Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

PHENOTYPE PREDICTION AND FEATURE SELECTION IN GENOME-WIDE ASSOCIATION STUDIES

by
Andrew Roberts

Genome wide association studies (GWAS) search for correlations between single nucleotide polymorphisms (SNPs) in a subject genome and an observed phenotype. GWAS can be used to generate models for predicting phenotype based on genotype, as well as aiding in identification of specific genes affecting the biological mechanism underlying the phenotype.

In this investigation, phenotype prediction models are constructed from GWAS training data and are evaluated for performance on test data. Three methods are used to rank SNPs by their correlation with the phenotype: the univariate Wald test, a multivariate, support vector machine (SVM) based technique, and a hybrid method where a subset of top ranked SNPs from the Wald test are used to train the SVM. Both case-control studies and quantitative phenotypes are examined. For each method and data set, a series of least squares linear regression models is generated from nested subsets of the best SNPs from each ranking method. The accuracy of these models is determined on a test data set, and a plot of prediction performance against the number of top ranked SNPs considered is generated.

The SVM and hybrid methods are found to be consistently superior to the Wald test in ranking predictive SNPs. The hybrid method allows a useful trade-off between increasing accuracy vs. using fewer SNPs to be optimized as desired.

PHENOTYPE PREDICTION AND FEATURE SELECTION IN GENOME-WIDE ASSOCIATION STUDIES

by
Andrew Roberts

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

May 2012

APPROVAL PAGE

PHENOTYPE PREDICTION AND FEATURE SELECTION IN GENOME-WIDE ASSOCIATION STUDIES

Andrew Roberts

Dr. Usman Roshan, Thesis Advisor — Date
Assistant Professor of Computer Science, NJIT

Dr. Jason T. L. Wang, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Assistant Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Andrew Roberts

Degree: Master of Science

Date: May 2012

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2012

- Bachelor of Science in Physics,
  University of Michigan, Ann Arbor, MI, 1990

Major: Bioinformatics

Dedicated to Michael Kaplan, whose delight after our discussion of a complicated scientific research topic was summed up when he said "you know, I bet it turns out to be really simple" -- in the hope that he turns out to be right.

ACKNOWLEDGMENT

I wish to thank my advisor, Dr. Usman Roshan, for his zeal, breadth of knowledge, guidance, and patience with my occasionally contrary and opinionated nature. I also thank Dr. Jason Wang and Dr. Zhi Wei for serving on my masters thesis committee as well as for the excellent instruction they have provided me during my bioinformatics studies at NJIT.

I wish to thank my employer, Merck, for providing financial support for my graduate education at NJIT. I thank my colleagues Gene Fluder, Matthew Walker, and Jim Tata not only for officially approving this financial support, but also for providing encouragement and understanding as I dealt with the difficulties of pursuing a graduate degree while working full time.

TABLE OF CONTENTS

TABLE OF CONTENTS
(continued)

LIST OF TABLES

Table **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Genome-wide association studies (GWAS) search the entire genetic profile of an organism for genetic variants, the result of mutation, that are correlated with phenotypes such as disease state. Identification of such variants can be used to predict unmanifested phenotypes, such as likelihood of developing a disease, in subjects. Knowledge of these variants can also shed light on the biological mechanisms underlying disease or other phenotypes. The ability of GWAS to search over whole genomes can greatly accelerate the discovery of this useful knowledge, but this very power makes accurate analysis of GWAS data difficult. Care must be taken to discern between useful results that reflect actual biological mechanisms and spurious findings caused by random correlations in the vast flood of raw data created.

This investigation examines the use of statistical algorithms to find truly predictive SNPs in GWAS. Three techniques are used to rank order SNPs by their significance to the phenotype: a univariate Wald test, which considers each SNP independently; a multivariate support vector machine (SVM), which considers all SNPs simultaneously; and a hybrid method which uses the Wald test to find a subset of best ranked SNPs which are then input to an SVM for final ranking. These techniques are applied to several datasets, including quantitative phenotypes for levels of blood chemicals in mice, simulated case/control data, and case/control studies for different types of cancer. Once a ranking of SNPs is generated, a series of least squares linear regression models is generated from successively larger subsets of the best ranked SNPs.

Predictive performance of these models is determined on a test data set and a plot of predictivity versus number of best ranked SNPs is created. This procedure allows evaluation of the ability of each of the methods to discriminate truly predictive SNPs from false positives, without having to know beforehand which SNPs are actually significant.

This thesis is organized as follows: first a general description of GWAS is provided. Then a description is given of the various statistical machine learning techniques used to find meaningful patterns in GWAS in order to identify significant SNPs and make predictions of unknown phenotypes from a genotype. These techniques included least squares linear regression, the univariate Wald test for determining feature significance, the multivariate support vector machine technique for predictive modeling and feature selection, general machine learning techniques of model building and validation, and model selection by evaluating series of models generated from nested subsets of SNP features.

After this basic introduction to concepts, the specific use of these techniques in this investigation is outlined in detail. The data sets used are briefly described. Finally, results of the techniques as applied to the data are given and the findings are discussed.

.

CHAPTER 2

GENOME-WIDE ASSOCIATION STUDIES

2.1 Genotypes and Phenotypes

An organism's genotype is the set of all variants in the DNA code contained within each of the organism's cells. Within a species, the vast majority of the DNA sequence is identical between individuals. A particular location in this consensus sequence is known as a locus, so individuals in a species can be said to have identical genotypes at the majority of their loci. However mutations create variants in this DNA sequence which can be inherited by offspring. The combination of inherited DNA from both of an organism's parents, repeated over many generations, ultimately leads to each individual having a distinct, usually unique set of variants in its DNA sequence. The different variants that exist in a population for a specific locus in a specie's DNA code are called the alleles of that locus [1,2].

Many kinds of genetic variants exist. Segments of DNA within a chromosome can be inserted, deleted, replicated, moved to a different location in the same or in a different chromosome, or inverted in place. One particularly simple and common type of variant is the single nucleotide polymorphism or SNP. A SNP is created when a single nucleotide base in a DNA sequence is replaced with a different nucleotide base. For example, an adenine "A" base at a specific location could be replaced with a thymine "T" base. The nucleotide base variant most common in a population is called the major allele, while the less common base is the minor allele. It is estimated that there are 10 million SNPs in the human genome for which the minor allele is observed in at least 5

percent of the population. Millions of these SNPs have been identified and published for the human genome [1-3].

For chromosomes other than the gender determining X and Y chromosomes, all humans have two copies of each genetic locus, one inherited from the mother and another from the father. Therefore, the genotype of a person for a particular SNP can include 0, 1, or 2 copies of the minor allele, corresponding to 2, 1, or 0 copies of the major allele respectively [1-3].

Recently developed DNA microarrays, also known as "gene chips," allow hundreds of thousands of SNPs to be reliably genotyped for an organism quickly and relatively inexpensively. This has allowed studies to be done where thousands of subjects are typed for very many SNPs located throughout the entire range of the genome. These genotypes can then be compared with observed phenotypes in the subjects in order to determine if any SNPs are statistically associated with the phenotype. These experiments are known as genome-wide association studies, or GWAS [1].

The phenotype for GWAS is most often simply the presence or absence of a specific disease state, for example presence or absence of a specific type of cancer, type 2 diabetes, heart disease, or other condition. The phenotype could also be whether a particular drug or medical treatment works or does not work for a patient. These phenotypes are binary, representable by the two discrete values 1 or 0. GWAS using such a phenotype include case/control studies. It is also possible to use a continuously varying, quantitative phenotype for GWAS, such as the level of a chemical in the blood [4]. This investigation will analyze data from both kinds of studies: blood chemical levels measured in mice are used for two examples of quantitative phenotypes, while

presence or absence of disease is used as a binary phenotype in several case/control studies in humans.

2.2 Linkage and Haplotypes

A genetic variant which affects a phenotype, for example by increasing the level of a blood chemical or altering the odds of getting a disease, is called a casual variant. Even though GWAS can genotype hundreds of thousands of SNPs, it is unlikely that any of these SNPs will be a causal variant. The number, variety, and complexity of genetic variation exceeds the capability of GWAS to capture directly. However, GWAS are still useful for associating genotypes with phenotypes because genetic variants are not distributed independently. Genetic variants located near each other on a chromosome tend to be inherited together as a block through many generations. This phenomenon is called genetic linkage, and a block of linked variants is called a haplotype. Because the variants in a haplotype are highly correlated, genotyping just a few of the SNPs in a haplotype is sufficient to identify it uniquely. This allows GWAS to discover associations between phenotypes and causal variants that they do not directly genotype. Variant alleles in those genotyped SNPs which lie in the same haplotype as a causal variant will share the statistical association with the disease phenotype in such studies [1,5].

Genetic linkage occurs because DNA is formed into chromosomes. A single mutation will create a novel genetic variant on an individual chromosome that already has an inherited pattern of variants. If such a chromosome was consistently reproduced as a single unit during meiosis, than all the distinct variants on the chromosome would remain in perfect linkage as the chromosome spread throughout the population with new

generations of organisms. The effect would be that the chromosome would form a single haplotype. However, the phenomenon of chromosomal recombination, in which the two members of a pair of homologous chromosomes exchange corresponding lengths of DNA, tends to divide the linked variants into groups residing on distinct members of the homologous chromosome pair, after which they will be independently inherited. The larger the number of base nucleotides lying between two variants, the higher the probability that recombination will occur between them and the genetic linkage between them will break. Variants very close to each other, on the other hand, may stay linked for a large number of generations, so that a set of such variants will form a distinct haplotype which can be distributed throughout the entire population and endure for many generations [1,2,5].

2.3 SNP Association Models

Given that a SNP can act as a statistical proxy for a causal variant in the same haplotype, a mathematical model for the association between the SNP and phenotype must be posited. Recall that a subject organism can have 0, 1, or 2 copies of the minor allele for a given SNPs. The simplest model, allele counting, assumes that the effect of the minor allele is directly proportional to the number of copies in an organism. In effect, it assigns a value to the minor allele and assumes that the total effect of the SNP is this value multiplied by the minor allele count. For binary phenotypes such as the absence or presence of disease, the value is interpreted as an odds ratio. A baseline odds for getting the disease, applicable to a subject with no copies of the minor allele, is multiplied by this odds ratio: once for heterozygotes with one copy of the minor allele, and twice for homozygotes with two copies. For quantitative phenotypes the value is interpreted as the

contribution each copy of the minor allele will make to the numeric value of the phenotype [1,4,6,7].

This model makes the assumption that two copies of an allele have twice the effect of one copy. The current investigation uses this model for the sake of simplicity, but it is important to note that more complex models are often more realistic. In classical genetics, a dominant gene is one where a single copy of the causal allele has full effect, so that having two copies of the allele does not increase the effect. A recessive gene is one where there is no effect unless both alleles are the minor allele. Allele counting will not properly represent either of these common types of causal variant [2,7].

Another important assumption in all these models is that the effect of one causal variant is independent of the effect of a different variant. In reality, different genetic variants often interact with each other in a non-linear fashion, so that the effect of one variant is highly dependent on the presence or absence of another variant. Allowing "gene-gene interactions" into an association model makes it more realistic, but the statistical and computational complexity introduced thereby is considerable. Therefore many GWAS analyses, including the current investigation, ignore such interactions [7].

2.4 Statistical Errors in GWAS

Even with a simple allele counting model, correct identification of SNPs linked to causal variants is difficult because of standard types of statistical error. Type II errors occur when the association between a SNP and a phenotype is real, but is not strong enough to pass a preset standard for statistical significance. The null hypothesis -- that there is no association – is not rejected when it ideally should be, resulting in "false negative." If a causal variant only has a small effect on the phenotype, or a SNP is only

in loose linkage with a causal variant, type II errors can result. They can be reduced by adding more subjects to the GWAS, but this raises the cost.

Type I errors result when there is no true association between a SNP and a phenotype, but there appears to be one purely because of chance. The random distribution of a SNP's alleles can correlate with phenotype even though no causal variant is involved. This leads to rejection of the null hypothesis when it is actually a correct assessment, resulting in a "false positive." GWAS are particularly prone to type I errors because of multiple testing. The most useful feature of GWAS, the genotyping of many thousands of SNPs, multiplies the opportunity for type I errors. If a p-value cutoff for rejecting the null hypothesis is set to 0.1%, which for many kinds of statistical studies is considered a very strict threshold, a GWAS will still produce tens or hundreds of false positives because so many SNPs are tested. Only by setting a very low p-value threshold for rejecting the null hypothesis can the presence of type I errors be rendered unlikely -- but this inevitably increases the incidence of type II errors, screening out all but the most strongly associated SNPs. These issues make correct interpretation of GWAS results difficult [1].

2.5 Uses of GWAS

There are two possible aims for performing GWAS. The most common aim is to identify locations in the genome that are associated with disease or other phenotype of interest to aid in biological discovery of the underlying genetic mechanisms. Knowing that a genetic variant affecting a phenotype is found in a small region of a specific chromosome is of great benefit to further investigation to reveal the exact location, type, and effect of this variant. Disease-associated variants may reveal promising drug targets, biomarkers

for medical investigation, or other desired information. The abundance of type I errors in GWAS complicates the accurate determination of causal variants, but follow-up studies examining only the positive results of an initial GWAS can resolve these errors. It is rare for a single GWAS to unambiguously find only SNPS truly associated with important casual variants, but it can find a set of tentatively associated SNPs for further investigation [1].

The other possible aim of GWAS is to create a predictive model that will allow prediction of hidden or undeveloped phenotypes from genotyping. For example, if a certain genotype is often found in subjects who have already have a condition such as type II diabetes or heart disease, currently healthy individuals with similar genotype may have higher risk for developing this condition in the future. Identification of such individuals can aid in assigning appropriate medical monitoring and preventative care to the persons who need it most. Although predictive models are best generated only from SNPs linked to true causal variants, models with some degree of usefulness can still be generated from a set of GWAS positives without having to determine which positives are true and which are false [4]. The current examination takes advantage of this to compare SNP selection algorithms without ever knowing which SNPs are truly linked to causal variants.

CHAPTER 3

FEATURE SELECTION AND PREDICTIVE MODELING

SNP selection and phenotype prediction from GWAS data utilize techniques from the field known as machine learning. A range of machine learning techniques exist, and determining which ones are most useful for GWAS data is a complicated problem in its own right. The purpose of the current investigation is to compare the SNP selection performance of two common machine learning algorithms -- the univariate Wald test and the multivariate SVM -- on GWAS data, and to explore whether a hybrid method combining these algorithms can provide the best features of both.

Since the causal SNPs in the GWAS data used are not known in advance (except for one simulated case/control data set), a least squares linear regression model is generated from the selected SNPs in a training data set, which is then used to predict the phenotype for a test data set that was not used during SNP selection and regression model building. The measurable performance of the regression model on the test set is used as an indicator for the underlying performance of the SNP selection algorithm. The least squares linear regression model is also the method underlying the univariate Wald test.

3.1 Least Squares Linear Regression

A predictive model generated for a GWAS will calculate the phenotype value as a function of the genotyped SNPs. The simple allele counting model described above, which ignores gene-gene interactions and which considers the effect of each SNP to be proportional to the number of minor alleles, is linear. This means that the output of a GWAS model that uses simple allele counting will be based on a linear function of the

SNP values, where each SNP value is coded as 0, 1, or 2 for the number of minor alleles present at the SNP locus. Linear regression is the process of generating such a linear function to model the relationship between certain variables in a data set.

Linear regression tries to find a set of coefficients that best expresses an output value y, the dependent variable, as a linear function of one or more input values, the independent variables x_1 through x_n. This is expressed by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{3.1}$$

or in vector notation

$$y = \beta_0 + \vec{\mathbf{x}}^T \vec{\boldsymbol{\beta}} \tag{3.2}$$

where $\vec{\mathbf{x}}$ and $\vec{\boldsymbol{\beta}}$ are n-dimensional vectors.

If the n-vector $\vec{\mathbf{x}}$ is transformed into the vector of length n+1 by prepending a 1, this equation can be more simply written as

$$y = \vec{\mathbf{x}}^T \vec{\boldsymbol{\beta}} \tag{3.3}$$

where $\vec{\boldsymbol{\beta}}$ is now also of length n+1, with a "0th" component representing β_0 [8-10].

To fit a linear model for a set of GWAS data, assume there are m subject organisms with known phenotype and genotype at n SNPs. The phenotypes can be represented as a vector $\vec{\mathbf{y}}$ of length m, and the genotypes as an m × n+1 matrix **X**, where the first column, labeled the 0th component, is all ones. **y**[i] is the phenotype for the ith

subject, and **X**[i,j] is the genotype for the jth SNP of the ith subject. **X**[i,*] is the row vector representing the full genotype over all SNPs for the ith subject. The task at hand is to find coefficients β_0 through β_n such that the predicted phenotype for the ith subject $\mathbf{X}[i,*]\vec{\boldsymbol{\beta}}$ -- written as $\hat{y}[i]$ -- is closest to the actual observed **y**[i] for all i subjects. If n < m there is no exact solution and a compromise solution must be generated that somehow minimizes the error [8-10].

The most common way to define the "best" $\vec{\boldsymbol{\beta}}$ is by the least squares criterion. This is the line that minimizes the error defined as the sum over all squared residuals, where a residual is the difference between the observed phenotype **y**[i] for subject i and the predicted phenotype $\hat{y}[i] = \mathbf{X}[i,*]\vec{\boldsymbol{\beta}}$ calculated from the observed SNP genotypes for the subject **X**[i,*]. This sum is called the RSS for the residual sum of squares. The equation for this sum is:

$$\sum_{i=1}^{m} (y[i] - \mathbf{X}[i,*]\vec{\boldsymbol{\beta}}) \tag{3.4}$$

To find the minimum RSS as a function of $\vec{\boldsymbol{\beta}}$, set the derivative of RSS with respect to $\vec{\boldsymbol{\beta}}$ to $\vec{0}$ and solve. This results in what is known as the "normal equations", which in matrix form are written as

$$\mathbf{X}^T\mathbf{X}(\vec{\mathbf{y}} - \mathbf{X}\vec{\beta}) = \vec{\mathbf{0}} \tag{3.5}$$

The solution can then be expressed by the equation

$$\vec{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{\mathbf{y}} \tag{3.6}$$

provided that $\mathbf{X}^T\mathbf{X}$ is nonsingular so that its inverse exists.

It can be shown that the solution calculated by the least squares algorithm is the maximum likelihood estimator for $\vec{\boldsymbol{\beta}}$ if the true y for each subject is accurately modeled by a linear function of the independent variables $\vec{\mathbf{x}}$ for the subject plus a normally distributed error term with mean of zero and constant variance [8-10].

Determining a least squares model by solving the normal equations can give unstable results if some of the components of **X** -- the columns of the matrix **X** -- are highly correlated with each other. In the extreme case where two components, say **X**[*,i] and **X**[*,j], are perfectly correlated with each other, the matrix $\mathbf{X}^T\mathbf{X}$ becomes singular, its inverse does not exist, and the normal equations cannot be solved [10,11]. This case is important to discuss here because one of the genotyped data sets used in this investigation contains many SNPs that correlate perfectly with at least one other SNP in the data set, owing to a high degree of linkage disequilibrium between them. In addition to this concern, directly solving the normal equations is computationally inefficient. Therefore other algorithms are usually used to fit a least squares linear regression model, which are faster, more stable, and capable of producing a solution even when $\mathbf{X}^T\mathbf{X}$ is singular [10,11].

One common method uses QR decomposition. This factors the **X** matrix into matrices **Q** and **R**, where **Q** is orthogonal and **R** is upper triangular [10,11]. This is the method used by the "lm" function in the widely used R statistical programming language. lm is the standard R way to perform least squares linear regression. This function was

used to generate regression models in the early phases of the current investigation, but this ended up being unsatisfactory, owing to a subtle problem that is now described.

As the QR factorization is being formed, columns in **X** which are linearly dependent can be detected and omitted from the model. The ith column of **X** can be omitted simply by setting the ith component of $\vec{\boldsymbol{\beta}}$ to zero. Since no new information is contributed by a redundant column, the residual sum of squares for a model that omits it is not altered by the omission [10,11]. However this can lead to a problem when the regression model so formed is used for making predictions on a data set other than the one used to generate the model. Two independent variables that are perfectly correlated in the initial training data set may not be so correlated in the new test data set. In this instance, the model predictions for the test set will depend on which of the linearly dependent columns in the original data set was omitted. In the lm function in R, columns are examined in order of increasing index, from "left-to-right" as the **X** matrix is factored. If a column is discovered that is identical to a column already considered, the new column is the one omitted. This means that the model generated is dependent on the order that the SNPs are presented in the data set. As the current investigation involves creating least squares regression models for sets of SNPs that have been differently rank ordered by the different algorithms being examined, this dependence on SNP order has a confounding effect on correctly comparing the algorithms' performance.

Fortunately there is yet another means of generating a least squares linear regression model which is immune to this effect. Performing a singular value decomposition, or SVD, factorization of the **X** matrix allows the computation of the pseudoinverse of the matrix. This in turn allows determination of the minimum norm

solution to the least squares regression problem. If the **X** is nonsingular the minimum norm solution is just the unique solution to the normal equations. However if **X** is singular, there is no unique solution to the least squares problem [11]. This can be seen from the previous discussion of the QR least squares method, where the residual sum of squares could be minimized to the same value by omitting either one or the other of two identical columns in **X**, by setting the corresponding component of $\vec{\boldsymbol{\beta}}$ to zero. The minimum norm solution determined by SVD chooses a different solution. It "splits the difference" and assigns an effect to both of the redundant columns that is exactly half of the effect that the QR method assigns to one and only one column. This solution is unique for a given matrix **X** and is independent of the order of the columns in the **X** matrix [11]. For a GWAS model, this means that the SVD algorithm does not arbitrarily choose one of two redundant SNPs to be associated, but rather spreads the observed association evenly between both SNPs.

The SVD algorithm is used by the scipy.linalg.lstsq function in the scipy scientific library extension for the python programming language. This function was used to do linear regression for all final results in the current investigation, providing stable regression models free of column order dependency issues.

3.2 Univariate Feature Ranking: the Wald Test

Assessing the phenotype association of SNPs in a GWAS can be conducted either by univariate testing, which analyzes each SNP individually, or by multivariate testing, which simultaneously analyzes the association of a set of SNPs. Univariate methods are popular because they are simple and scale well to large data sets.

For binary phenotypes and the simple allele counting model of SNP association, the "gold standard" statistical test is chi-square, based on the 2 x 2 contingency table for the two phenotypes and the two alleles. For quantitative phenotypes it is necessary to perform a Wald test, which generates a least squares linear regression model using a single SNP for an independent variable and tests the significance of the model against the null hypothesis that there is no association between genotype and phenotype. For a single SNP equation (1) reduces to

$$y = \beta_0 + \beta_1 x \tag{3.7}$$

where x is the minor allele count for the SNP.

The null hypothesis is expressed numerically by setting β_1 to zero. The distribution of β_1 under the null hypothesis is a Student's t distribution, with n-2 degrees of freedom when the regression is performed on n data points. The t statistic is the value of β_1 calculated using the least squares algorithm divided by its standard error. This standard error can be estimated from the data sample by the equation

$$\frac{\sqrt{\sum_{i=1}^{n}[y[i]-(\beta_0+\beta_1 x[i])]^2}}{(n-2)\sqrt{\sum_{i=1}^{n}[x[i]-\bar{x}]^2}} \tag{3.8}$$

The p-value for this t statistic can then be used in a test of statistical significance for the association between phenotype and the genotyped SNP [7,8,10].

In this investigation the Wald test is performed for all genotyped SNPs in a GWAS, and the SNPS are ranked in order of increasing p-value, so that SNPS most

likely to be truly associated with phenotype and placed first and SNPs that appear not to be associated with phenotype are placed last. Case/control studies are adapted to enable the use of the Wald test by setting the phenotype to be 1.0 for cases and -1.0 for controls, thus converting the binary phenotype to a quantitative one so that a least squares regression model may be fitted. While it is rare to analyze case/control studies in this way, it is done here so that a single consistent technique is used for SNP ranking in all GWAS data sets considered in the current investigation, regardless of whether the original phenotype is quantitative or binary.

3.3 Multivariate Feature Ranking: Support Vector Machines

A popular multivariate machine learning technique is the support vector machine, or SVM. SVMs are most commonly used for classification tasks, such as modeling case/control GWAS with binary phenotypes, but an SVM regression algorithm for modeling quantitative outcomes exists and is used in this investigation. SVMs are flexible and powerful, and a large body of mathematical theory exists concerning them. The current discussion is limited to a brief overview of the SVM regression technique.

Least squares linear regression, as stated previously, minimizes the RSS, the sum of squared residuals, where a residual is the difference between the predicted and actual values of the dependent variable. In SVM regression this is replaced by the sum of error terms, with this term defined as 0 if the absolute value of the residual is below a threshold ε, and as the absolute value of the residual minus ε if the absolute value exceeds ε. The contribution of the error is therefore a linear rather than a quadratic function of the residual, and there is a "grace zone" where the residual does not contribute to the error at all if it is small enough [9,10,12].

The other prominent feature of the SVM method is that, in addition to minimizing the error, it also attempts to minimize the norm of $\vec{\boldsymbol{\beta}}$, written $\left\|\vec{\boldsymbol{\beta}}\right\|^2$. This is interpreted as forming a simpler model, that ignores some of the variation in the training set data. This trade-off between reducing error and reducing model complexity can ideally be used to create a regression model that captures the essential "signal" of the training set, such as the true association of phenotype to a SNP, while ignoring the random "noise" of false associations.

The defining equations that must be satisfied are as follows. The function

$$\frac{1}{2}\left\|\vec{\boldsymbol{\beta}}\right\|^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\xi_i^+ + \xi_i^-) \tag{3.9}$$

must be minimized with respect to $\vec{\boldsymbol{\beta}}$, subject to the constraints

$$y[i] - \mathbf{X}[i,*]\vec{\beta} \le \varepsilon + \xi_i^+ \tag{3.10}$$

$$\mathbf{X}[i,*]\vec{\beta} - y[i] \le \varepsilon + \xi_i^- \tag{3.11}$$

where all ξ_i^+ and ξ_i^- are greater than or equal to 0 [9,10,12].

The ξ_i^+ and ξ_i^- represent errors, where the model predicted $\hat{y}[i]$ is different from the actual y[i] be more than ε. The first term in equation (9) is one-half the norm of the $\vec{\boldsymbol{\beta}}$ vector that defines the actual linear regression model, while the second term is the sum of the errors when using that $\vec{\boldsymbol{\beta}}$ vector to predict the training set. The C constant is

predetermined and defines the relative weight for the trade-off between model simplicity and model error.

A mathematical transformation allows this problem to be recast in what is called the dual form, which allows the solution for $\vec{\boldsymbol{\beta}}$ to be defined in terms of a linear combination of a subset of the subject row vectors in **X**. These row vectors are called "support vectors" because they "support" the solution, and hence the method determining and using them for regression modeling is the "support vector machine." [9,10,12]

Once the $\vec{\boldsymbol{\beta}}$ vector is determined, it can be used to rank features in order of importance. This is done simply by ordering each feature by the absolute value of the corresponding coefficient in $\vec{\boldsymbol{\beta}}$ in decreasing order. This allows a SVM regression model to be used as a SNP selection method in GWAS. [13,14]

Because SVM is a multivariate method, it has some advantages over a univariate method. Unlike univariate methods, it can take into account redundancy between features. This is similar to the fashion in which least squares linear regression will determine the coefficient β_i for the ith feature in a training set in the context of all other features, producing a different coefficient than the least squares algorithm produces when only considering that single feature. Another advantage of SVM, not shared by ordinary least squares regression, is that it attempts to simplify the model by minimizing the norm of $\vec{\boldsymbol{\beta}}$, which can only be reasonably done by trading off the contributions of different features, an inherently multivariate process.[9,10,12]

However SVM, like all multivariate methods, also has drawbacks. If the number of features equals or exceeds the number of subjects -- if **X** has as many or more columns than rows -- a multivariate method can exactly fit the training data, but at the expense of

generalization to new data. In effect it fits the "noise" as well as the "signal" of the training data. The minimization of $\vec{\beta}$ helps overcome this effect, but if the feature count is very large, it is difficult to find a truly optimum $\vec{\beta}$ amidst all the possible solutions. Computation cost also increases with feature count, and for GWAS with hundreds of thousands of SNP features it can be extremely difficult to find an SVM solution with limited time and computing resources.

3.4 Model Bias, Variance, and Cross-Validation

Predictive performance of least squares regression and SVM models, along with that of other machine learning techniques, can be examined in a more abstract sense.

Machine learning methods analyze a set of data in order to discover a model which accurately reflects the process that created the data. Most methods begin with an assumed general form for the model that includes a number of unspecified parameters. The specific model is then generated by analyzing the training data to determine estimates for the value of these parameters. For example, least squares linear regression assumes that the dependent variable is determined by a linear function of the independent variables plus a random, normally distributed error. The regression algorithm analyzes the training data to estimate a coefficient for each independent variable in this linear function -- these coefficients constitute the $\vec{\beta}$ vector that the algorithm generates. Other machine learning techniques assume other kinds of underlying models and therefore estimate different parameters.

When finding an optimum model, two contrary tendencies must be taken into account. If the assumed model is too simple, it will not be able to accurately mirror the

actual underlying process which generates the data. For example, if a dependent variable is actually determined by a non-linear function of the independent variables, then a linear model will not fully capture this relationship. Similarly, if not all of the important independent variables are captured in the model, it will predict badly even if the underlying process really is linear. This shortcoming is referred to as bias; the model is making simplifying assumptions about the underlying process which do not accurately reflect its true complexity. To avoid bias, a more complex and flexible general model with more parameters can be chosen, so that the model is powerful enough to accurately represent the underlying process.

On the other hand, a complex and flexible model with many parameters is more susceptible to capturing random associations in the data along with the true associations reflecting the underlying process. Such a model can fit the original training data extremely well, but will badly predict any new test data. It has learned to model the "noise" as well as "signal" in the training set, and so generalizes badly precisely because it captures the specifics of one particular set of data too well. This shortcoming is called variance; the model is generating an overly complex model which does not best capture the truly significant patterns in the data.

Finding a good model means finding an optimal trade-off between bias and variance. This was mentioned previously in the context of the C parameter in equation (3.9) defining the SVM model. A higher value of C emphasizes better fitting the training data at the expense of simplicity of the model. This is decreasing the bias -- allowing the model to be more complicated -- but potentially increasing the variance by modeling the random specifics of the training data. Lowering the value of C increases the bias by

simplifying the model, but can decrease the variance as the simpler model is less susceptible to fitting random noise.

Another factor in the bias vs. variance balance is the number of features used to build the model. Adding more features to a model can decrease bias by giving the model more power to capture the actual process underlying the data, but it can also increase the variance by making the model too easily influenced by random associations. The current investigation addresses this issue by creating a sequence of regression models based on an increasing number of SNP features, thereby examining a wide range of possibilities in the bias-variance trade-off.

An important consequence of these concerns is that a model's performance should never be tested on the data set used to build the model. In order to truly test how well a model generalizes beyond the data it is from which it is built, the model should be generated from a "training set" of data and then tested for performance against a completely separate "test set" of data, which is totally uninvolved in the model generation process. In order to maximize the use of the available data, the full available data set is usually randomly divided to create the training and test sets. This division, along with the subsequent model generation and testing procedure, is then repeated several times to get a robust estimate of performance.

One popular technique for this, which is used in the current investigation, is cross-validation. The data set is divided into N equal parts. All but one of these parts are combined into a training set that is used to generate the model. The model's performance is then tested on the omitted part, which forms the test set. This procedure is repeated N times, with a different part segregated as the test set each time. At the end of cross-

validation, each subject in the data has been predicted as part of a test set exactly once, and has been part of a training set N-1 times. Each model is built using a fraction of N/(N-1) of the complete data set. This whole process is called "N-fold" cross-validation. The current investigation uses 10-fold cross-validation for all data sets. [9,10]

3.5 Model selection: Nested Subsets of Ranked Features

Given that many different models, using different algorithms and feature sets, can be generated, choosing the best model can be a complicated process. Even if only one general modeling technique is used, the number of possible sets of features that can be chosen increases exponentially with the number of features available. For GWAS with many thousands of SNP features, it is computationally impossible to build models for all possible combinations of SNPs.

Various strategies exist for searching the space of all possible feature sets for modeling. For example, in forward selection a model is built and tested for each available feature in isolation, creating N univariate models for N available features. The feature for the most predictive model is chosen. Then N-1 models are built and tested, with each model being based on two features: the feature selected from the previous step along with one of the N-1 remaining features. From this round of testing, the additional feature associated with the best model is chosen and put in a "chosen set" with the first selected feature. The process is then repeated using these two features in combination with each of the N-2 remaining features, allowing selection of a third chosen feature. This process is repeated, adding a new feature to the chosen set each time, until the model performance ceases to improve more than some predefined criterion. The model based on the chosen set is then selected as the optimum model.

Backward selection is similar but works in the opposite direction. A single initial model is built including all available features. Then N models are built with all the features except for a single feature which is left out, and the omitted feature for the model that has the least decrease in performance is removed from consideration. The process is iterated, removing one feature at a time, until the removal of any feature leads to an unacceptable decrease in model performance. The model based on the set of features left at this point is selected as the optimum model. [9,10]

An even simpler method is used in the current investigation, which is not seeking the best possible model, but instead using models to compare feature ranking techniques. The feature ranker, a machine learning algorithm of some sort, is used to assign a statistical significance value to each of the N features. The features are then placed in a list rank ordered by decreasing significance. Then a set of N models are built, with each model including the i best ranked features where i ranges from 1 to N. This forms a sequence of models built on nested subsets of features. Each model contains all the features of the previous model plus one additional feature, the next one in rank order. This technique is simple to understand and computationally feasible even for many features, but its success depends on the ability of the ordering algorithm to correctly rank features by their significance. [13]

A particular caution must be taken with feature selection. Many models are built on a training set and then used to predict a test set. The model with the best predictive performance is then chosen. However this process is actually a form of training, because choosing which of the many possible feature sets generates the best model is a form of model parameter fitting. Since the test set was used in this process, it effectively

becomes a kind of adjunct training set, violating the prohibition that models should not be trained on test data. In order to determine the true performance of the selected feature set, the model generated from it must be used to predict another, completely novel test set of data that was not involved in the feature selection process. In the current investigation, this final step is not taken, so no conclusions about the absolute performance of the best model can be drawn. However the relative performance of the models generated, especially when examined as a trend rather than a single result, can still provide insight into the best strategies for feature selection. The current investigation takes advantage of this to evaluate the relative performance of different feature ranking algorithms. [10]

CHAPTER 4

METHODS AND DATA

4.1 Purpose of the Investigation

The current investigation explores the relative performance of three kinds of feature ranking methods: the univariate Wald test ranking of all SNPs, the multivariate SVM regression ranking of all SNPs, and a hybrid method in which a number of the best ranked SNPs from the Wald test are used as training input to the SVM, which then produces a ranking of just this subset of SNPs. The interesting hypothesis examined is that the hybrid technique of using a univariate method to filter a subset of SNPs, which are then ranked by a multivariate method, may capture the best properties of both methods. [14] The univariate filter allows very large numbers of SNPs to be ranked with reasonable computational resources, and since each SNP is analyzed independently, the test for association cannot be degraded by trying to process too many features at once. Then the best results of this filter are passed to a multivariate method, which can accommodate redundancy between SNPs and can also properly balance the bias vs. variance trade-off to generate a superior ranking of this subset of SNPs. Because the multivariate method only sees a limited number of filtered SNPs, it is computed faster, and is less likely to be overwhelmed by too many features, compared to SVM run on all genotyped SNPs.

4.2 Data Sets

Several different data sets are used in this investigation. They are briefly described here and summary information about subject and SNP counts are given in table 4.1. Note that

for all case/control studies, a phenotype of 1.0 is assigned to cases and -1.0 is assigned to controls, so that they can be treated as quantitative phenotypes by the univariate Wald test and multivariate least squares regression models. All data sets, except for one simulated set, are taken from publicly available GWAS data.

One master data set is taken from a study genotyping a heterogeneous stock of laboratory mice, which were bred by mixing 8 inbred strains to increase genetic diversity. In addition to genotyping, multiple quantitative phenotypes were measured. This investigation uses two of these phenotypes: mean cellular hemoglobin (MCH), a measurement of hemoglobin content in red blood cells, and percentage of CD8 cells (CD8), the relative proportion of a type of immune system lymphocyte. [4]

A simulated case/control data set is generated used the freely available GWASimulator software. A set of 15 "causal SNPs" are each assigned an odds ratio of 2.0 per minor allele for increasing the base risk of disease case status. Using these parameters, 1000 case subjects and 1000 control subjects are generated, using the defined odds ratios for causal SNPs to generate these subjects' genotypes according to the corresponding disease risk probability distribution. For each causal SNP except one, 2000 non-causal SNPs are simulated, with the causal SNP in the precise middle of a run of consecutive SNPs. The remaining SNP is near the end of a chromosome, so only 1187 non-causal SNPs are simulated around it.

GWASimulator simulates the linkage in each run of consecutive SNPs according to a chromosomal profile based on the International HapMap Project findings. This produces simulated GWAS data with realistic SNP linkage characteristics. Unlike most real GWAS data, in this simulated data the identities and effects of the variants causing

disease status are precisely known. This allows SNP selection methods to be evaluated directly, as well as indirectly via phenotype prediction performance. [15]

Three case/control data sets are taken from GWAS conducted on humans, with different types of cancer as the phenotype which is present or absent. The disease types are breast cancer [16], pancreatic cancer [17], and prostate cancer [18]. These data sets are genotyped for many more SNPs than the mouse and simulated data sets, preventing the use of SVM with all of the SNPs as input features owing to computational constraints.

Table 4.1 Subject and SNP Counts for Data Sets

Data Set	# of Subjects	# of SNPS (pre-filtering)
Mouse	1591	12545
Mouse CD8	1521	12545
Simulated	2000	28498
Breast Cancer	2287	513363
Pancreatic Cancer	3937	538099
Prostate Cancer	2300	317503

Certain data clean-up and formatting operations are initially applied to all data sets. The plink software package [6,7], specifically designed for managing and analyzing GWAS data, is used. Initially, the data set is filtered. SNPs with missing data in more than 1% of subjects are omitted. SNPs with the same genotype for all subjects are also excluded as no association with phenotype can be determined in this case.

Next, the data is analyzed to determine the counts of each genotype (0, 1, or 2 minor alleles) for each SNP over all subjects. Then any missing genotype data is

replaced with the most frequently occurring genotype, a very simple kind of data imputation. This step, along with the initial exclusion of SNPs with a more than one percent missing data, generates a data set where all genotypes are defined for all subjects with minimal perturbation of the original statistical structure of the data.

Finally, now that the data set is properly filtered and adjusted, it is output in two alternate formats: a binary format specific to plink, which increases performance for Wald testing, and a text format where the genotypes are numerically encoded by the number of minor alleles so that the SVM and linear regression software packages can model them.

4.3 Analysis Procedure

All data sets are modeled and tested using 10-fold cross-validation. SNP ranking and least squares regression model building are conducted using only the training set, creating a panel of different regression models for each ranking method and SNP count. All prediction performance testing for these models is done using only the test set. In this way each subject in the data set is predicted once by each type of generated model, with the models built from a training set comprising 90% of the subjects in the data set but not including the subject being predicted.

For each training set a group of SNP rank orderings are generated. First the univariate Wald test is run for all SNPs, and a "WALD_ALL" ranking is generated where SNPS are ordered by increasing p-value -- which is by decreasing statistical significance of association with the phenotype -- as determined by the Wald test. For the mouse and simulated case/control datasets, an SVM ranking is then generated for all SNPS, producing an "SVM_ALL" ranking. The cancer case/control datasets have too many

SNPs for an SVM_ALL ranking to be generated, so this ranking is omitted for these datasets.

Next a set of hybrid Wald-SVM rankings is generated. For the mouse and simulated case/control data sets, the number of Wald ranked SNPS input to the SVM is determined by the number of SNPs with a Bonferroni-corrected p-value less than 0.05, a quantity called "R" in this investigation. (Note that "R" for the number of SNPs passing this p-value filter is not to be confused with "r" for the Pearson's correlation coefficient, which is determined by comparing predicted and actual phenotype values.) A p-value output by the Wald test is Bonferroni-corrected by division by the total number of ranked SNPs, which compensates for the increased probability of type I errors caused by multiple testing. This means that there is only a 5% chance that any of the SNPs in the R top ranked SNPs are falsely associated with the phenotype.

The hybrid Wald-SVM rankings are generated by only using the best ranked N x R SNPs as input to the SVM training set, where N is a small integer. SVM rankings are generated based on the best 2 x R, 5 x R, and 10 x R ranked SNPs from the WALD_ALL ranking. This generates the rankings denoted as SVM_02R, SVM_05R, and SVM_10R.

For the cancer case/control datasets, there are frequently no SNPs with Bonferroni-corrected p-value under the 0.05 cutoff, so that R is zero. Therefore a pre-determined number of best SNPs from the WALD_ALL ranking is input to the SVM. In this way SVM rankings are generated for the best 5, 10, 50, 100, 500, 1000, 5000, and 10000 Wald ranked SNPs.

Once all the rankings are generated, a sequence of least squares linear regression models is generated for each ranking. Each model is built from the first n SNPs in the

ranking, where n is set in turn to each integer from 1 up to N. Here N is the maximum of the total number of SNPs in the ranking or 1000. Each model therefore includes all the SNPs used in models previous to it in the sequence, plus the next SNP in order from the ranking. For each of these models, all subjects in the test set are predicted, and the Pearson's correlation coefficient r is calculated for the correlation between predicted and actual phenotype values of the test set. The value of r for each SNP count is saved.

The average of the correlation r for each SNP count, for each ranking method, is calculated over all ten cross-validation runs and plotted. Statistics for the mean, standard deviation, maximum, and minimum over all ten cross-validation runs are calculated for the maximum r correlation value obtained by each ranking method.

CHAPTER 5

RESULTS

5.1 Mouse Data

The mouse data sets provide the best test bed for this investigation. The phenotypes are truly quantitative, not just "quantified" case/control studies, so the linear regression and SVM methods are highly appropriate to model them. The heritability of MCH is estimated at 0.55 and that of CD8 at 0.99 [4], which means there is a significant genetic component to these traits. The plots reveal the trends of the various ranking methods quite clearly.

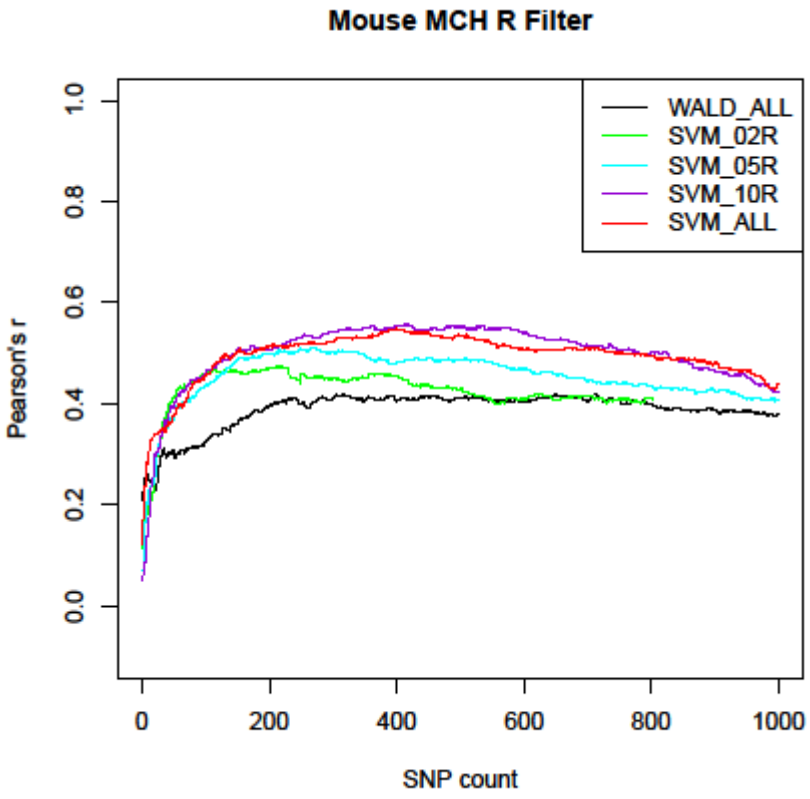

Figure 5.1 Plot of Pearson's correlation coefficient for predictions on mouse MCH data as a function of number of best ranked SNPs included in the regression model.

In Figure 5.1 it can be seen that the Wald ranking performance is substantially worse than any of the SVM ranking. SVM_02R is somewhat better, SVM_05R better than that, and SVM_10R and SVM_ALL are about evenly the best overall. Also of note is the rate of prediction performance increase as the SNP count increases. For the first few SNPs, all rankings gain predictivity very quickly; presumably this represents a few highly significant SNPs that are easily found and best ranked by all methods. After that, the Wald ranking "levels off" and only slowly increases performance as more SNPs are added. All the SVM methods, on the other hand, continue to gain predictivity quickly until eventually reaching a maximum. Since a primary purpose of GWAS is to identify a small set of SNPs for further investigation, in order to aid in the discovery of underlying biological mechanisms, this ability of the SVM methods to place a relatively small set of predictive SNPs at the top of the ranking is a useful property.

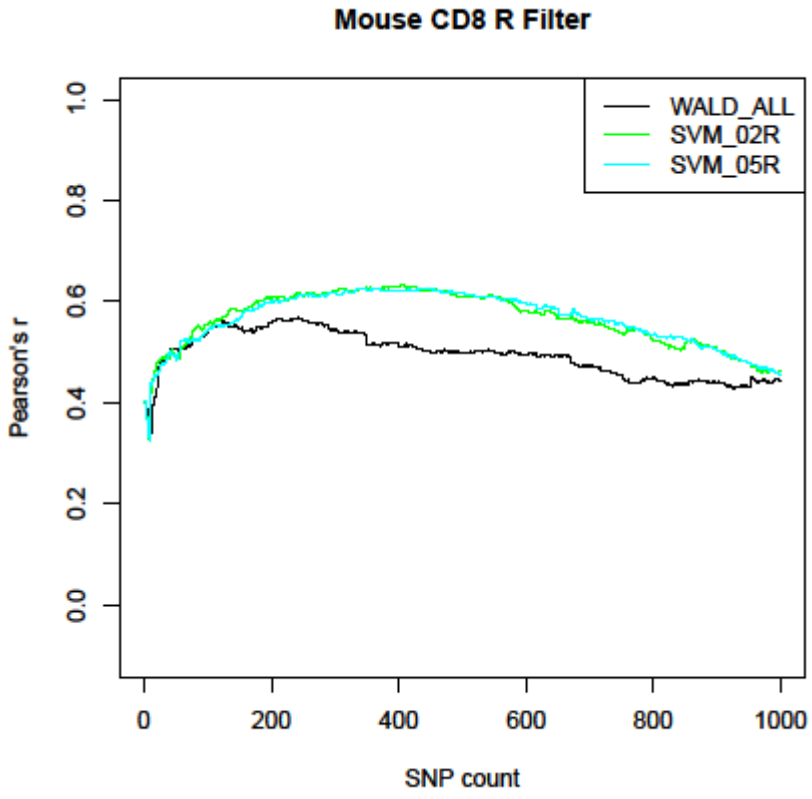

Figure 5.2 Plot of Pearson's correlation coefficient for predictions on mouse CD8 data as a function of number of best ranked SNPs included in the regression model.

In Figure 5.2 the SVM_10R and SVM_ALL ranks are missing. This is because the "R" value is very high, so much so that SVM_05R is in fact SVM_ALL: fewer than 5 x R SNPs are genotyped. Therefore SVM_10R and SVM_ALL would be redundant with SVM_05R. The signal is quite strong, with very many contributing SNPs. Similarly to MCH, we see that the SVM rankings perform better than the Wald ranking. However the difference is not as significant as for MCH. Also, the rapid "early rise" of the predictivity curve is not as strongly differentiated between Wald and SVM rankings as for MCH. It appears that the CD8 signal is strong enough that all the rankings can usefully identify significant SNPs.

5.2 Simulated Data

The simulated data set demonstrates what happens when a particularly strong signal is contributed by a relatively small number of SNPs.

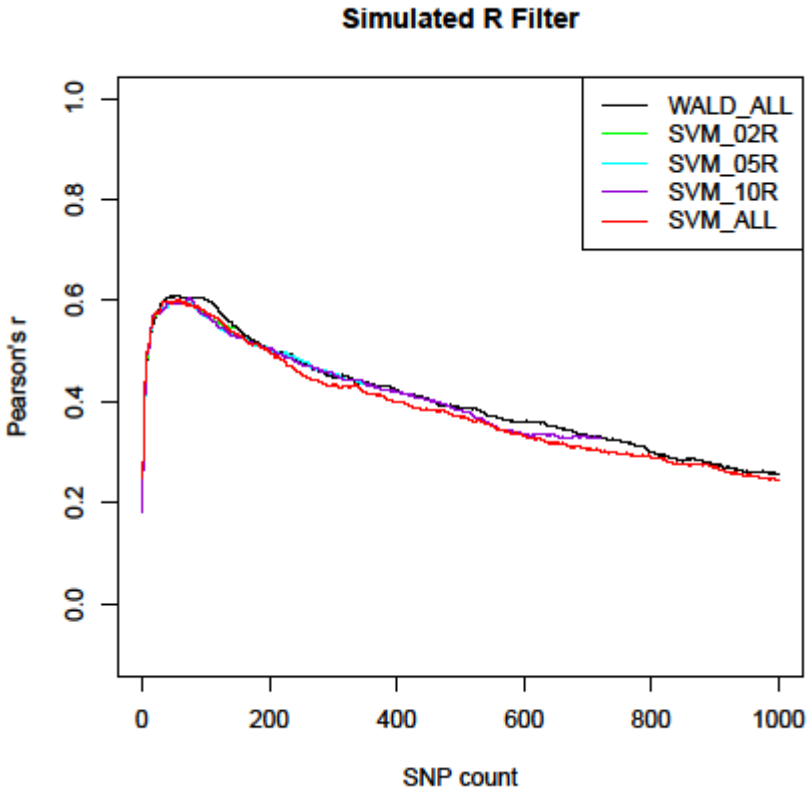

Figure 5.3 Plot of Pearson's correlation coefficient for predictions on simulated case/control data as a function of number of best ranked SNPs included in the regression model.

All rankings very quickly find all "causal SNPs" as well as neighboring SNPs that are strongly linked to the causal SNPs. The Wald test very slightly outperforms the SVM methods, but not significantly. The SVM methods are so similar that it is difficult to discriminate them in the plot. The SVM_02R curve is plotted, but is basically overwritten by the other SVM curves and is thus invisible. The SVM_05R and SVKM_10R plots are tightly intertwined and can hardly be discriminated.

It should be noted that the underlying data process in this data set is in perfect accord with the model assumptions that the machine learning techniques implicitly apply. There are no interactions between SNPs, and the effect of each SNP is perfectly modeled as an odds ratio multiplied by the number of minor alleles. This fact, combined with the small number of causal SNPs and the relatively high odds ratio for each one, makes this data set very easy to correctly analyze for all techniques. Inspection of the actual SNPs rankings reveals that all highly ranked SNPs are indeed causal or near neighbors of causal SNPs – a finding that can only be made for a simulated data set where the true underlying process is known ahead of time.

5.3 Cancer Data

The three cancer data sets genotype far more SNPs, and include more subjects, than the mouse data. However the genetic signal is very small. The initial GWAS that collected this data identified a few significant SNPs, but could only confirm them using additional data sets from other studies. The results of the current investigation with these data sets are on the edge of significance.

Because the R values for these data sets are almost always zero, it is not plausible to use multiples of R as SNP counts for creating SNP subsets for the SVM methods. An alternate method is used where fixed number of best ranked SNPs from the Wald test are used to train the SVM models. The results are displayed in two plots for each data set: one plot for the smaller SNP counts of 5, 10, 50, and 100, and another plot for the larger SNP counts of 500, 1,000, 5,000, and 10,000. The Wald test ranking results of displayed in both plots. SVM models for all SNPs could not be built as the total SNP count for

these studies was too large to be processed by the SVM with available computing resources.

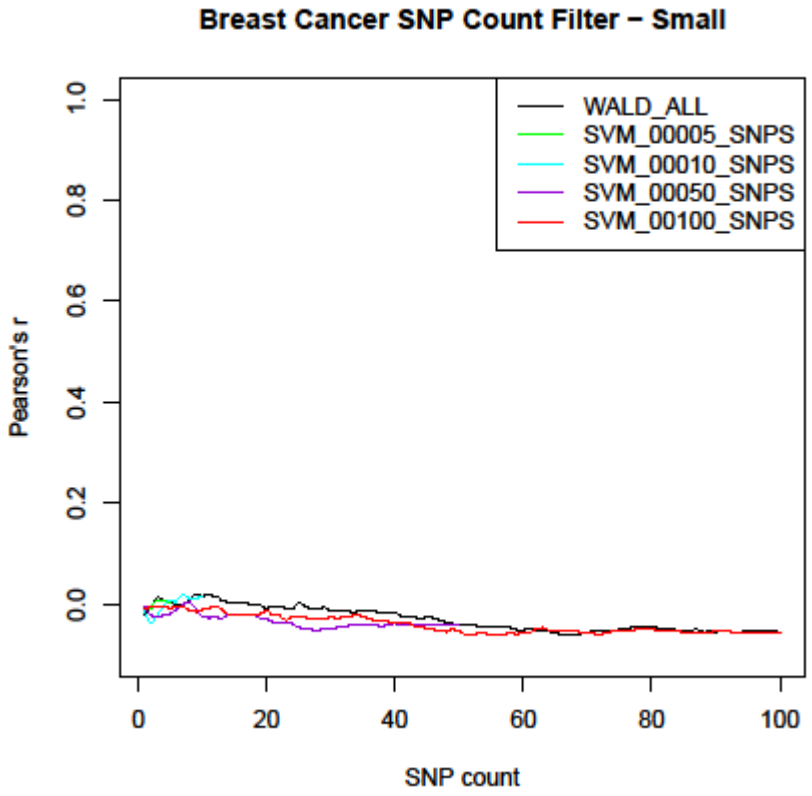

Figure 5.4 Plot of Pearson's correlation coefficient for predictions on breast cancer case/control data as a function of number of best ranked SNPs included in the regression model, for small numbers of SVM ranked SNPs and Wald ranked SNPs

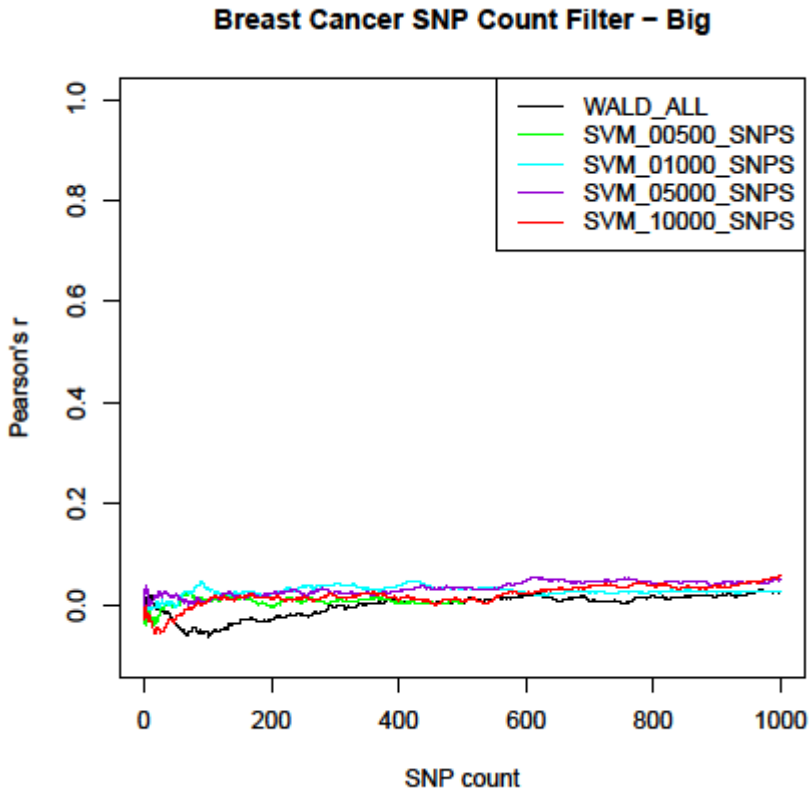

Figure 5.5 Plot of Pearson's correlation coefficient for predictions on breast cancer case/control data as a function of number of best ranked SNPs included in the regression model, for large numbers of SVM ranked SNPs and Wald ranked SNPs.

For the breast cancer data set, all ranking methods have essentially no predictive power. While looking at the plots may create the illusion that the SVM methods predict better for large numbers of input SNPs, the statistical variation between individual cross-validation trials shows this to be an unjustifiable conclusion.

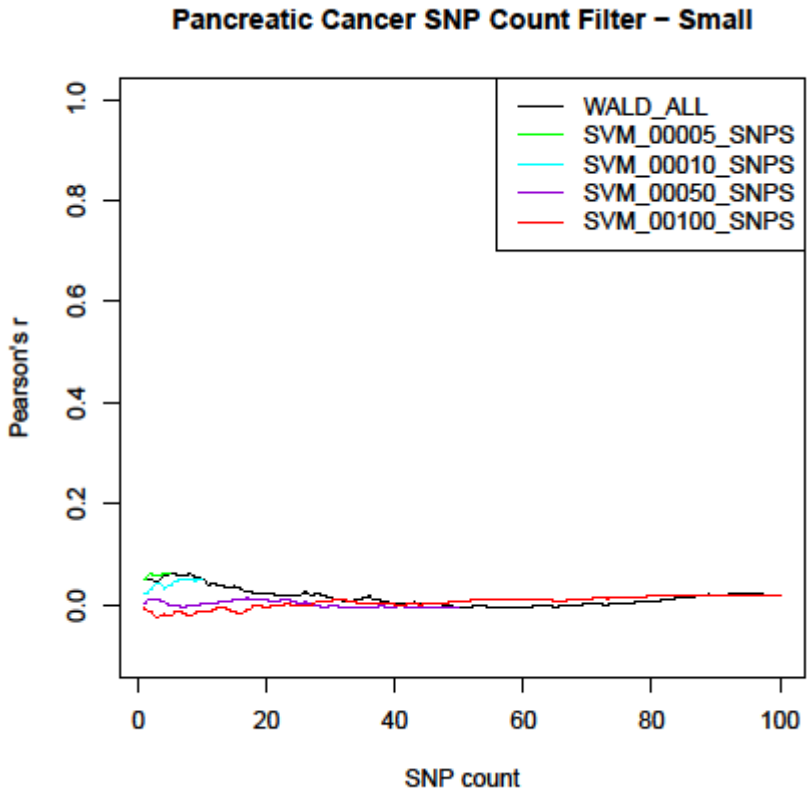

Figure 5.6 Plot of Pearson's correlation coefficient for predictions on pancreatic cancer case/control data as a function of number of best ranked SNPs included in the regression model, for small numbers of SVM ranked SNPs and Wald ranked SNPs.

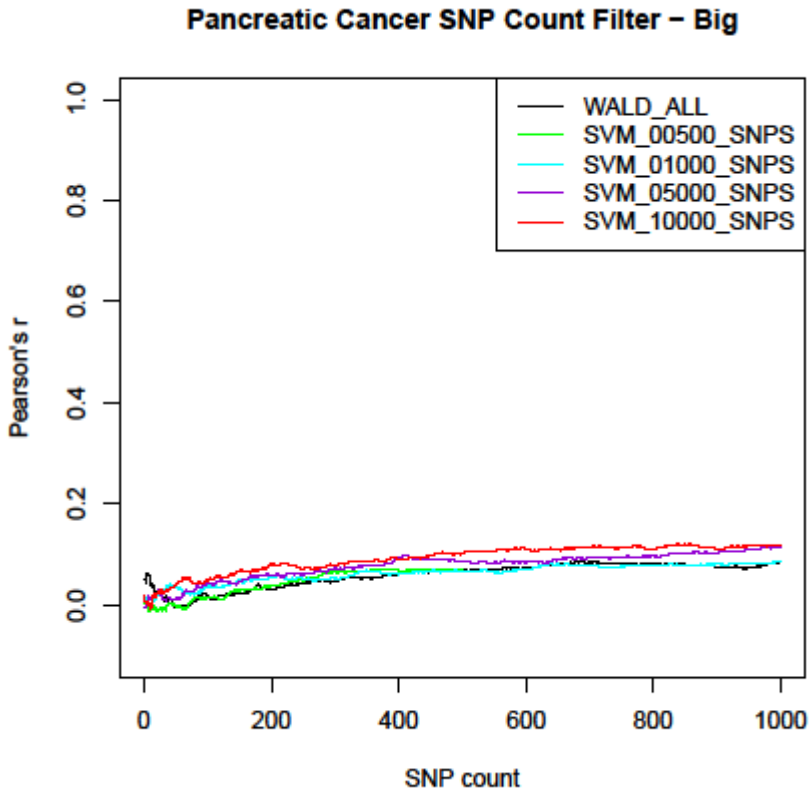

Figure 5.7 Plot of Pearson's correlation coefficient for predictions on pancreatic cancer case/control data as a function of number of best ranked SNPs included in the regression model, for large numbers of SVM ranked SNPs and Wald ranked SNPs.

For pancreatic cancer the predictivity is still very small, but a bit of a trend can be discerned. There is a small predictive power in the first ten or twenty best ranked SNPs from the Wald test. SVM offers no benefit when applied using these few SNPs. After this initial signal, predictivity decreases towards zero as more SNPs are added, confusing the initial "signal" with "noise." However when the SNP counts increase, additional signal starts to appear again, as if a large quantity of SNPs each with very small contribution to disease are acting in concert. Both Wald and SVM rankings with many SNPs show a slight increase as the number of SNPs provided to linear regression

increases. When five or ten thousand SNPs are input to the SVM ranking, it appears to provide a small advantage over the Wald test.

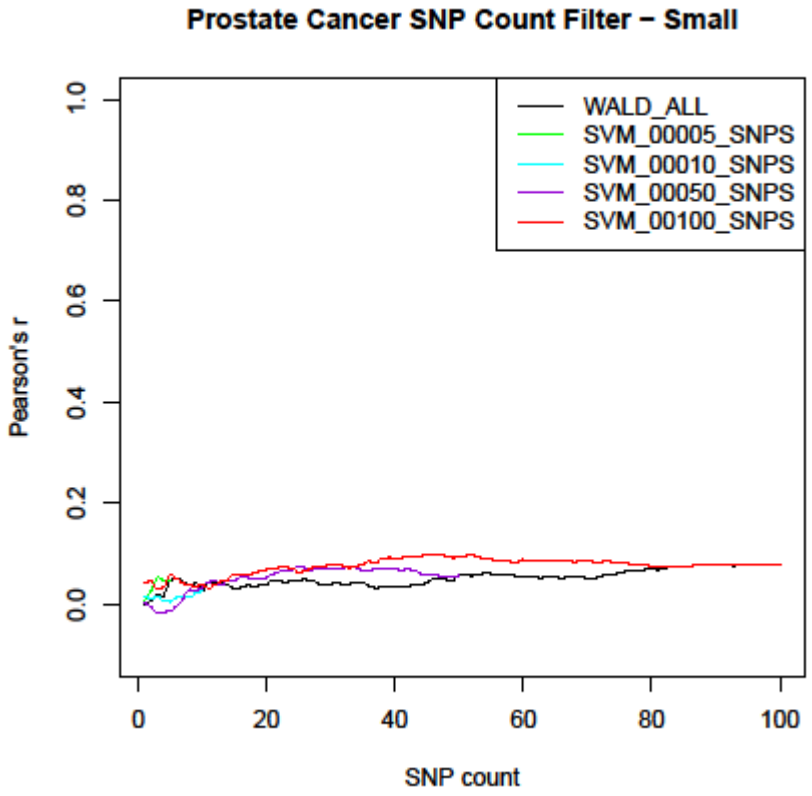

Figure 5.8 Plot of Pearson's correlation coefficient for predictions on prostate cancer case/control data as a function of number of best ranked SNPs included in the regression model, for small numbers of SVM ranked SNPs and Wald ranked SNPs.

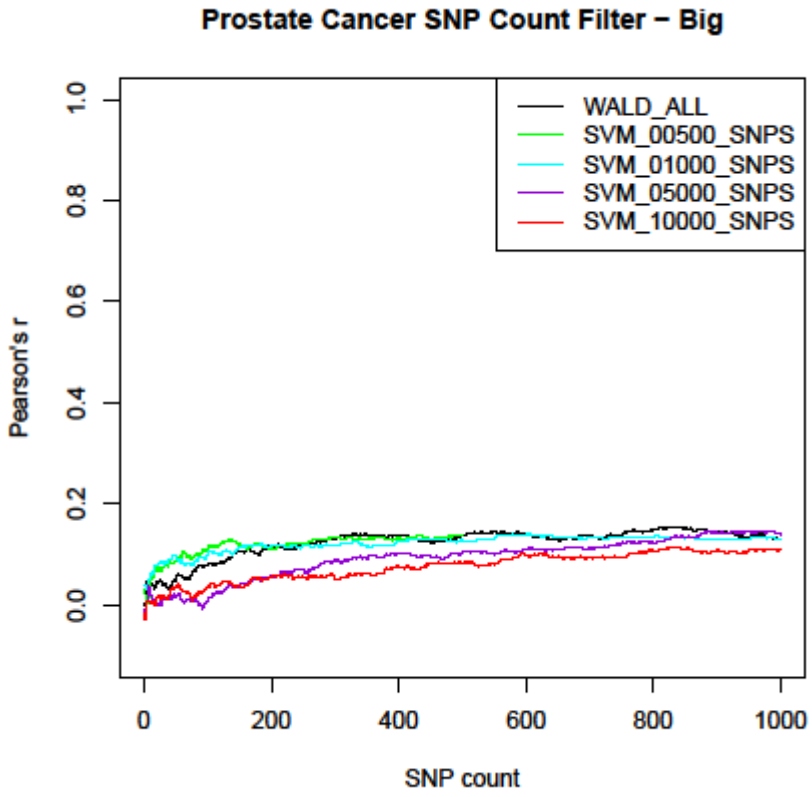

Figure 5.9 Plot of Pearson's correlation coefficient for predictions on prostate cancer case/control data as a function of number of best ranked SNPs included in the regression model, for large numbers of SVM ranked SNPs and Wald ranked SNPs.

The prostate cancer data provides the strongest signal of any of the cancer data, although one that is still very small. In this case the Wald test has the best overall performance, with all of the SVM rankings appearing to diminish predictivity overall. However for small numbers of SNPs -- from about 50 to 1,000 SNPs as input to SVM -- the SVM rankings initially perform better than the Wald ranking, so for identification of a small number of most strongly associated SNPs the SVM method could provide an advantage.

CHAPTER 6

DISCUSSION

6.1 Findings

This investigation shows that multivariate SVM ranking can definitely improve on univariate Wald ranking in many circumstances. However, this is not universally true, as some data sets and ranges of SNPs under consideration are better predicted with the Wald ranking. The behavior of different data sets varies widely, reflecting the complexity of the genetic architecture influencing phenotypes. The simulated data produces clear, easily comprehensible results, and all ranking methods were essentially equal in discovering the statistical patterns in the data set. The data from actual GWAS is much more problematic. The best ranking algorithm varies almost on a case by case basis.

Often a trade-off must be made between achieving the maximum possible overall predictivity of phenotype and identifying a small number of SNPs with the most effect. The most common use of GWAS is for the latter purpose, so methods that rank a few good SNPs highly should receive attention even if their overall predictivity is less than optimal. This investigation unfortunately cannot offer a definite prescription for achieving the desired end in GWAS, but it does show that multivariate methods such as SVM are well worth trying. The hybrid method using a univariate filter with a multivariate ranking can sometimes provide advantage in the identification of a few most significant SNPs, but the filtering does not seem to improve overall predictivity compared to running SVM on all available SNPs. However when the GWAS genotypes a very large number of SNPs this is computationally infeasible, so using a univariate filter can enable the use of SVM for as many SNPs as it can handle.

In addition to examining the ranking methods themselves, this investigation provides insight into issues with least squares linear regression applied to SNPs with a high degree of linkage. The issues arising from different algorithms used to form a least squares model when some independent variables are perfectly correlated are subtle. Much time was expended in this investigation discovering, understanding, and resolving this phenomenon. Many discussions of least squares regression in the literature do not examine these issues in depth, because most statistical data sets are set up to only use independent variables that are very unlikely to be perfectly correlated. GWAS are an exceptional case where a very large quantity of SNP features, each being described by one of a few discrete values (0, 1, or 2 minor allele count), can have high correlation due to linkage. The inability of the very popular statistical programming language R to adequately handle this issue in its default least squares functionality proves that GWAS investigators need to be especially aware of limitations and assumptions of the software they choose, as the data sets they examine have distinctive properties not found in many other statistical investigations.

6.2 Future Directions

As in all research projects, very many interesting avenues of investigation could not be pursued in the time available for the current investigation. Far more research is called for to provide truly meaningless assessment of univariate and multivariate machine learning tools applied to GWAS data. Some possibilities for additional research are suggested here.

It is standard when applying SVM methods to try several different values of the C factor from equation 3.9, in order to determine the best trade-off between model bias and

variance. [9,10,12] Time did not allow for this basic procedure to be applied in the current investigation. Instead, a single C factor is chosen that allowed the SVM to complete in a reasonable amount of time, using the default value set by the SVM light software as a rough guide. The cancer data sets require a different value of C than the mouse and simulated data sets in order for the SVM model to finish in reasonable time, but no meaningless investigation of the appropriate value for this parameter is performed. Further investigation of the effect of different values of C is recommended.

SVM in the current investigation is only used for ranking, with all actual prediction carried out by the least squares linear regression technique. SVM is primarily known as a predictive modeling method, not a feature ranker, so comparison of actual predictions from SVM models to the least squares results would be of interest.

The Wald test and linear regression are techniques designed for quantitative outcomes which vary continuously. Their application to case/control studies with binary outcomes certainly works well, but other techniques are generally preferred for modeling such data. Logistic regression is a technique related to linear regression but modified to produce probabilities for binary outcomes. [9,10]

In addition to SVM, many other multivariate machine learning techniques are available which provide control over the trade-off of bias vs. variance. Techniques of regularized regression augment the least squares linear or logistic regression algorithms with penalties on model complexity that can be adjusted. Examples include ridge regression, lasso, and elastic net techniques. Lasso in particular is optimized for identifying a small subset of highly predictive features. [10] SNP ranking using some of these techniques should be compared with SVM ranking.

The current investigation must be brought to a close, but hopefully the software and methods developed for it, and its ambiguous but intriguing findings, can provide a platform and starting point for future research.

CHAPTER 7

CONCLUSIONS

GWAS are a major innovation in the study of genetic causes of disease and other phenotypes, but the very power of this data collection technique makes accurate analysis of the data difficult. Innovation in genotyping methods continues at an extremely rapid pace, and already it is expected that complete sequencing of subject genomes will soon replace SNP-based techniques in many studies. In this event issues such as less than perfect linkage between causal variants and SNPs, and the effect of rare variants not detected by a panel of SNPs identifying standard haplotypes, will cease to be a concern. Causal variants of any type or frequency will be directly genotyped, with no need for SNPs to act as statistical proxies for the true biological associations. However this vast increase in detailed genetic data will not obviate the essential statistical issues involved in associating genotype and phenotype, but rather magnify them in proportion to the increase in raw data available.

Therefore, even if traditional SNP-based GWAS are on their way to becoming obsolete, the data analysis issues examined in the current investigation will not follow them into history. Assessment of the range of machine learning techniques and their application to the particularly complex field of genetic studies will only increase in importance. It is hoped that the current investigation will be a very small contribution toward this important and fascinating research.

REFERENCES

1. Feero WG, Guttmacher A (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 2010; 363: 166-176.
2. Curtis H (1979) Biology, 3rd edition. New York: Worth Publishers, Inc. 1065 p.
3. Lesk A (2008) Introduction to bioinformatics, 3rd edition. New York: Oxford University Press Inc. 432 p.
4. Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet 4(10): e1000231.
5. International HapMap Project (2012) The origin of haplotypes. Available: http://hapmap.ncbi.nlm.nih.gov/originhaplotype.html.en. Accessed 1 April 2012.
6. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et. al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.
7. Purcell S (2009) PLINK V1.07: Whole genome data analysis toolkit. Available: http://pngu.mgh.harvard.edu/purcell/plink/. Accessed 1 April 2012.
8. Pagano M, Gauvreau K (2000) Principles of biostatistics, 2nd edition. Pacific Grove: Duxbury Thomson Learning.
9. Alpaydin E (2004) Introduction to machine learning. Cambridge: The MIT Press. 445 p.
10. Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning: data mining, inference, and prediction, 2nd edition. New York: Springer. 763 p.
11. Press W, Teukolsky S, Vetterling W, Flannery B (1992) Numerical recipes in C: the art of scientific computing, 2nd edition. Cambridge: Cambridge University Press. 994 p.
12. Smola A, Scholkopf B (2004) A tutorial on support vector regression. Statistics and Computing; Volume 14, Number 3: 199-222.
13. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46: 389-422.
14. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H (2011) Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Res. 2011 May; 39(9): e62.

REFERENCES
(Continued)

15. Li C, Li M (2008) GWAsimulator: A rapid whole genome simulation program. Bioinformatics 24:140-142.

16. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et. al. (2007) A genome-wise association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet. 2007 June; 39(6): 870-874.

17. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, et. al. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet. 2009 Sep; 41(9): 986-990.

18. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, et. al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet. 2007 May; 39(5): 645-649.