	1.27
TT/GG/CT	37	5.34	37	2.82	1.89
TT/GG/TT	13	1.88	10	0.76	2.46
Total	285	41.13	375	28.60	1.44

Table 12 - Frequency differences of 9 high-risk genotype combinations of INSIG-SCAP-SREBP pathway between obese and nonobese children [96]

This table provides and overview of the 9 high-risk genotype combinations that were identified in this study. [This figure is reused in accordance with the open access permissions granted by BioMed Research International that is applicable to this content]

Another recent example of the use of Multifactor Dimensional Analysis in epistasis is the study by Kebir et al [97]. This study sought to identify key genes that are associated with schizophrenia [97]. The genes that were analyzed in this pathway were all Histone deacetylases (HDAC) as previous literature had identified the possible association of some genes in this pathway with schizophrenia [97]. In particular, this study analyzed 601 different variants across 10 loci. The analysis performed in this study was a generalized multifactor dimensionality

reduction without any pruning or filtering [97]. This analysis found that the loci HDAC9, HDAC10 and HDAC11, had significant interaction [97]. This study further demonstrates the significance of MDR analysis on the study of gene-gene interaction as it was able to identify a gene interaction relationship that was previously undiscovered and new genes that potentially interact to influence the development of schizophrenia [97].

In a recent study by Kumar et al additional applications of MDR for detecting epistasis were identified [98]. This study used MDR to analyze the interaction of various genetic combinations at seven loci with respect to essential hypertension, systolic and diastolic blood pressure, mean arterial blood pressure and body mass index [98]. Expression of the genes FTO and GNB3 were found in previous studies to be associated with essential hypertension in previous studies, however, this study sought to analyze the interaction of FTO and GNB3 variants with other loci and produced disease models for analysis with 2-7 loci involved [98]. The study found that a three locus model comprised of the polymorphisms *FTO* rs8050136C/A, *GNB3* rs1129649T/C, and rs5443C/T was the best disease prediction model. The results of their analysis is presented in Table 13. This is significant as this study further demonstrates the use of MDR for identifying gene-gene interactions and specific genotype combinations that are associated with the disease of interest [98]. This study shows the potential use of MDR to identify specific genotype combinations that potentially cause or have an effect on certain diseases. In addition, this study showed how MDR analysis can be used in a follow-up genetic association study that analyzes genotype combinations in significant genes that were identified in previous studies to both confirm their significance for the disease of interest and identify additional associations between previously identified genes.

FTO+GNB3	Best models	TB	TA	CVC	P value	OR(95% CI)
2L	rs1129649T/C rs5443C/T	0.62	0.59	8/10	0.067	2.1(0.9–4.8)
3L†	rs8050136C/A rs1129649T/C rs5443C/T	0.63	0.62	9/10	0.0005	3.9(1.8–8.5)
4L	rs9930506A/G rs9932754C/T rs1129649T/C rs5443C/T	0.65	0.61	4/10	0.006	3.0(1.4–6.5)
5L	rs9939609T/A rs9930506A/G rs9932754C/T rs1129649T/C rs5443C/T	0.64	0.61	8/10	0.0002	4.9(2.0–11.7)
6L	rs8050136C/A rs9939609T/A rs9926289G/A rs9932754C/T rs1129649T/C rs5443C/T	0.63	0.59	6/10	0.0165	3.0(1.2–7.7)
7L	rs8050136C/A rs9939609T/A rs9926289G/A rs9930506A/G rs9932754C/T rs1129649T/C rs5443C/T	0.61	0.58	10/10	0.012	3.6(1.3–10.4)

Table 13 - Interaction between genotypes of *FTO* and *GNB3* using MDR [98]

This table shows the best multilocus models for 2- to 7-locus models (left hand column). The 3 locus model was found to be the best model. TB represents the test balance accuracy, TA training accuracy, and CVC cross validation consistency. P-values were calculated by permuting the case and controls 100 times [98]. [This figure is reused in accordance with the Creative Commons Attribution (CC BY) license granted by PLOS One that is applicable to this content]

Lai et al performed a study that further illustrates in detail the potential for performing a multifactor dimensionality reduction analysis to identify genetic associations between genes of interest identified in previous studies that were associated with a disease [99]. This study focused on 492 variants for three loci that were associated with bipolar disorder [99]. MDR was used to perform an epistatic interaction analysis on the genetic variants of these genes [99]. The MDR analysis found that there were gene-gene interactions with specific variants of the RORA and RORB genes that were associated with bipolar disorder [99]. In addition, the authors also found that there was a four way interaction with NR1D1, RORA, and RORB [99]. A table summarizing

the MDR results can be found in Table 14. This is significant because it confirmed previous studies suggesting the significance of these genes and provided evidence that variants of these genes were interacting to affect the development of bipolar disorder [99]. In addition to identifying these new gene interactions and the significance of certain gene variants, the study also found that the genotype combinations identified involved genes that were members of the circadian pathway, which was previously associated with bipolar disorder [99]. In particular, this study demonstrates the use of MDR to analyze genes that were previously associated with a disease of interest in an effort to identify gene-gene interactions that occur between these genes. In addition, the identification of critical genotype combinations can also identify specific metabolic pathways that are associated with the disease of interest, as this study demonstrated, and help researchers identify specific genes or pathways that should be targeted in future research.

Gene1	SNP1	Gene2	SNP2	Gene3	SNP3	Gene4	SNP4	Training Bal. Acc. (%)	Testing Bal. Acc. (%)	Cross validation consistency
RORB	rs1327836							0.5778	0.5200	7/10
RORB	rs1327836	*NR1D1*	rs12941497					0.5971	0.5239	3/10
RORB	rs1327836	*RORB*	rs3750420	*RORA*	rs809736			0.6355	0.4807	2/10
RORB	rs3750420	*RORB*	rs11144047	*RORA*	rs4774388	*NR1D1*	rs2071427	0.7016	0.5325	8/10

Table 14 - Summaries of multifactor dimension reduction gene-gene interaction results [99]

This table shows the multilocus MDR models. The columns on the right represent the training balance accuracy, testing balance accuracy, and cross validation consistency. The four way gene-gene interaction model was identified as the optimal model [99]. [This figure is reused in accordance with the Creative Commons Attribution (CC BY) license granted by PLOS One that is applicable to this content]

MDR and Other Challenges in Genetic Association Studies

As was demonstrated to be true for neural networks, multifactor dimensionality reduction has also been used extensively specifically to deal with the challenges in genetic association studies that were summarized in Table 1. Genetic heterogeneity is a major challenge in MDR studies and several authors have published studies demonstrating a very low statistical power associated with the MDR method when it's used to analyze datasets that involve genetic heterogeneity [32]-[36]. However, Edwards et al. have challenged the definition of "statistical power" and suggest that a more relaxed definition is more informative with regard to the utility of MDR when dealing with heterogeneous data [100]. Edwards et al. used simulation data and modeled genetic heterogeneity using two separate two-locus epistatic disease models wherein each model independently led to trait status. Using the strictest definition of statistical power in which both two-locus models must be identified as causative, MDR did not achieve greater than 10% statistical power in any simulation scenario tested [100]. However, if the definition of statistical power included the correct discovery of *either* epistatic model or *any* causative locus from either model, power increased dramatically to 100% in most scenarios tested with the lowest value being 94% which occurred when 10,000 loci were considered as predictors in the model [100]. Interestingly, a different group of authors, Polonikov et al., actually used MDR directly to identify gender-based genetic heterogeneity by identifying genotype combinations that are significantly associated with asthma and then afterwards stratifying the results based on gender [101]. As such, the usefulness of even classical MDR when it comes to dealing with

genetic heterogeneity is noteworthy even if the method fails to achieve high statistical power under the strictest definition of the term.

Additionally, variations to the MDR method do exist that have been successfully employed to increase the statistical power associated with MDR when dealing with heterogeneous data. Cattaert et al performed a study comparing the performance of classical MDR with Model-Based MDR (MB-MDR), a recently developed variant, when dealing with a generated dataset that involves significant heterogeneity [38]. This study used simulation data for a four-locus interaction in the presence of significant heterogeneity and found that MDR had a power of only 8% to identify all four of the loci involved in the epistatic interaction, whereas MB-MDR had a power of 80% [38]. The same authors in [38] published an additional study comparing the statistical power of MB-MDR to MDR by using each method to analyze another generated heterogeneous dataset [102]. Comparing over six different two-locus disease models, the authors report that MB-MDR had a two- to ten-fold increase in statistical power over MDR for all but two of the disease models wherein MDR actually outperformed MB-MDR (0% for MB-MDR compared to 2% for MDR in one model, and 5% for MB-MDR compared to 9% MDR in the other model) [102]. Mei et al. proposed a variant method of MDR called MDR-Phenomics that integrates phenotypic covariates in order to account for genetic heterogeneity [103]. The authors achieved a two- to three-fold increase in statistical power (up to 99.6%) as compared to other MDR variants to detect all simulated causative disease models as significant when the analysis was performed on a generated heterogeneous dataset [103]. As can be seen, the machine learning method that is MDR is very dynamic in that it can be modified extensively and combined with other algorithms in order to add functionality and increase the statistical power that it can achieve when dealing with highly heterogeneous data.

Modifications to the basic MDR algorithm also exist to mitigate the issues in computational complexity that arise when analyzing high-dimensional data. One variation of MDR that has the capability to reduce the computational complexity of the calculations while also actually increasing the predictive power of the developed model is Robust MDR (RMDR) [32]. As discussed earlier in this paper, RMDR performs a Fisher's exact test to determine if a genotype combination should be included in the final model. The study presented earlier in this paper by Gui et al performed a comparison of RMDR and MDR. This study found that RMDR had a statistically significant higher success rate than MDR (see Table 10) and a faster computational time when pre-calculated "lookup" tables were used during the Fisher's test [32]. However, the computational time required for RMDR was higher than MDR when a large (3+) number of loci were analyzed. Therefore, to decrease the computational complexity of the analysis when 3+ loci were included as predictors in the model, the use of pre-calculated "lookup tables" for the Fisher's test were utilized. This was found to significantly decrease the calculation time required for an RMDR analysis to that of a regular MDR analysis [32].

However, other methods do exist to reduce the computational complexity involved in an MDR analysis that do not impose such a low limit on the number of predictors that can be used in order for the benefit to be realized. One method involves using filtration methods such as TuRF, SURF, or ReliefF to pre-filter noisy SNPs that are determined to be unlikely to interact or significantly increase the predictive power of the model [104]. This method can involve global statistical inference tests to allow for the comparison of p-values for each of the potential predictors in the model so that any predictor with a p-value greater than some user-defined threshold is filtered out from inclusion in the model. Such pre-filtering significantly decreases the computational complexity involved in the MDR analysis by reducing the number of

predictors under consideration. The computational complexity of a MDR analysis can also be significantly reduced using the 3-way split method in conjunction with pruning [32]. The 3-Way Split technique involves splitting the data into three groups as normal: training, testing, and validation. However, a pruning technique using a regression based statistical model is performed in place of cross validation. This results in a significantly reduced computational complexity for the analysis and therefore greatly lowers the computational time and resources necessary [33]. Finally, generalized MDR (GMDR) is another machine learning technique that can be utilized to account for computational complexity [96][105][106]. The GMDR method also utilizes pruning techniques in place of cross-validation in order to decrease the computational complexity of the analysis. In addition to computational complexity issues, another issue previously mentioned for machine learning techniques and their application in genetic association studies is inadequate sample size. This can be partially compensated for using the model-based MDR strategy that was previously discussed [36][37][38]. This method increases the statistical power of MDR when dealing with unbalanced data which can be a common issue in analyses that involve small sample sizes [107]. These useful modifications show that MDR is adaptable to accommodate for issues with computational complexity as well as unbalanced datasets that are often associated with small sample sizes.

MDR is also well known to have the ability to detect interactions among even those genes that demonstrate a total absence of statistically significant main effects. This is a consequence of the very nature of the MDR algorithm which performs an exhaustive search over every possible genotype combination among the included predictors with no consideration for the significance of any one particular gene's main effect or any particular type of interaction. Numerous examples exist in the literature that use simulation data to demonstrate the ability of

MDR to detect interactions among genes even in the total absence of main effects, and even the ability to detect causative genes that may interact additively but with a statistically insignificant main effect on an individual level [108]-[115]. By its very nature, MDR is a particularly effective method for detecting interactions among genes with statistically insignificant main effects on an individual, but it also does not naturally allow for the distinguishment between epistatic and additive interactions among genes and instead only allows for the explicit identification of generalized interaction among predictors [116][117].

As demonstrated throughout this section on MDR, there are many challenges that arise when performing genetic association studies that MDR has the ability to directly address. MDR inherently addresses certain issues by its very nature such as that of undetectable main effects among predictors, and there also exists many modifications to the traditional MDR method that compensate for many of the additional challenges that are faced in genetic association studies. The versatility of the MDR method and machine learning techniques in general are further proof of their usefulness in genetic association studies as they can be easily adapted to the many challenges that investigators face.

CONCLUSION

Two major machine learning methods, neural networks and multifactor dimensionality, were thoroughly explored (summaries of selected studies provided in Tables 15 and 16) and a fundamental overview was provided for several other machine learning methods. Such discussion took place within the context of and in a manner relevant to an overarching theme that is the utility of machine learning in the modern state of genetic association studies. The missing heritability problem has garnered significant academic interest in recent years and attempts to find the missing heritability for complex diseases have been a dominant force in genetic association studies. Gene-gene interactions have been proposed as a major contributor to this missing heritability, although it is difficulty to take such interactions into account when analyzing genotype-phenotype models with standard methods. As discussed, such difficulty results from the involved curse of dimensionality that arises when studying interactions among many predictors. Another obstacle to identifying the missing heritability is inadequate statistical power to detect causative genes that have small main effects on an individual level even though they may have a large effect when considered collectively due to interactions among them. Other difficulties, such as genetic heterogeneity, further complicate genetic association studies and may also be a significant contributor to the missing heritability problem at large.

As explicitly described in this paper, machine learning methods such as neural networks, multifactor dimensionality reduction, cluster analysis, random forests, and many others can be utilized to address some of the primary challenges that exist in genetic association studies. Along with the large number of major machine learning methods available to address these challenges, many variants exist for each of these major methods that add functionality to the basic algorithm.

As such, a large plethora of machine learning techniques exist to address a wide variety of challenges and circumstances, and the utility of machine learning in genetic association studies is significant. As computational power continues to expand with evolving technology, some of the issues that arise with computational complexity in GWAS will certainly be mitigated. However, access to greater computational power will likely result in the development of more complicated statistical models (e.g. a greater number of predictors under consideration) given that such models can then be feasibly analyzed where before we lacked the computational ability to. The increase in the complexity of the statistical models under investigation will continue to necessitate machine learning methods that address the challenge of computational complexity if advances in computational power lag behind that which is necessary to thoroughly analyze the increasingly complex statistical models being investigated. Such an effect is exemplified by the relatively recent transition from single-locus association studies to multi-locus association studies that followed the advent of genome-wide sequencing technologies and increasing computational power.

Additionally, our ability to account for confounding issues such as heterogeneity and small effect sizes will not inherently become enhanced solely as a result of access to greater computational power. Such issues require continued advances in computer science and algorithm engineering in order for us to improve our ability to address these challenges in our investigations and statistical models. Advances in computing technology will also directly enhance our ability to use machine learning algorithms by providing the algorithms in use with more computational resources (as exemplified by the aforementioned random jungle variant of the random forest method) and mitigating some of the issues in scalability seen in machine learning algorithms such as multifactor dimensionality reduction. Machine learning has proven

itself to be a valuable tool for genetic analysis, particularly with regard to GWAS, and the utility that machine learning algorithms provide to the field of genetic association studies continues to increase with ongoing advancements in machine learning algorithms rapidly being made.

Summary of Selected Neural Network Studies	
Study Reference Number	Summary
70	Gunther et. al utilized a standard feedforward multilayer perceptron neural network to predict disease penetrance for several different gene-gene interaction disease models. One of the reasons this study was notable is that it demonstrated the ability of even the most basic neural network algorithm to accurately account for multiple kinds of gene-gene interactions, such as additive, multiplicative, epistatic, and mixed.
68	Beam et. al proposed a neural network variant called a "Bayesian neural network" that takes advantage of Bayesian probability theory to build its architecture and perform feature selection via the statistical quantification of variable importance. They report statistical power in the face of genetic interaction that is comparable if not superior to other top-performing methods such as MDR which is notable given that Bayesian neural networks scale very well to high-dimensional datasets.
71	Beam et. al performed a gene interaction analysis using Bayesian Neural Networks. This study further demonstrated the utility of Bayesian neural networks for detecting and accounting for gene-gene interactions in high-dimensional data (this study used 1000 SNPs as predictors). This study also demonstrated the utility of Bayesian neural networks for investigations of gene-gene interactions in a model that involves multiple possible outcomes.
28	Tong et. al performed a gene interaction analysis utilizing a standard multilayer perception neural network with a backpropagation algorithm. The central hypothesis of the study was that the expression level of a gene represented by the output node in a neural network could be predicted based on the expression of the set of genes used as the input nodes, if the set of genes used as the input nodes were able to explain a particular categorical outcome such as disease status or type of blue cell tumor. This study is significant in that it utilized neural networks to investigate gene-gene interactions by using the expression data of a set of genes as input to predict the expression of a gene represented by the output node.
81	Liang and Keleman proposed and utilized a Hierarchical Bayesian Neural Network model to assess the temporal patterns of gene expression in yeast cell cycles. Their analysis was significant as it was able to classify yeast cell cycles with an accuracy of 88% in spite of the massive heterogeneity associated with the dataset.
82	Ando et. al utilized a fuzzy neural network system that was based on majoritarian decision using multiple noninferior models. The authors used this approach to compensate for the challenges arising from data heterogeneity. Using gene transcription data, the authors were able to obtain a statistical power of 91% for predicting patient outcome after treatment with chemotherapy for cancer.

83	Ando et. al performed an additional study utilizing fuzzy neural networks on data with a high degree of heterogeneity. This analysis found that the fuzzy neural network was very useful for addressing challenges arising from data heterogeneity as they were able to predict with an accuracy of 93% the prognosis of patients with diffuse large B-cell lymphoma.
85	Ritchie et. al utilized a genetic programming neural network to detect genetic interactions in simulated data wherein the predictors had minimal to nonexistent main effects. GPNN demonstrated significantly superior performance (achieving a statistical power up to 100%) to detect causative genes, particularly in the epistatic models, compared to stepwise logistic regression, explicit logistic regression, and CART. This study demonstrates the utility of neural networks for detecting gene-gene interactions when the main effects themselves are statistically insignificant.

Table 15 – Summary of Selected Neural Network Studies Included in this Review

Summary of Selected MDR Studies	
Study Reference Number	Summary
96	Liu et. al used a MDR variant called Generalized MDR with the addition of logistic regression pruning to identify signficant genotype combinations that are associated with obesity. This study demonstrates the effectiveness of generalized MDR for identifying gene-gene interactions in genetic association studies, and is particularly interesting given that the authors used MDR to confirm the importance of the lipid metabolism pathway in the development of obesity.
97	Kebir et. al used MDR analyzed 601 variants across 10 loci to identify key genotype combinations that are associated with schizophrenia. This study was notable as it identified previously unknown gene interactions that were associated with schizophrenia.
98	Kumar et. al utilized MDR to analyze genotype combinations at seven loci with respect to hypertension, systolic and diastolic blood pressure, mean arterial blood pressure and body mass index. This study was significant as it studied genotype combinations for genes that were previously identified as being associated with the disease of interest and identified previously unknown interactions between these genes.
99	Lai et. al performed an epistatic interaction analysis on 492 genetic variants for three loci using MDR. The three loci studied in this analysis were previously associated with bipolar disorder. This study outlined the potential use of MDR and genetic association analysis to identify genotype combinations that are significant as a followup to previous studies that identify genes associated with a disease.
100	Edwards et. al performed an MDR analysis using simulated data using two separate two-locus epistatic disease models that both independently led to trait status to model genetic heterogeneity. Their analysis of the simulated data demonstrated that MDR can accurately compensate for data heterogeneity.
101	Polonikov et. al utilized MDR to identify significant genotype combinations that are associated with asthma. This study performed a post hoc stratification of the results based on gender in an effort to account for heterogeneity

38	Cattaert et. al utilized a MDR variant called Model-Based MDR on simulated data. In particular, they found that MB-MDR significantly outperforms MDR when dealing with significant heterogeneity in the data, with MB-MDR having a power of 80% for a four-locus model compared to only 8% for classical MDR in simulated data with significant heterogeneity.
32	Gui et. al utilized a MDR variant called Three Way Split MDR on simulated data in conjunction with classical MDR in order to compare the variation to classical MDR. They found that Three Way Split MDR significantly reduced the computational complexity of the analysis due to its use of a regression based statistical model pruning technique in place of cross validation.
109	Gui et al. proposed a variant of MDR called Survival MDR that modifies MDR's constructive induction algorithm to use the log-rank test. The authors used simulation data involving a disease model in which two SNPs interact epistatically with no main effects to produce the disease and reported a statistical power of up to 70% to detect the causative genes when the causative genes were embedded in a set of 10 predictor genes under consideration. This study demonstrates the ability of MDR to detect gene-gene interactions even in the total absence of main effects.

Table 16 - Summary of Selected MDR Studies Included in this Review

APPENDICES

Appendix A: Literature Queries Used to Learn Background Information on Complex Diseases

Query	Database or Search Engine	Filter(s) applied
(complex[Title]) AND (disease[Title] OR diseases[Title] OR trait[Title] OR traits[Title])	PubMed	Review
"complex disease" OR "complex trait" +review	Google	N/A

Appendix B: Literature Queries Used to Learn Background Information on Epistasis

Query	Database or Search Engine	Filter(s) applied
epistasis[Title/Abstract] OR epistatic[Title/Abstract]	PubMed	Review
epistasis	Google	N/A

Appendix C: Literature Queries Used to Learn Background Information on Machine Learning as Applied to Genetic Association Studies

Query	Database or Search Engine	Filter(s) applied
(machine learning[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	Review
(machine learning[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	published in the last 5 years

Appendix D: Literature Queries to Learn Background Information on the Missing Heritability Problem

Query	Database or Search Engine	Filter(s) applied
(missing heritability[Title/Abstract])	Pubmed	published in the last 5 years
(missing heritability[Title/Abstract]) AND gwas	Pubmed	published in the last 5 years
(missing heritability[Title/Abstract]) AND epistasis	Pubmed	published in the last 5 years
(missing heritability[Title/Abstract]) AND (main effect OR main effects)	Pubmed	published in the last 5 years
"missing heritability"	Google	N/A

Appendix E: Literature Queries Used to Learn Background Information on Neural Networks

Query	Database or Search Engine	Filter(s) applied
"neural network" theory	Google	N/A
(neural network[Title])	Pubmed	N/A
(neural network[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract])	Pubmed	published in the last 5 years

Appendix F: Literature Queries Used to Learn Background Information on Multifactor Dimensionality Reduction

Query	Database or Search Engine	Filter(s) applied
"multifactor dimensionality reduction" theory	Google	N/A
(multifactor dimensionality reduction[Title])	Pubmed	N/A
(multifactor dimensionality reduction[Title/Abstract])	Pubmed	N/A
(multifactor dimensionality reduction[Title/Abstract])	Pubmed	published in the last 5 years

Appendix G: Literature Queries Used to Learn Background Information on Random Forests

Query	Database or Search Engine	Filter(s) applied
"random forest" theory	Google	N/A
(random[Title] AND (forest[Title] OR forests[Title]))	Pubmed	N/A
(random[Title/Abstract] AND (forest[Title/Abstract]OR forests[Title/Abstract]))	Pubmed	N/A
(random[Title/Abstract] AND (forest[Title/Abstract] OR forests[Title/Abstract]))	Pubmed	published in the last 5 years

Appendix H: Literature Queries Used to Learn Background Information on Support Vector Machines

Query	Database or Search Engine	Filter(s) applied
"support vector machine" theory	Google	N/A
(support vector[Title] AND (machine[Title] OR machines[Title]))	Pubmed	N/A
(support vector[Title/Abstract] AND (machine[Title/Abstract] OR machines[Title/Abstract]))	Pubmed	N/A
(support vector[Title/Abstract] AND (machine[Title/Abstract] OR machines[Title/Abstract]))	Pubmed	published in the last 5 years

Appendix I: Literature Queries Used to Learn Background Information on K-Nearest Neighbors

Query	Database or Search Engine	Filter(s) applied
"k nearest neighbor" theory	Google	N/A
(nearest[Title] AND (neighbor[Title] OR neighbors[Title]))	Pubmed	N/A
(nearest[Title/Abstract] AND (neighbor[Title/Abstract] OR neighbors[Title/Abstract]))	Pubmed	N/A
(nearest[Title/Abstract] AND (neighbor[Title/Abstract] OR neighbors[Title/Abstract]))	Pubmed	published in the last 5 years

Appendix J: Literature Queries Used to Learn Information on Neural Networks and Genetic Association Studies

Query	Database or Search Engine	Filter(s) applied
(neural network[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract]) AND heterogeneity	Pubmed	N/A
(neural network[Title/Abstract]) AND (effect size OR main effect OR main effects)	Pubmed	N/A
(neural network[Title/Abstract] AND Bayesian[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract] AND genetic programming[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract] AND grammatical evolution[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract] AND Bayesian[Title/Abstract])	Pubmed	N/A

(neural network[Title/Abstract] AND fuzzy[Title/Abstract])	Pubmed	N/A
(neural network[Title/Abstract] AND (variant[Title/Abstract] OR variation[Title/Abstract] OR alternative[Title/Abstract] OR modification[Title/Abstract]))	Pubmed	published in the last 5 years
(neural network[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(neural network[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(neural network[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	published in the last 5 years

Appendix K: Literature Queries Used to Learn Information on Multifactor Dimensionality Reduction and Genetic Association Studies

Query	Database or Search Engine	Filter(s) applied
(multifactor dimensionality reduction[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(multifactor dimensionality reduction[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(multifactor dimensionality reduction[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	N/A
(multifactor dimensionality reduction[Title/Abstract]) AND heterogeneity	Pubmed	N/A
(multifactor dimensionality reduction[Title/Abstract]) AND (effect size OR main effect OR main effects)	Pubmed	N/A

(multifactor dimensionality reduction[Title/Abstract] AND (variant[Title/Abstract] OR variation[Title/Abstract] OR alternative[Title/Abstract] OR modification[Title/Abstract]))	Pubmed	published in the last 5 years
(multifactor dimensionality reduction[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(multifactor dimensionality reduction[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(multifactor dimensionality reduction[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	published in the last 5 years

Appendix L: Literature Queries Used to Learn Information on Random Forests and Genetic Association Studies

Query	Database or Search Engine	Filter(s) applied
(random forest[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(random forest[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(random forest[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	N/A
(random forest[Title/Abstract]) AND heterogeneity	Pubmed	N/A
(random forest[Title/Abstract]) AND (effect size OR main effect OR main effects)	Pubmed	N/A
(random forest[Title/Abstract] AND (variant[Title/Abstract] OR variation[Title/Abstract] OR alternative[Title/Abstract] OR modification[Title/Abstract]))	Pubmed	published in the last 5 years
(random forest[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND	Pubmed	published in the last 5 years

(epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])		
(random forest[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(random forest[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	published in the last 5 years

Appendix M: Literature Queries Used to Learn Information on Support Vector Machines and Genetic Association Studies

Query	Database or Search Engine	Filter(s) applied
(support vector machine[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(support vector machine[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(support vector machine[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	N/A
(support vector machine[Title/Abstract]) AND heterogeneity	Pubmed	N/A
(support vector machine[Title/Abstract]) AND (effect size OR main effect OR main effects)	Pubmed	N/A
(support vector machine[Title/Abstract] AND (variant[Title/Abstract] OR variation[Title/Abstract] OR alternative[Title/Abstract] OR modification[Title/Abstract]))	Pubmed	published in the last 5 years

(support vector machine[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(support vector machine[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(support vector machine[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	published in the last 5 years

Appendix N: Literature Queries Used to Learn Information on K-Nearest Neighbors and Genetic Association Studies

Query	Database or Search Engine	Filter(s) applied
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	N/A
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	N/A
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND heterogeneity	Pubmed	N/A
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND (effect size OR main effect OR main effects)	Pubmed	N/A
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract] AND (variant[Title/Abstract] OR	Pubmed	published in the last 5 years

variation[Title/Abstract] OR alternative[Title/Abstract] OR modification[Title/Abstract]))		
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND (epistasis[Title/Abstract] OR interaction[Title/Abstract] OR epistatic[Title/Abstract])	Pubmed	published in the last 5 years
(nearest neighbor[Title/Abstract] OR nearest neighbors[Title/Abstract]) AND (gene[Title/Abstract] OR genetic[Title/Abstract] OR genome[Title/Abstract])	Pubmed	published in the last 5 years

REFERENCES

1. E. S. Lander, “Initial impact of the sequencing of the human genome,” Nature, vol. 470, no. 7333, pp. 187–197, Feb. 2011
2. T. A. Manolio, L. D. Brooks, and F. S. Collins, “A HapMap harvest of insights into the genetics of common disease,” J. Clin. Invest., vol. 118, no. 5, pp. 1590–1605, May 2008.
3. J. N. Hirschhorn, “Genomewide association studies--illuminating biologic pathways,” N. Engl. J. Med., vol. 360, no. 17, pp. 1699–1701, Apr. 2009.
4. O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, “The mystery of missing heritability: Genetic interactions create phantom heritability,” Proc Natl Acad Sci U S A, vol. 109, no. 4, pp. 1193–1198, Jan. 2012.
5. T. F. C. Mackay, “Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene Interactions,” Nat Rev Genet, vol. 15, no. 1, pp. 22–33, Jan. 2014.
6. J. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human disease," *Hum Hered.,* vol. 56, no. 1-3, pp. 73-82, 2003.
7. J. Craig and Others, "Complex Diseases: Research and Applications," *Nature Education,* p. 1, 2008.
8. H. J. Cordell, “Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans,” Hum. Mol. Genet., vol. 11, no. 20, pp. 2463–2468, Oct. 2002.
9. R. A. Fisher, “The Correlation Between Relatives on the Supposition of Mendelian

Inheritance.," 1918.

10. M. A. Wenzel and S. B. Piertney, "Digging for gold nuggets: uncovering novel candidate genes for variation in gastrointestinal nematode burden in a wild bird species," J. Evol. Biol., p. n/a–n/a, 2015.
11. D. Lvovs, O. Favorova and A. Favorov, "A Polygenic Approach to the Study ?of Polygenic Diseases," *Acta Naturae,* vol. 4, no. 3, pp. 59-71, 2012.
12. R. Culverhouse, B. Suarez, J. Lin and T. Reich, "Perspective on Epistasis: Limits of Models Displaying No Main Effect," *American Journal of Human Genetics,* vol. 70, no. 2, pp. 461-71, 2002.
13. J. Moore and S. Williams, "Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis," *Bioessays,* vol. 27, no. 6, pp. 637-46, 2005.
14. N. Yi, "Statistical Analysis of Genetic Interactions," *Genet Res (Camb),* vol. 92, no. 5–6, pp. 443–459, Dec. 2010.
15. X. Shen, M. Pettersson, L. Rönnegård, and Ö. Carlborg, "Inheritance Beyond Plain Heritability: Variance-Controlling Genes in Arabidopsis thaliana," PLoS Genet, vol. 8, no. 8, Aug. 2012.
16. R. M. Nelson, M. E. Pettersson, X. Li, and Ö. Carlborg, "Variance Heterogeneity in Saccharomyces cerevisiae Expression Data: Trans-Regulation and Epistasis," PLoS One,

vol. 8, no. 11, Nov. 2013.

17. T. Thornton-Wells, J. Moore and J. Haines, "Genetics, statistics and human disease: analytical retooling for complexity," *Trends in Genetics,* vol. 20, no. 12, pp. 640-7, 2004.
18. Y. Chan, O. Holmen, A. Dauber, L. Vatten, A. Havulinna, F. Skorpen and e. al., "Common Variants Show Predicted Polygenic Effects on Height in the Tails of the Distribution, Except in Extremely Short Individuals," *PLoS Genetics,* vol. 7, no. 12, p. e1002439, 2011.
19. R. Wilcox and R. Bellman, "Adaptive control processes--A guided tour," *Prince University Press: Naval Research Logistics Quarterly,* vol. 8, no. 3, pp. 315-6, 1961.
20. Y. Benjamin, D. Drai, G. Elmer, N. Kafkafi and I. Golani, "Controlling the false discovery rate in behavior genetics research," *Behavioral Brain Research,* vol. 125, no. 1-2, pp. 279-84, 2001.
21. W. Bush and J. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Computational Biology,* vol. 8, no. 12, 2012.
22. S. Garcia, A. Fernandez, J. Luengo and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability," *Soft Computing,* vol. 13, no. 10, pp. 959-77, 2009.
23. G. Koch, "The use of non-parametric methods in the statistical analysis of the two-period change-over design. Biometrics," *Biometrics,* vol. 28, no. 2, pp. 577-84, 1972.

24. K. G. Sheela and S. N. Deepa, “Review on Methods to Fix Number of Hidden Neurons in Neural Networks,” Mathematical Problems in Engineering, vol. 2013, p. e425740, Jun. 2013.

25. A. A. Motsinger, S. L. Lee, G. Mellick, and M. D. Ritchie, “GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease,” BMC Bioinformatics, vol. 7, no. 1, p. 39, Jan. 2006.

26. S. D. Turner, S. M. Dudek, and M. D. Ritchie, “Grammatical Evolution of Neural Networks for Discovering Epistasis among Quantitative Trait Loci,” in Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, C. Pizzuti, M. D. Ritchie, and M. Giacobini, Eds. Springer Berlin Heidelberg, 2010, pp. 86–97.

27. B. Djavan, M. Remzi, A. Zlotta and e. al, "Novel artificial neural network for early detection of prostate cancer," Journal of Clinical Oncology, vol. 20, pp. 921-929, 2002.

28. D. L. Tong, D. J. Boocock, G. Dhondalay and e. al., "Artificial neural network inference (ANNI): a study on gene-gene interaction for biomarkers in childhood sarcomas," PLoS One, vol. 9, no. 7, p. e102483, 2014.

29. D. GK, T. DL and B. GR, "Estrogen receptor status prediction for breast cancer using artificial neural network," Proceedings of the 2011 International Conference on Machine Learning and Cybernetics (ICMLC 2011), vol. 2, pp. 727-731, 2011.

30. L. Lancashire, B. Rees and G. Ball, "Identification of gene transcript signatures

predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach," Artificial Intelligence Medicine, vol. 43, pp. 99-111, 2008.

31. B. Matharoo-Ball, C. Hughes, L. Lancashire, D. Tooth, G. Ball and e. al., "Characterization of biomarkers in polycystic ovary syndrome (PCOS) using multiple distinct proteomic platforms," Journal Of Proteome Research, vol. 6, pp. 3321-3328, 2007.
32. J. Gui, A. S. Andrew, P. Andrews, H. M. Nelson, K. T. Kelsey, M. R. Karagas, and J. H. Moore, "A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility," Ann Hum Genet, vol. 75, no. 1, pp. 20–28, Jan. 2011.
33. S. J. Winham, A. J. Slater, and A. A. Motsinger-Reif, "A comparison of internal validation techniques for multifactor dimensionality reduction," BMC Bioinformatics, vol. 11, no. 1, p. 394, Jul. 2010.
34. P. S. Silva, V. Fontana, M. R. Luizon, R. Lacchini, W. A. Silva, C. Biagi, and J. E. Tanus-Santos, "eNOS and BDKRB2 genotypes affect the antihypertensive responses to enalapril," Eur. J. Clin. Pharmacol., vol. 69, no. 2, pp. 167–177, Feb. 2013.
35. M. R. Luizon, V. C. Sandrim, A. C. Palei, R. Lacchini, R. C. Cavalli, G. Duarte, and J. E. Tanus-Santos, "Epistasis among eNOS, MMP-9 and VEGF maternal genotypes in

hypertensive disorders of pregnancy," Hypertens. Res., vol. 35, no. 9, pp. 917–921, Sep. 2012.

36. J. M. Mahachie John, F. Van Lishout, and K. Van Steen, "Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data," Eur J Hum Genet, vol. 19, no. 6, pp. 696–703, Jun. 2011.
37. C.-F. Li, F.-T. Luo, Y.-X. Zeng, and W.-H. Jia, "Weighted Risk Score-Based Multifactor Dimensionality Reduction to Detect Gene-Gene Interactions in Nasopharyngeal Carcinoma," Int J Mol Sci, vol. 15, no. 6, pp. 10724–10737, Jun. 2014.
38. T. CATTAERT, M. L. CALLE, S. M. DUDEK, J. M. MAHACHIE JOHN, F. VAN LISHOUT, V. URREA, M. D. RITCHIE, and K. VAN STEEN, "A detailed view on Model-Based Multifactor Dimensionality Reduction for detecting gene-gene interactions in case-control data in the absence and presence of noise," Ann Hum Genet, vol. 75, no. 1, pp. 78–89, Jan. 2011.
39. J. Gui, A. S. Andrew, P. Andrews, H. M. Nelson, K. T. Kelsey, M. R. Karagas, and J. H. Moore, "A Simple and Computationally Efficient Sampling Approach to Covariate Adjustment for Multifactor Dimensionality Reduction Analysis of Epistasis," Hum Hered, vol. 70, no. 3, pp. 219–225, Oct. 2010.
40. G.-B. Chen, J. Zhu, and X.-Y. Lou, "A faster pedigree-based generalized multifactor dimensionality reduction method for detecting gene-gene interactions," Stat Interface,

vol. 4, no. 3, pp. 295–304, Jan. 2011.

41. M. G. M. Braem, M. Voorhuis, Y. T. van der Schouw, P. H. M. Peeters, L. J. Schouten, M. J. C. Eijkemans, F. J. Broekmans, and N. C. Onland-Moret, "Interactions between Genetic Variants in AMH and AMHR2 May Modify Age at Natural Menopause," PLoS One, vol. 8, no. 3, Mar. 2013.
42. F. Di Cianni, D. Campa, F. Tallaro, C. Rizzato, F. De Rango, R. Barale, G. Passarino, F. Canzian, F. Gemignani, A. Montesanto, S. Landi, and G. Rose, "MAP3K7 and GSTZ1 are associated with human longevity: a two-stage case–control study using a multilocus genotyping," Age (Dordr), vol. 35, no. 4, pp. 1357–1366, Aug. 2013.
43. J. S. Park, S. B. Choi, J. W. Chung, S. W. Kim, and D. W. Kim, "Classification of serous ovarian tumors based on microarray data using multicategory support vector machines," Conf Proc IEEE Eng Med Biol Soc, vol. 2014, pp. 3430–3433, 2014.
44. A. T. Lu and R. M. Cantor, "Identifying rare-variant associations in parent-child trios using a Gaussian support vector machine," *BMC Proc*, vol. 8, no. Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo, p. S98, 2014.
45. Z. Gillani, M. Akash, M. Rahaman, and M. Chen, "CompareSVM: supervised, Support Vector Machine (SVM) inference of gene regularity networks," *BMC Bioinformatics*, vol. 15, no. 1, p. 395, Nov. 2014.
46. S. B. Choi, J. S. Park, J. W. Chung, T. K. Yoo, and D. W. Kim, "Multicategory

classification of 11 neuromuscular diseases based on microarray data using support vector machine," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2014, pp. 3460–3463, 2014.

47. H. Saberkari, M. Shamsi, M. Joroughi, F. Golabi, and M. H. Sedaaghi, "Cancer Classification in Microarray Data using a Hybrid Selective Independent Component Analysis and υ-Support Vector Machine Algorithm," *J Med Signals Sens*, vol. 4, no. 4, pp. 291–298, Oct. 2014.
48. J. C. Cuevas Tello, D. Hernández-Ramírez, and C. A. García-Sepúlveda, "Support vector machine algorithms in the search of KIR gene associations with disease," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2053–2062, Dec. 2013.
49. Y. Lee and C.-K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," Bioinformatics, vol. 19, no. 9, pp. 1132–1139, Jun. 2003.
50. X. Jiang, B. Cai, D. Xue, X. Lu, G. F. Cooper, and R. E. Neapolitan, "A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets," Journal of the American Medical Informatics Association, vol. 21, no. e2, pp. e312–e319, Oct. 2014.
51. H. Han and X. Jiang, "Overcome Support Vector Machine Diagnosis Overfitting," Cancer Inform, vol. 13, no. Suppl 1, pp. 145–158, Dec. 2014.
52. Y. Shen, Z. Liu, and J. Ott, "Support vector machines with L1 penalty for detecting gene-

gene interactions," Int J Data Min Bioinform, vol. 6, no. 5, pp. 463–470, 2012.

53. K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, "Screening large-scale association study data: exploiting interactions using random forests," BMC Genet., vol. 5, p. 32, 2004.

54. S. J. Winham, C. L. Colby, R. R. Freimuth, X. Wang, M. de Andrade, M. Huebner, and J. M. Biernacka, "SNP interaction detection with Random Forests in high-dimensional genetic data," BMC Bioinformatics, vol. 13, p. 164, Jul. 2012.

55. N. R. Cook, R. Y. L. Zee, and P. M. Ridker, "Tree and spline based association analysis of gene-gene interaction models for ischemic stroke," Stat Med, vol. 23, no. 9, pp. 1439–1453, May 2004.

56. X. Chen and H. Ishwaran, "Random Forests for Genomic Data Analysis," Genomics, vol. 99, no. 6, pp. 323–329, Jun. 2012.

57. D. F. Schwarz, I. R. König, and A. Ziegler, "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data," Bioinformatics, vol. 26, no. 14, pp. 1752–1758, Jul. 2010.

58. B. A. McKinney, B. C. White, D. E. Grill, P. W. Li, R. B. Kennedy, G. A. Poland, and A. L. Oberg, "ReliefSeq: A Gene-Wise Adaptive-K Nearest-Neighbor Feature Selection Tool for Finding Gene-Gene Interactions and Main Effects in mRNA-Seq Gene Expression Data," PLoS ONE, vol. 8, no. 12, p. e81527, Dec. 2013.

59. N. Salari, S. Shohaimi, F. Najafi, M. Nallappan, and I. Karishnarajah, "A novel hybrid classification model of genetic algorithms, modified k-Nearest Neighbor and developed backpropagation neural network," PLoS ONE, vol. 9, no. 11, p. e112987, 2014.

60. A. Suratanee and K. Plaimas, "Identification of inflammatory bowel disease-related proteins using a reverse k-nearest neighbor search," J Bioinform Comput Biol, vol. 12, no. 4, p. 1450017, Aug. 2014.

61. R. M. Parry, W. Jones, T. H. Stokes, J. H. Phan, R. A. Moffitt, H. Fang, L. Shi, A. Oberthuer, M. Fischer, W. Tong, and M. D. Wang, "k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," Pharmacogenomics J., vol. 10, no. 4, pp. 292–309, Aug. 2010.

62. K. Ning, H. K. Ng, S. Srihari, H. W. Leong, and A. I. Nesvizhskii, "Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology," BMC Bioinformatics, vol. 11, p. 505, 2010.

63. C. F. Pasluosta, P. Dua, and W. J. Lukiw, "Nearest hyperplane distance neighbor clustering algorithm applied to gene co-expression analysis in Alzheimer's disease," Conf Proc IEEE Eng Med Biol Soc, vol. 2011, pp. 5559–5562, 2011.

64. P. Lucek and J. Ott, "Neural Network Analysis of Complex Traits," *Genetic Epidemiology*, vol. 14, pp. 1101-6, 1997.

65. N. Hardison and A. Motsinger-Reif, "The power of quantitative grammatical evolution

neural networks to detect gene-gene interactions," *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation,* pp. 299-306, 2011.

66. K. Hornik, M. Stinchcombe and H. White, "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks,* vol. 2, no. 5, pp. 359-66, 1989.
67. M. Ritchie, B. White, J. Parker, L. Hahn and J. Moore, " Optimization of neural network architecture using genetic programming improves detection of gene-gene interactions in studies of human diseases.," BMC Bioinformatics, vol. 4, 2003.
68. A. L. Beam, A. Motsinger-Reif, and J. Doyle, “Bayesian neural networks for detecting epistasis in genetic association studies,” BMC Bioinformatics, vol. 15, no. 1, p. 368, Nov. 2014.
69. M.M Gupta and D.H Rao, “On the principles of fuzzy neural networks,” Fuzzy Sets and Systems, vol. 61, no. 1, pp. 1-18, Jan. 1994
70. F. Günther, N. Wawro, and K. Bammann, “Neural networks for modeling gene-gene interactions in association studies,” BMC Genetics, vol. 10, no. 1, p. 87, Dec. 2009.
71. A. L. Beam, A. A. Motsinger-Reif, and J. Doyle, “An investigation of gene-gene interactions in dose-response studies with Bayesian nonparametrics,” BioData Mining, vol. 8, no. 1, p. 6, Feb. 2015.
72. S. D. Turner, S. M. Dudek, and M. D. Ritchie, “ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering

epistasis among quantitative trait Loci," BioData Mining, vol. 3, no. 1, p. 5, Sep. 2010.

73. A. G. Heidema, J. M. Boer, N. Nagelkerke, E. C. Mariman, D. L. van der A, and E. J. Feskens, "The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases," *BMC Genetics*, vol. 7, no. 1, p. 23, Apr. 2006.

74. K.-S. Lynn, L.-L. Li, Y.-J. Lin, C.-H. Wang, S.-H. Sheng, J.-H. Lin, W. Liao, W.-L. Hsu, and W.-H. Pan, "A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data," Bioinformatics, vol. 25, no. 8, pp. 981–988, Apr. 2009.

75. R. Xu, S. Damelin, B. Nadler, and D. C. Wunsch, "Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps," Artif Intell Med, vol. 48, no. 2–3, pp. 91–98, Mar. 2010.

76. M. L. Gonzalgo, S. Yegnasubramanian, G. Yan, C. G. Rogers, T. L. Nicol, W. G. Nelson, and C. P. Pavlovich, "Molecular profiling and classification of sporadic renal cell carcinoma by quantitative methylation analysis," Clin. Cancer Res., vol. 10, no. 21, pp. 7276–7283, Nov. 2004.

77. K. H. Chang, N. Miller, E. A. H. Kheirelseid, C. Lemetre, G. R. Ball, M. J. Smith, M. Regan, O. J. McAnena, and M. J. Kerin, "MicroRNA signature analysis in colorectal cancer: identification of expression profiles in stage II tumors associated with aggressive disease," Int J Colorectal Dis, vol. 26, no. 11, pp. 1415–1422, Nov. 2011.

78. S. Warkentin, M. Ohlsson, P. Wollmer, L. Edenbrandt, and L. Minthon, "Regional cerebral blood flow in Alzheimer's disease: classification and analysis of heterogeneity," Dement Geriatr Cogn Disord, vol. 17, no. 3, pp. 207–214, 2004.

79. D. M. Abd El-Rehim, G. Ball, S. E. Pinder, E. Rakha, C. Paish, J. F. R. Robertson, D. Macmillan, R. W. Blamey, and I. O. Ellis, "High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses," Int. J. Cancer, vol. 116, no. 3, pp. 340–350, Sep. 2005.

80. M. D'Alessandro, G. Vachtsevanos, R. Esteller, J. Echauz, S. Cranstoun, G. Worrell, L. Parish, and B. Litt, "A multi-feature and multi-channel univariate selection process for seizure prediction," Clin Neurophysiol, vol. 116, no. 3, pp. 506–516, Mar. 2005.

81. Y. Liang and A. G. Kelemen, "Hierarchical Bayesian Neural Network for Gene Expression Temporal Patterns," Stat Appl Genet Mol Biol, vol. 3, p. Article20, 2004.

82. T. Ando, M. Suguro, T. Kobayashi, M. Seto, and H. Honda, "Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling," Cancer Sci., vol. 94, no. 10, pp. 906–913, Oct. 2003.

83. T. Ando, M. Suguro, T. Hanai, T. Kobayashi, H. Honda, and M. Seto, "Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma," Jpn. J. Cancer Res., vol. 93, no. 11, pp. 1207–1212, Nov. 2002.

84. E. R. Holzinger, C. C. Buchanan, S. M. Dudek, E. C. Torstenson, S. D. Turner, and M. D. Ritchie, "Initialization Parameter Sweep in ATHENA: Optimizing Neural Networks for Detecting Gene-Gene Interactions in the Presence of Small Main Effects," Genet Evol Comput Conf, vol. 12, pp. 203–210, 2010.

85. M. D. Ritchie, A. A. Motsinger, W. S. Bush, C. S. Coffey, and J. H. Moore, "Genetic Programming Neural Networks: A Powerful Bioinformatics Tool for Human Genetics," Appl Soft Comput, vol. 7, no. 1, pp. 471–479, Jan. 2007.

86. W.-H. Wei, G. Hemani, and C. S. Haley, "Detecting epistasis in human complex traits," Nat Rev Genet, vol. 15, no. 11, pp. 722–733, Nov. 2014.

87. A. A. Motsinger-Reif, S. M. Dudek, L. W. Hahn, and M. D. Ritchie, "Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology," Genet. Epidemiol., vol. 32, no. 4, pp. 325–340, May 2008.

88. S. Ledesma, G. Cerda, G. Aviña, D. Hernández, and M. Torres, "Feature Selection Using Artificial Neural Networks," in MICAI 2008: Advances in Artificial Intelligence, A. Gelbukh and E. F. Morales, Eds. Springer Berlin Heidelberg, 2008, pp. 351–359.

89. G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

90. H. Wang, S.-H. Lo, T. Zheng, and I. Hu, "Interaction-based feature selection and

classification for high-dimensional biological data," Bioinformatics, vol. 28, no. 21, pp. 2834–2842, Nov. 2012.

91. E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *In Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 601–608.
92. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
93. M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer," Am J Hum Genet, vol. 69, no. 1, pp. 138–147, Jul. 2001.
94. R. L. Collins, T. Hu, C. Wejse, G. Sirugo, S. M. Williams, and J. H. Moore, "Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis," BioData Mining, vol. 6, no. 1, p. 4, Feb. 2013.
95. Y. Chung, S. Y. Lee, R. C. Elston, and T. Park, "Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions," Bioinformatics, vol. 23, no. 1, pp. 71–76, Jan. 2007.
96. F.-H. Liu, J.-Y. Song, X.-R. Shang, X.-R. Meng, J. Ma, and H.-J. Wang, "The Gene-Gene Interaction of INSIG-SCAP-SREBP Pathway on the Risk of Obesity in Chinese

Children," Biomed Res Int, vol. 2014, 2014.

97. O. Kebir, B. Chaumette, M. Fatjó-Vilas, A. Ambalavanan, N. Ramoz, L. Xiong, F. Mouaffak, B. Millet, N. Jaafari, L. E. DeLisi, D. Levinson, R. Joober, L. Fañanás, G. Rouleau, C. Dubertret, and M.-O. Krebs, "Family-based association study of common variants, rare mutation study and epistatic interaction detection in HDAC genes in schizophrenia," Schizophr. Res., vol. 160, no. 1–3, pp. 97–103, Dec. 2014.

98. R. Kumar, S. Kohli, P. Alam, R. Barkotoky, M. Gupta, S. Tyagi, S. K. Jain, and M. A. Q. Pasha, "Interactions between the FTO and GNB3 Genes Contribute to Varied Clinical Phenotypes in Hypertension," PLoS ONE, vol. 8, no. 5, p. e63934, May 2013.

99. Y.-C. Lai, C.-F. Kao, M.-L. Lu, H.-C. Chen, P.-Y. Chen, C.-H. Chen, W. W. Shen, J.-Y. Wu, R.-B. Lu, and P.-H. Kuo, "Investigation of Associations between NR1D1, RORA and RORB Genes and Bipolar Disorder," PLoS ONE, vol. 10, no. 3, p. e0121245, Mar. 2015.

100. T. L. Edwards, K. Lewis, D. R. Velez, S. Dudek, and M. D. Ritchie, "Exploring the performance of Multifactor Dimensionality Reduction in large scale SNP studies and in the presence of genetic heterogeneity among epistatic disease models," Hum. Hered., vol. 67, no. 3, pp. 183–192, 2009.

101. A. V. Polonikov, V. P. Ivanov, A. D. Bogomazov, M. B. Freidin, T. Illig, and M. A. Solodilova, "Antioxidant Defense Enzyme Genes and Asthma Susceptibility: Gender-

Specific Effects and Heterogeneity in Gene-Gene Interactions between Pathogenetic Variants of the Disease," Biomed Res Int, vol. 2014, 2014.

102. T. Cattaert, M. L. Calle, S. M. Dudek, J. M. Mahachie John, F. Van Lishout, V. Urrea, M. D. Ritchie, and K. Van Steen, "Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise," Ann. Hum. Genet., vol. 75, no. 1, pp. 78–89, Jan. 2011.

103. H. Mei, M. L. Cuccaro, and E. R. Martin, "Multifactor Dimensionality Reduction–Phenomics: A Novel Method to Capture Genetic Heterogeneity with Use of Phenotypic Variables," Am J Hum Genet, vol. 81, no. 6, pp. 1251–1261, Dec. 2007.

104. H. Dai, M. Bhandary, M. Becker, J. S. Leeder, R. Gaedigk, and A. A. Motsinger-Reif, "Global tests of P-values for multifactor dimensionality reduction models in selection of optimal number of target genes," BioData Min, vol. 5, p. 3, May 2012.

105. J. M. Mahachie John, T. Cattaert, L. De Lobel, F. Van Lishout, A. Empain, and K. Van Steen, "Comparison of genetic association strategies in the presence of rare alleles," BMC Proc, vol. 5, no. Suppl 9, p. S32, Nov. 2011.

106. J. Choi and T. Park, "Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions," BMC Syst Biol, vol. 7, no. Suppl 6, p. S15, Dec. 2013.

107. H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions*

on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

108. R. C. Culverhouse, "A comparison of methods sensitive to interactions with small main effects," Genet. Epidemiol., vol. 36, no. 4, pp. 303–311, May 2012.

109. J. Gui, J. H. Moore, K. T. Kelsey, C. J. Marsit, M. R. Karagas, and A. S. Andrew, "A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis," Hum. Genet., vol. 129, no. 1, pp. 101–110, Jan. 2011.

110. N. O. Oki and A. A. Motsinger-Reif, "Multifactor dimensionality reduction as a filter-based approach for genome wide association studies," Front Genet, vol. 2, p. 80, 2011.

111. M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer," Am J Hum Genet, vol. 69, no. 1, pp. 138–147, Jul. 2001.

112. H. He, W. S. Oetting, M. J. Brott, and S. Basu, "Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study," BMC Med. Genet., vol. 10, p. 127, 2009.

113. J. M. Mahachie John, F. Van Lishout, E. S. Gusareva, and K. Van Steen, "A robustness study of parametric and non-parametric tests in model-based multifactor

dimensionality reduction for epistasis detection," BioData Min, vol. 6, p. 9, Apr. 2013.

114. J. Xu, J. Lowey, F. Wiklund, J. Sun, F. Lindmark, F.-C. Hsu, L. Dimitrov, B. Chang, A. R. Turner, W. Liu, H.-O. Adami, E. Suh, J. H. Moore, S. L. Zheng, W. B. Isaacs, J. M. Trent, and H. Grönberg, "The Interaction of Four Genes in the Inflammation Pathway Significantly Predicts Prostate Cancer Risk," Cancer Epidemiol Biomarkers Prev, vol. 14, no. 11, pp. 2563–2568, Nov. 2005.

115. J. Choi and T. Park, "Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions," BMC Systems Biology, vol. 7, no. Suppl 6, p. S15, Dec. 2013.

116. J. Gui, J. H. Moore, K. T. Kelsey, C. J. Marsit, M. R. Karagas, and A. S. Andrew, "A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis," Hum. Genet., vol. 129, no. 1, pp. 101–110, Jan. 2011.

117. H. Mei, D. Ma, A. Ashley-Koch, and E. R. Martin, "Extension of multifactor dimensionality reduction for identifying multilocus effects in the GAW14 simulated data," BMC Genet., vol. 6 Suppl 1, p. S145, 2005.