5-2015

An Integrated Approach to Exploit Linkage Disequilibrium for Ultra High Dimensional Genome-wide Data

Michelle Carlsen
Utah State University

AN INTEGRATED APPROACH TO EXPLOIT LINKAGE DISEQUILIBRIUM FOR ULTRA HIGH DIMENSIONAL GENOME-WIDE DATA

by

Michelle Carlsen

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Guifang Fu Major Professor	David Brown Committee Member
Daniel Coster Committee Member	

UTAH STATE UNIVERSITY
Logan, Utah

2015

ABSTRACT

An Integrated Approach to Exploit Linkage Disequilibrium

For Ultra High Dimensional Genome-Wide Data

by

Michelle Carlsen, Master of Science

Utah State University, 2015

Major Professor: Dr. Guifang Fu
Department: Mathematics and Statistics

Genome-wide data with millions of single nucleotide polymorphisms (SNPs) can be highly correlated due to linkage disequilibrium (LD). The ultra high dimensionality of big data brings unprecedented challenges to statistical modeling such as noise accumulation, curse of dimensionality, computational burden, spurious correlation, and a processing and storing bottleneck. The traditional statistical approaches lose their power due to $p >> n$ and the complex correlation structure among SNPs. In this article, we propose an integrated DCRR approach to accommodate the ultra high dimensionality, joint polygenic effects of multiple loci, and the complex LD structures. Initially, a distance correlation (DC) screening approach is used to extensively remove noise, after which LD structure is addressed using ridge penalized multiple logistic regression (LRR) model. The false discovery rate, true positive discovery rate, and computational cost were simultaneously assessed through a large number of simulations. The binary trait of *Arabidopsis thaliana*, hypersensitive response to the bacterial elicitor *AvrRpm1*,was analyzed on 84 inbred lines (28 controls and 56 cases) and 216,130 SNPs were analyzed and significant SNPs were detected. Compared to previous SNP discovery methods implemented on the same dataset, the dCRR approach successfully detected the causative SNP while dramatically reducing spurious assocoiations and computational time.

(38 pages)

PUBLIC ABSTRACT

An Integrated Approach to Exploit Linkage Disequilibrium

For Ultra High Dimensional Genome-Wide Data

by

Michelle Carlsen, Master of Science

Utah State University, 2015

Major Professor: Dr. Guifang Fu
Department: Mathematics and Statistics

This paper presents improved methods for analysis of genome-wide association studies in contemporary genetic research. Thanks to current sequencing methods, half to one million single-nucleotide polymorphisms (SNPs) can be feasibly generated within any given population, and there are often correlations among SNPs that cause truly causative loci to be confounded by correlated neighboring loci. Additionally, complex traits are often jointly affected by multiple genetic variants with each having small or moderate individual effects. To address these issues in genome-wide association studies, we propose a novel statistical approach, DCRR, to detect significant associations between large numbers of SNPs and phenotypes. We applied DCRR on simulations of that varied in marker allele frequencies, linkage disequilibrium, and the numbers of SNPs considered; and we analyzed a previously published Arabidopsis thaliana dataset of an AvrRpm1 binary trait. Our distance correlation was effective in ranking SNPs while the logistic ridge regression detected causative SNPs without including spurious correlated neighbors. Our results indicate that DCRR is an effective and reliable method that can improve the accuracy and efficiency of large association datasets.

(38 pages)

ACKNOWLEDGEMENTS

This work is supported by NSF research grant (DMS-1413366).

I would like to thank my incredible advisor Dr. Guifang Fu for the immense amount of work she put into all aspects of this project. The many wonderful teachers I have had during my time at Utah State who have taught me so much. And my marvelous parents who have supported me through it all.

Dedicated to Mildred Mortensen Stromberg

Michelle Carlsen

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Page

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

With recent developments in high-throughput genotyping techniques and dense maps of polymorphic loci within genomes, an ultrahigh dimension of SNPs (typically more than half a million) is increasingly common in contemporary genetics, computational biology, and other fields of research [1, 6, 10, 74, 91]. Despite the fact large-scale genome-wide association studies (GWAS) provide great power to unravel the genetic etiology of complex traits by taking advantage of extremely dense sets of genetic markers [8, 9, 79, 85], they bring concomitant challenges in computational cost, estimation accuracy, statistical inference, and algorithm stability [24, 27]: Firstly, the number of SNPs p, in units of hundreds of thousands or millions, far exceeds the number of observations n, in units of hundreds or thousands. Statistically, when the number of predictors is much larger than the number of samples, commonly referred to as "small n big p", the power of many traditional statistical models is disabled [18, 25]. The unique problems that belong only to ultrahigh dimensional big data, such as storage bottleneck, noise accumulation, spurious correlation, incidental endogeneity, and so on, are pointed out by Fan et al. 2014 [24]. Computationally, the combinatorial explosive search space grows exponentially with the number of predictors, called the "Curse of Dimensionality". Secondly, most complex traits are mediated through multiple genetic variants each conferring a small or moderate marginal effect with low penetrance on the traits, which obscures the individual significance of each variant [65, 75, 87, 88]; Thirdly, multicollinearity grows with dimensionality. As a result, the number and extent of spurious associations between genetic loci and phenotype increases rapidly with increasing p due to non-causal SNPs highly correlated with causative ones [23, 24, 26]

Linkage Disequibrium (LD), the genetic term of nonrandom association of alleles at nearby loci, may be caused by frequent recombination, physically linked genetic variants, population admixture, or genetic drift [5, 15, 16, 30, 31, 34, 61, 66, 72, 82]. LD is one of

the most important, extensive, and widespread features in genomes, with approximately $70-80\%$ of genomes having regions of high LD [15, 30, 61, 80, 82]. Additionally the LD patterns among the whole genome vary, with the average length of 60-200 kb in general populations [50, 61, 82]. Excessive LD may hinder our ability to detect any causative genetic variants truly influencing the phenotype. Strong LD existing among the loci of extremely dense panels means that the correlated SNPs in the vicinity share substantial amounts of information and introduce heterogeneity that can partially mask the effects of other SNPs. Strong LD leads to inflated variance, incorrect statistical inferences, inaccurate tests of significance for the SNP, unstable parameter estimates, diminished significance for the truly influential SNP, and false scientific identifications [7, 11, 14, 69].

Many statistical models have been used to assess the association between genetic variant and phenotype in GWAS. The prevailing strategies of GWAS focus on single-locus models (for example logistic regression with a single SNP as the predictor, Cochran-Armitage test for trend [2], or Fisher's exact test) which assess the potential association of each SNP in isolation from the others [4, 17, 44, 47–49, 60, 63, 73, 86]. Although widely used for its simplicity, the single-locus model is inefficient with limited power because it neglects the combined multiple joint effects of SNPs, inappropriately separates SNPs in LD, fails to differentiate potentially causative from non-causative variants, struggles with multiple correction due to the extremely large number of simultaneous tests involved, and yields both high false-positive and false-negative results [6, 13, 58, 59]. The standard multiple regression approaches, albeit accommodating the joint effects of multiple SNPs and allowing for control of small LD, breaks down when moderate-to-strong LD exists among the SNPs and are infeasible when the number of SNPs is larger than the number of observations [19, 20, 37, 38, 75]. In addition, multiple regression models involve a large number of degrees of freedom and lack parsimony. The conditional logistic regression was proposed to accommodate the LD effects but, it does not allow for the simultaneous quantification of each SNP individually along with the combined effects of other SNPs [89]. Principal component analysis (PCA) or other clustering methods group SNPs according to their LD patterns. However, these approaches may miss the truly causative variant, undervalue the complexity of LD, and not allow for the interpretation of the individual significance of each SNP. The Partial Least Squares (PLS) method has also been used to address the correlation among predictors, but the theoretical properties of PLS (such as mean squared error) have not been established as thoroughly as in other

approaches [28, 43].

Ridge regression (RR) [46], fitting a penalized likelihood with the penalty defined as the sum of the square of each parameter estimate, was used extensively to deal with the situation that the predictors are highly correlated and also handle the situation when the number of predictor exceeds the number of subjects [13, 29, 36, 42, 46, 57, 58, 75, 84, 84, 94]. The RR has been shown to be preferable to the Ordinary Least Square (OLS), PCA, or other approaches in many contexts and achieves the smallest prediction error among a number of regression approaches after head-to-head comparisons [28]. Through several simulations with varied LD strength, allele frequency, and effect size, Malo et al. compared the performance of RR, standard multiple regression, and single-locus regression for the continuous phenotype. They reported that RR performs best in every combination and the advantages of RR are more obvious when the LD strength is strong. They also reported that the single-locus regression was the worst among three approaches because it failed to differentiate causative SNPs from those spurious SNPs that are merely in LD with the causative SNPs. Sun et al. identified a new genetic locus associated with a continuous trait, anti-CCP, by RR that was not detected by single-locus model [75]. Cule et al. extended the test proposed by Halawa and EI Bassiouni [39] and proposed an asymptotic test of significance forRR, and demonstrated that the test is comparable to permutation test but with much reduced computational cost for both continuous and binary phenotype [13].

Though RR is powerful for addressing correlation and multiple joint effects, it is extremely time consuming and is only designed for a moderate number of predictors (i.e. $p > n$ but not $p >> n$). Many approaches that are powerful for moderate dimension are either computationally infeasible or perform no better than random guessing for ultrahigh dimensional data due to noise accumulation;and RR is no exception [21, 24, 41]. The signal-to-noise ratio in GWAS is often very low, with only a small portion of SNPs contributing to the phenotype and the number of non-causative and causative SNPs showing great disparity. In light of these sparsity assumptions, feature screening has proven to be highly effective and pivotal for its speed and accuracy to handle ultrahigh dimensional data [22, 26, 40, 55, 56, 92]. Feature screening forcefully filters a large amount of noise and decreases the original large-scale to moderate-scale, overcomes noise accumulation difficulties, greatly improves estimation accuracy, and dramatically reduces the computational burden. The distance correlation screening approach (DC)

has an additional agreeable theoretical sure screening property: all truly important predictors can be selected with the probability tending to 1 as the sample size diverges to ∞ [56]. Although the feature screening approach is powerful in handling the ultrahigh dimension data, it cannot provide any closer analysis such as parameter estimation and significance tests for each predictor.

In this article, we propose an integrated DCRR approach designed for the case-control cohort whole genome data, with a binary phenotype and a half to one million of SNPs. The DCRR first extensively filters noise with a loose threshold using DC, and then intensively examines the significance of the remaining informative SNPs by ridge penalized multiple logistic regression (LRR). DCRR integrates the benefits of both DC and RR, while avoiding the drawbacks of both approaches [60]. It is computationally efficient, reliable, and flexible, with a goal of accommodating LD between variants at different loci and hence differentiating the causative variants from the spurious variants that are in LD with the causative ones. It quantifies the significance of each SNP individually as well as accounts for the joint effects of all other SNPs in a multivariate sense, and stabilize the parameter estimates in the presence of strong LD and ultrahigh dimension of SNPs in GWAS. The traditional RR involving a $O(np^2+p^3)$ calculation [43], which needs an intractable amount of time when p approaches one million. The DCRR approach that we proposed dramatically decreases the calculation burden to $O(p+n^3)$, with a substantial saving for ultra high dimension $p >> n$ and the computational speed mainly depends on the number of observations rather than the number of SNPs.

We demonstrate that our approach is uniformly and consistently powerful under a wide spectrum of different simulations of minor allele frequency (MAF), LD strenght, and the number of SNPs, while controlling the false discovery rate (FDR) at less than 0.05, via simulation. We compare our approach with the popular single-locus Cochran-Armitage model and the traditional LRR model, and demonstrate that the stronger the LD or the larger the dimension, the better performance of the DCRR approach; which power remains high even for low MAF. To further validate our approach, we analyze a real binary whole genome *Arabidopsis thaliana* data.

CHAPTER 2

Materials and Methods

2.1 Measurement of LD

Consider two biallelic loci in the same chromosome, with A/a representing the alleles of the first loci and B/b representing the alleles of the second loci. These two biallelic loci form four possible haplotype, AB, Ab, aB, and ab. Let $f(A), f(a), f(B)$, and $f(b)$ denote the corresponding allele frequencies, and $f(AB), f(Ab), f(aB)$, and $f(ab)$ denote the corresponding haplotype frequencies. LD, the non-independence structure of the alleles for a pair of polymorphic loci at a population level, is generally measured as $D = f(AB) - f(A)f(B) = f(AB)f(ab) - f(Ab)f(aB)$ [54]. A D value close to zero corresponds to no LD. Although D quantifies how much haplotype frequencies deviate from the equilibrium state, it is highly dependent on allele frequencies and hence difficult to compare for different regions. Therefore, the normalized measure, $D' = D/D_{\max}$ is more widely used by removing the sensitivity to allele frequencies [33, 51, 54, 64], where

$$D_{\max} = \begin{cases} \max\{-f(A)f(B), -f(a)f(b)\} & if\ D < 0 \\ \min\{f(A)f(b), f(a)f(B)\} & if\ D \geq 0 \end{cases}$$

The range of D' is between -1 and 1, with $|D'| = 1$ corresponding to complete LD and $D' = 0$ corresponding to no LD. Another widely used measure of LD is the statistical coefficient of determination, r^2 [5, 33, 51, 64, 67, 82], defined as

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}.$$

Mueller reviewed the different properties and applications of these two measures of LD [64]. The statistical significance test on D is performed by the Pearson's independence test for the 2×2 contingency table generated by the possible combinations of the alleles

of a pair loci, which is also equal to

$$X^2 = \frac{nD^2}{f(A)f(a)f(B)f(b)} = nr^2, \tag{2.1}$$

following a χ^2 distribution with 1 degree of freedom [51, 83, 90]

2.2 Distance correlation sure independence screening

The main framework of the DCRR approach is to first extensively remove the noise via a distance correlation based feature screening approach, and then intensively address the correlation structure using the ridge penalized multiple logistic regression model. Finally the significance test of each individual SNP is performed.

Let $\mathbf{y}$ be the binary phenotype with 1 representing case and 0 representing control. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ be the genotype vector of all SNPs, where p is the number of SNPs. For each biallelic locus, the three possible genotypes can be coded as 0 (for aa), 1 (for Aa), and 2 (for AA).

The dependence strength between two random vectors can be measured by the distance correlation (Dcorr) [76]. Szekely et al. showed that the Dcorr of two random vectors equals zero if and only if these two random vectors are independent. The distance covariance is defined as

$$dcov^2(\mathbf{y}, \mathbf{X}) = \int_{R^{1+p}} ||\phi_{\mathbf{y},\mathbf{X}}(t,s) - \phi_{\mathbf{y}}(t)\phi_{\mathbf{X}}(s)||^2 \, w(t,s)dtds, \tag{2.2}$$

where $\phi_{\mathbf{y}}(t)$ and $\phi_{\mathbf{X}}(s)$ are the respective characteristic functions of $\mathbf{y}$ and $\mathbf{X}$, $\phi_{\mathbf{y},\mathbf{X}}(t,s)$ is the joint characteristic function of $(\mathbf{y}, \mathbf{X})$, and

$$w(t,s) = \{c_1 \; c_p \; ||t||^2 \; ||s||_p^{1+p}\}^{-1},$$

with $c_1 = \pi$, $c_p = \pi^{(1+p)/2}/\Gamma\{(1+p)/2\}$ and $||\cdot||$ stands for the Euclidean norm. Then the Dcorr is defined as

$$dcorr(\mathbf{y}, \mathbf{X}) = \frac{dcov(\mathbf{y}, \mathbf{X})}{\sqrt{dcov(\mathbf{y}, \mathbf{y}) \; dcov(\mathbf{X}, \mathbf{X})}}. \tag{2.3}$$

From Equation (2.2) and (2.3), we notice that the DC approach does not assume any parametric model structure and works well for both linear and nonlinear association. In

addition, it works well for both categorical and continuous data.

Szekely et al. gave a numerically easier estimator of $\hat{dcov}^2(\mathbf{y}, \mathbf{X})$ as

$$\hat{dcov}^2(\mathbf{y}, \mathbf{X}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3. \tag{2.4}$$

Let $\mathbf{y}_i$ and $\mathbf{X}_i$ denote a random sample of the populations $\mathbf{y}$ and $\mathbf{X}$, respectively. Then

$$\begin{aligned}
\hat{S}_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{y}_i - \mathbf{y}_j||\ ||\mathbf{X}_i - \mathbf{X}_j||_p \\
\hat{S}_2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{y}_i - \mathbf{y}_j||\ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||\mathbf{X}_i - \mathbf{X}_j||_p, \\
\hat{S}_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} ||\mathbf{y}_i - \mathbf{y}_k||\ ||\mathbf{X}_j - \mathbf{X}_k||_p
\end{aligned} \tag{2.5}$$

Finally, the point estimator $\hat{dcorr}(y, X)$ can be estimated by Equation (2.3), (2.4), and (2.5).

Let $\mathbf{X}_{\mathcal{C}} = \{X_j | X_j, j = 1, \ldots, d$, is a causative SNP, i.e. truly associated with the phenotype$\}$ and let $\mathbf{X}_{\mathcal{N}} = \{X_k | X_k, k = 1, \ldots, p-d$, is a noise SNP that is not relevant to the phenotype$\}$. The idea of feature screening is to filter $\mathbf{X}_{\mathcal{N}}$ and keep all true causative SNPs in the subset $\mathbf{X}_{\mathcal{C}}$. By decreasing the values of $\hat{dcorr}(y, X_i), i = 1, \ldots, p$, we are able to rank the importance of SNPs from the highest to lowest [56], with $\mathbf{X}_{\mathcal{C}}$ located in front of $\mathbf{X}_{\mathcal{N}}$. Li et al. theoretically proved that the DC feature screening has an additional helpful theoretical sure screening property, all truly important predictors can be selected with probability tending to one as the sample size diverges to ∞, if the tuning parameter d is sufficiently large. The watershed between importance and unimportance, i.e. value of d, like other tuning parameters, is not trivial to determine. Li et al. suggested either set $d = [n/logn]$ ($[\cdot]$ is the integer part) or choose the top d SNPs such that $\hat{dcorr}(y, X_d)$ is greater than a prespecified constant.

Although the DC approach is very powerful at filtering noise and recognizing the truly important SNPs from millions of candidates, it may neglect some important SNPs which are individually uncorrelated yet jointly correlated with the phenotype, or it may highly rank some unimportant SNPs that are spuriously correlated with the phenotype due to their strong LD with other causative SNPs. To overcome these shortcomings, we use iterative distance correlation (IDC) to address possible complex situations that

can exist. The main difference between DC and IDC is that DC finalizes the first d members of $\mathbf{X}_{\mathcal{C}}$ in a single step but IDC builds up $\mathbf{X}_{\mathcal{C}}$ gradually over several steps, i.e. $\mathbf{X}_{\mathcal{C}} = \mathbf{X}_{\mathcal{C}1} \bigcup \mathbf{X}_{\mathcal{C}2} \bigcup \ldots \bigcup \mathbf{X}_{\mathcal{C}k}$, with $d = d_1 + d_2 + \ldots + d_k$, where $\mathbf{X}_{\mathcal{C}i}$ stands for the members selected at i^{th} step and d_i is the size of each set $\mathbf{X}_{\mathcal{C}i}$, for $i = 1, \ldots, k$. The main idea of IDC is to iteratively adjust residuals obtained from regressing all remaining SNPs onto the selected members contained in $\mathbf{X}_{\mathcal{C}}$. Regressing unselected on selected and adjusting residuals effectively breaks down the original complex correlation structure among SNPs. To be more specific, the iterative steps of IDC can be summarized as [93]

Step 1: Choose the first d_1 members of $\mathbf{X}_{\mathcal{C}}$ (i.e. $\mathbf{X}_{\mathcal{C}} = \mathbf{X}_{\mathcal{C}1}$) using DC to rank all candidates of $\mathbf{X}$ for $\mathbf{y}$, where $d_1 < d$.

Step 2: Define $\mathbf{X}_r = \{I_n - \mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T \mathbf{X}_{\mathcal{C}})^{-1} \mathbf{X}_{\mathcal{C}}^T\} \mathbf{X}_{\mathcal{C}}^C$, where $\mathbf{X}_{\mathcal{C}}^C$ is the complement set of $\mathbf{X}_{\mathcal{C}}$. Then choose the second d_2 members into $\mathbf{X}_{\mathcal{C}}$ (i.e. $\mathbf{X}_{\mathcal{C}} = \mathbf{X}_{\mathcal{C}1} \bigcup \mathbf{X}_{\mathcal{C}2}$) using DC to rank all candidates of $\mathbf{X}_r$ for $\mathbf{y}$, where $d_1 + d_2 \leq d$.

Step 3: repeat step 2 until the size of $\mathbf{X}_{\mathcal{C}}$ reaches the pre-specified number d.

Whether or not these d_i at each step are equal ehibits a negligble effect on the results, but their magnitudes will appreciably affect results. Theoretically, smaller d_i will yield better results, but also cause a dramatically lower computational speed. Therefore, we need a combination of DC and IDC to accommodate the computational cost and model performance simultaneously.

2.3 Ridge penalized logistic regression

For LRR, $\mathbf{y}$ is still the binary phenotype and $\mathbf{X}_{\mathcal{C}}$ being the selected important SNPs with moderate dimension ($d = [n]$). For simplicity of notation, we still use $\mathbf{X}$ to denote $\mathbf{X}_{\mathcal{C}}$. To address the correlation among SNPs, stabilize the model estimates, and test for significance of each individual SNP while accommodating the joint effects of others, we impose a ridge penalized logistic multiple regression model [53, 78]. In traditional logistic regression, the probability of case is related to predictors by the inverse logit function

$$p(\mathbf{y}_i = 1|\mathbf{X}) = \frac{e^{\mathbf{X}_i \beta}}{1 + e^{\mathbf{X}_i \beta}}.$$

The parameter vector β^{λ} of ridge logistic regression can be estimated by maximizing the log likelihood subject to a size constraint on L_2 norm of the coefficients via the Newton

- Raphson algorithm

$$l(\mathbf{X}, \beta^{\lambda}) = \sum_{i=1}^{n} \mathbf{y}_i \log[p(\mathbf{y}_i = 1|\mathbf{X})] + \sum_{i=1}^{n} (1 - \mathbf{y}_i) \log[1 - p(\mathbf{y}_i = 1|\mathbf{X})] - \lambda||\beta||^2.$$

The first derivative of the penalized likelihood yields

$$\hat{\beta}^{\lambda} = (\mathbf{X}^T\mathbf{W}\mathbf{X} + 2\lambda I)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Z},$$

where $\mathbf{W} = diag[\hat{p}(\mathbf{y}_i = 1|\mathbf{X})(1 - \hat{p}(\mathbf{y}_i = 1|\mathbf{X}))]$, and $\mathbf{Z}$ is an $n \times 1$ vector with elements

$$z_i = logit[\hat{p}(\mathbf{y}_i = 1|\mathbf{X})] + \frac{\mathbf{y}_i - \hat{p}(\mathbf{y}_i = 1|\mathbf{X})}{\hat{p}(\mathbf{y}_i = 1|\mathbf{X})(1 - \hat{p}(\mathbf{y}_i = 1|\mathbf{X}))}.$$

The tuning parameter λ controls the strength of shrinkage of the norm of β.A few methods have been proposed to choose the tuning parameter λ [32, 45, 52]. One common approach is the ridge trace [46]. The ridge trace is a plot of the parameter estimates over increasing λ values. The ideal λ is where all parameter estimates have stabilized. A suitable choice of $\lambda > 0$ introduces a little bias but decreases the variance and hence minimizes the mean squared error [53, 78]

$$MSE(\hat{\beta}) = Tr[Var(\hat{\beta})] + [bias(\hat{\beta})]^T[bias(\hat{\beta})].$$

The asymptotic variance of $\hat{\beta}^{\lambda}$ can be derived as

$$Var(\hat{\beta^{\lambda}}) = \{\mathbf{X}^T\mathbf{W}\mathbf{X} + 2\lambda\mathbf{I}\}^{-1}\{\mathbf{X}^T\mathbf{W}\mathbf{X}\}\{\mathbf{X}^T\mathbf{W}\mathbf{X} + 2\lambda\mathbf{I}\}^{-1}.$$

2.4 Hypothesis Testing

The significance of each individual SNP, while accounting for the joint and correlated effects of other SNPs, is assessed via the hypothesis test

$$H_{0j} : \beta_j^{\lambda} = 0 \text{ vs } \; H_{1j} : \beta_j^{\lambda} \neq 0, \text{ for } \; j = 1, \ldots, d. \tag{2.6}$$

The corresponding 'non-exact' test statistic is

$$T^{\lambda} = \frac{\hat{\beta}_j^{\lambda}}{se(\hat{\beta}_j^{\lambda})}.$$

Halawa and Bassiouni investigated this 'non-exact' t-type test under two different λs via simulation of 84 different models and concluded that it has considerably larger power in many cases, or slightly less power in a few cases, compared to the test of traditional regression estimates via maximum likelihood [39]. Cule et al. extended this test from a continuous response to the binary response and claimed that the asymptotic standard normal distribution of the test statistic T^{λ} under the null performs as well as that of permutation test [13]. Therefore, we also assume $T^{\lambda} \sim N(0,1)$ under the null and use the standard normal distribution to perform the significance test of each SNP.

Since multiple SNPs are usually tested simultaneously, and the dimension of tests is small or moderate after the feature screening procedure ($d << p$), we use the simplest Bonferroni correction to control the family wise error rate. Whereas the traditional single-locus model use p for multiple correction but we use d instead because the actual number of tests involved is d. We set the SNPs who are filtered out by DC to have p-value of 1 because they are not informative and are not considered for significance testing.

CHAPTER 3

Numerical Simulations

To assess the performance of our approach, we conducted a large number of simulations to obtain the power and type I error rates under varied combinations of the number of SNPs (p), correlation strength (ρ), and minor allele frequency (MAF). We also compare our DCRR approach with the most popular single-locus model, Cochran - Armitage trend test (CA), and the traditional LRR approach.

The correlated haplotype vector was simulated from a multivariate normal distribution with the mean vector randomly generated from Unif$(0, 5)$ and covariance structure designed as $AR(1)$. The variance was fixed to be 1 and the correlation parameter ρ was used to control the strength of LD among SNPs. Next, the individual allele of each haplotype was generated by dichotomizing the continuous haplotype values based on the MAF, and the corresponding percentile obtained from the cumulative density function of the marginal normal distribution of each SNP. For each SNP, we generated two independent haplotypes and the sum of each pair of haplotypes was used to create the genotype, which yields the $n \times p$ dimensional matrix $\mathbf{X}$ [81]. To clearly describe all possible effects and roles of each SNP, we introduce four definitions [62]

- rSNP (risk SNP): a truly causative SNP that functionally affects the phenotype.
- LD.rSNP: a non-causative SNP that has no effect on the phenotype but is in LD with rSNP.
- nSNP: a noise SNP that is neither important for the phenotype nor in LD with any rSNP.
- LD.nSNP: a nSNP that has no effect on the phenotype but is in LD with other nSNPs

From the index set of the SNPs, $S = \{1, \ldots, p\}$, we randomly chose 5 rSNPs. Due to the property of AR(1), the SNPs in the closest neighborhood of these rSNPs were the LD.rSNP with strongest correlations with the rSNPs and hence substantially increased the difficulty in detecting the true rSNPs, which affect both type I error and power. Among the $S \backslash rSNP$ set containing all $p-5$ nSNPs, those far away from these 5 rSNPs had negligible LD with the rSNP and acted as noise. The other nSNPs located in close proximity to each nSNP were LD.nSNP and this correlation among noise had the potential to act as confounders of the rSNPs.

The binary phenotype was generated based on the genotype matrix $\mathbf{X}$ and the effect size. Setting the β values of all 5 rSNP at 1, and all other SNPs as 0, the probability of case was computed as

$$logit[p(\mathbf{y}_i = 1|\mathbf{X})] = \mathbf{X}\beta + \epsilon,$$

where $\epsilon \sim N(0, 1)$.

The four criteria used to evaluate the performance of the models are defined as

- **Strict Power**: the percentage of replications where all 5 rSNPs were simultaneously rejected,
- **Power**: the proportion of rSNPs rejected among all simulation replicates of rSNPs,
- **Type I Error**: the proportion of $p-5$ LD.SNPs, nSNPs, and LD.nSNPs rejected among all simulation replicates of these non-causative SNPs,
- **Time**: total time for all simulation replicates to be completely finished for each different setting and each different approach.

3.1 Simulation design 1

We set $p = 10$ (signal/noise=2), 100 (signal/noise=20), $1,000$ (signal/noise=200), and $10,000$ (signal/noise=2,000) to consider small, medium, high, and ultra high dimension of SNPs. We also controlled the strength of LD from small to large as $\rho = 0.2,\ 0.4,\ 0.6,\ \text{or}, 0.8$. A total of 48 combination of MAF (MAF = 0.1, 0.3, or, 0.5), ρ, and p provided a comprehensive assessment on how our model performed under different conditions. We performed 100 replications for 40 of the simulations, but only 10 replicates for the last 8 settings when $p = 10,000$ and MAF=0.3, or 0.5, due to the

extremely lengthy computational time of LRR. Different λ values were chosen according to different data requirements based on the ridge trace plots. After λs were determined, we used exactly the same λ values to compare both DCRR and LRR for the same data to ensure the comparisons were accurate. During the DC selection procedure, we chose $d = 8$ for $p = 8$, $d = 20$ for $p = 100$, and $d = n/ln(n) \simeq 80$ for $p = 1,000$ and $10,000$. To minimize other possible factors, equal numbers of case and control were generated and sample size n fixed to be 500.

Simulation results of the 48 settings are summarized in Table 3.1 (MAF=0.1), Table 3.2 (MAF=0.3), and Table 3.3 (MAF=0.5). When MAF=0.3 or 0.5, all three approaches achieve satisfactorily high power and strict power for any dimension of SNPs and any LD strength(Figure 3.1). However, the high power of CA came at the cost of extremely inflated type I error, which indicates that the single-SNP model neglected the correlations and joint effects among SNPs. Comparing the three tables simultaneously, we noticed that the type I error of CA kept increasing as ρ increases from 0.2 to 0.8 consistently for any MAF and p. In particular, when $p = 10$ and $\rho = 0.8$, the false discovery rate of CA was as large as 100% for all three different MAF values. Compared to CA, the type I errors of LRR and DCRR did not show an increasing trend as ρ increased, and almost all were below $\alpha = 0.05$.

When MAF=0.1, the possible range of D spanned from 0.01 to 0.81 and hence greatly increased the difficulty level of SNP being detected. As a result, when comparing the power and strict power of MAF=0.1 with the other two MAF values, we noticed that both power and strict power exhibited the smallest value in MAF=0.1 for all three approaches (Figure 3.1). In particular, when the signal/noise ratio or dimension of SNPs increased dramatically, the strict power of MAF=0.1 severely dropped for both CA and LRR for any given ρ (Figure 3.2). Indeed, the strict power of LRR and CA approximated as 40% for $p = 10,000$ and 70% for $p = 1,000$. However, the strict power of DCRR more than doubled compared to that of CA and LRR for any ρ when MAF=0.1 and $p = 10,000$ (Figure 3.1 and Figure 3.2). Figure 3.3 displays the comparisons of strict power (in orange), power (in purple), and type I error (in light blue) simultaneously for all three approaches and four dimensions when $\rho = 0.8$. The strict power and power of CA and LRR decreased dramatically as p increased, but those of DCRR are relatively stable at a value above 90%. Additionally, the type I error of CA was as high as 100% for $p = 10$ while all other approaches had type I error rates less than 5%. The type I

error decreased as p increased for each approach because the ratio of n.SNP to LD.rSNP was increasing.

Of the 48 combinations of varied MAF, LD strength, and dimension, the DCRR method performed consistently and uniformly more powerful than the other approaches, and the superiority of DCRR was striking under harsh conditions such as ultra high dimensions or complex correlations. Among the 48 simulated comparisons there were only two exceptions; when $p = 10$, $\rho = 0.8$, and MAF=0.3 or 0.5, the power and strict power of DCRR was inferior to the other two approaches. This accidental drop was caused by one causative r.SNP that was not successfully selected from the top 8, but rather ranked 9^{th} or 10^{th}. By choosing the tuning parameter d large enough, we were able to avoid this type of error. Since the DC feature screening approach is chiefly designed for ultra high dimensional cases, a dimension as low as 10 did not leave sufficient space for DC to select freely. We believe that the power of DCRR will be manifested for large dimension problems, as occurred in the other 46 simulation comparisons.

We recorded the total computational time of each approach, completing 100 simulation replicates for each fixed simulation setting. From Figure 3.4, we noticed that the computational cost of DCRR dramatically decreased compared to LRR as dimension increased. The computational benefits of DCRR were manifested at $p = 1,000$ and became more remarkable for $p = 10,000$. The computational time of DCRR was similar to that of CA, which indicates that DCRR does not increase the computation cost despite considering multiple joint effects and correlation effects that were neglected by the single-SNP models.

TABLE 3.1: Simulation results for MAF = .1

		p = 10			p = 100		
		CA	LRR	DCRR	CA	LRR	DCRR
$\rho = .2$	Strict Power	1	1	1	0.91	0.91	0.97
	Power	1	1	1	0.982	0.982	0.994
	Type1	0.016	0.014	0.016	0.00032	0.00032	0.0026
	Time	16.34s	11.79s	78.89s	2.4m	.50m	6.52m
$\rho = .4$	Strict Power	1	1	1	0.93	0.93	0.98
	Power	1	1	1	0.984	0.984	0.996
	Type1	0.05	0.036	0.04	0.0022	0.0022	0.0068
	Time	16.82s	24.20s	158.46s	2.44m	.54m	6.54m
$\rho = .6$	Strict Power	1	0.98	0.99	0.94	0.94	0.99
	Power	1	0.996	0.998	0.988	0.988	0.998
	Type1	0.39	0.01	0.02	0.0088	0.0085	0.0195
	Time	15.96s	13.48s	80.45s	2.59m	.50m	7.81m
$\rho = .8$	Strict Power	1	0.94	0.98	0.94	0.96	0.99
	Power	1	0.988	0.996	0.988	0.992	0.998
	Type1	0.99	0.018	0.044	0.0546	0.0287	0.0522
	Time	16.17s	14.58s	79.49s	2.6m	.59m	7.12m
		p = 1000			p = 10,000		
		CA	LRR	DCRR	CA	LRR	DCRR
$\rho = .2$	Strict Power	0.74	0.72	0.92	0.37	0.57	0.99
	Power	0.944	0.94	0.984	0.832	.896	0.998
	Type1	0.00004	0.00005	0.0005	0.000007	0.000004	0.00049
	Time	48.48m	35.96m	73.91m	95.71h	422.41h	107.08h
$\rho = .4$	Strict Power	0.68	0.67	0.91	0.40	0.48	0.91
	Power	0.93	0.93	0.982	0.836	0.846	0.982
	Type1	0.00003	0.0003	0.0005	0.000004	0.000006	0.0005
	Time	47.34m	33.68m	69.86m	97.87h	443.53h	111.42h
$\rho = .6$	Strict Power	0.77	0.78	0.96	0.39	0.42	0.93
	Power	0.95	0.952	0.992	0.834	0.874	0.986
	Type1	0.00016	0.0002	0.001	0.000009	0.00001	0.00051
	Time	48.71m	32.50m	72.18m	97.57h	420h	105h
$\rho = .8$	Strict Power	0.68	0.69	0.89	0.40	0.43	0.93
	Power	0.932	0.942	0.978	0.856	0.854	0.986
	Type1	0.0012	0.0011	0.0037	0.00003	0.000036	0.00073
	Time	53.02m	33.55m	69.52m	94.93h	379.62h	64.88h

TABLE 3.2: Simulation results for MAF = .3

		p = 10			p = 100		
		CA	LRR	DCRR	CA	LRR	DCRR
$\rho = .2$	Strict Power	1	1	1	1	1	1
	Power	1	1	1	1	1	1
	Type1	0.046	0.028	0.034	0.00052	0.0053	0.0034
	Time	18.04s	12.41s	78.30s	2.43m	.58m	7.56m
$\rho = .4$	Strict Power	1	1	1	0.99	0.99	0.99
	Power	1	1	1	0.998	0.998	0.998
	Type1	0.228	0	0.014	0.0086	0.0083	0.018
	Time	17.93s	13.14s	80.23s	2.40m	.59m	7.55m
$\rho = .6$	Strict Power	1	1	1	1	1	1
	Power	1	1	1	1	1	1
	Type1	0.856	0.004	0.012	0.0354	0.0341	0.0508
	Time	18.43s	12.81s	77.97s	2.41m	.58m	8.13m
$\rho = .8$	Strict Power	1	1	0.87	1	1	1
	Power	1	1	0.974	1	1	1
	Type1	1	0.006	0.028	0.1358	0.0107	0.0188
	Time	17.73s	13.23s	78.09s	2.44m	.657m	7.16m
		p = 1000			p = 10,000		
		CA	LRR	DCRR	CA	LRR	DCRR
$\rho = .2$	Strict Power	0.96	0.96	0.97	0.9	0.9	1
	Power	0.992	0.992	0.994	0.98	0.98	1
	Type1	0.00008	0.00008	0.0006	0	0	0.0005
	Time	57.32m	36.59m	49.36m	9.33h	42.36h	11.21h
$\rho = .4$	Strict Power	0.98	0.98	0.99	1	1	1
	Power	0.996	0.996	0.998	1	1	1
	Type1	0.00014	0.0001	0.0009	0.00001	0.00001	0.0005
	Time	50.78m	34.13m	73.3m	10.35h	46.21h	10.22h
$\rho = .6$	Strict Power	0.98	0.98	1	1	1	1
	Power	0.996	0.998	1	1	1	1
	Type1	0.00086	0.0008	0.0027	0.00005	0.00006	0.0006
	Time	49.02m	35.33m	71.10m	10.94h	41.42h	10.99h
$\rho = .8$	Strict Power	0.97	0.97	1	1	1	1
	Power	0.994	0.994	1	1	1	1
	Type1	0.0055	0.0051	0.0104	0.0004	0.0004	0.0016
	Time	50.55m	32.55m	69.95m	10.65h	38.35h	10.20h

TABLE 3.3: Simulation results for MAF = .5

		p = 10			p = 100		
		CA	LRR	DCRR	CA	LRR	DCRR
$\rho = .2$	Strict Power	1	1	1	1	1	1
	Power	1	1	1	1	1	1
	Type1	0.036	0.018	0.024	0.0015	0.0014	0.0043
	Time	18.82s	11.95s	78.62s	2.42m	.57m	7.72m
$\rho = .4$	Strict Power	1	1	1	1	1	1
	Power	1	1	1	1	1	1
	Type1	0.296	0.0006	0.048	0.0105	0.0102	0.0189
	Time	17.55s	12.47s	79.92s	2.49m	.57m	7.69m
$\rho = .6$	Strict Power	1	1	1	1	1	1
	Power	1	1	1	1	1	1
	Type1	0.908	0.008	0.036	0.0379	0.0259	0.0391
	Time	18.36s	13.64s	78.46s	2.42m	.60m	7.51m
$\rho = .8$	Strict Power	1	1	0.81	1	1	1
	Power	1	1	0.962	1	1	1
	Type1	1	0.012	0.054	0.1581	0.0124	0.0215
	Time	17.91s	13.85s	78.31s	2.44m	.67m	10.89m
		p = 1000			p = 10,000		
		CA	LRR	DCRR	CA	LRR	DCRR
$\rho = .2$	Strict Power	1	1	1	0.9	0.9	1
	Power	1	1	1	.98	.98	1
	Type1	0.00005	0.00005	0.0006	0.00001	0.00001	0.0004
	Time	54.31m	35.62m	73.38m	10.65h	43.16h	10.68h
$\rho = .4$	Strict Power	1	1	1	0.9	0.9	1
	Power	1	1	1	0.98	0.98	1
	Type1	0.00017	0.0002	0.0009	0.00001	0.00001	0.0006
	Time	48.07m	33.62m	71.57m	11.12h	43.24h	11.47h
$\rho = .6$	Strict Power	0.99	1	1	1	1	1
	Power	0.998	1	1	1	1	1
	Type1	0.0011	0.001	0.0036	0.00006	0.00007	0.00077
	Time	46.66m	32.48m	71.13m	11.09h	39.40h	11.47h
$\rho = .8$	Strict Power	1	1	1	1	1	1
	Power	1	1	1	1	1	1
	Type1	0.0011	0.001	0.0036	0.00047	0.00046	0.0020
	Time	47.85m	34.67m	72.65m	10.87h	38.91h	10.48h

3.2 Simulation Design 2

Although our above simulation design achieved agreeable results, we further tested the power of our DCRR approach by increasing the difficulty of the simulations. The previous simulation focused on an equal number of cases and controls, identical MAF values, identical LD structures for all SNPs, and limited noise to signal ratios under each fixed simulation design. In this simulation, we approximated a real life scenario with several complications that simultatiously occured in one data set.

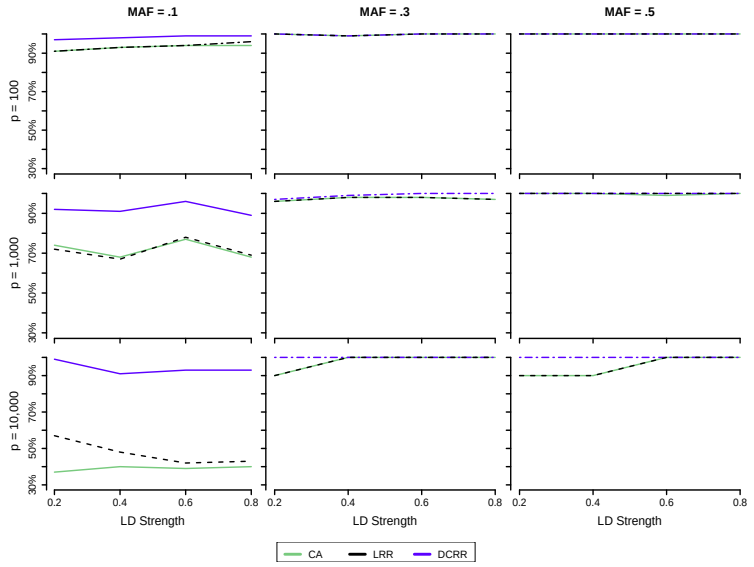

FIGURE 3.1: The changing pattern of strict power of three approaches as increasing ρ under combinations of varied MAF and dimension

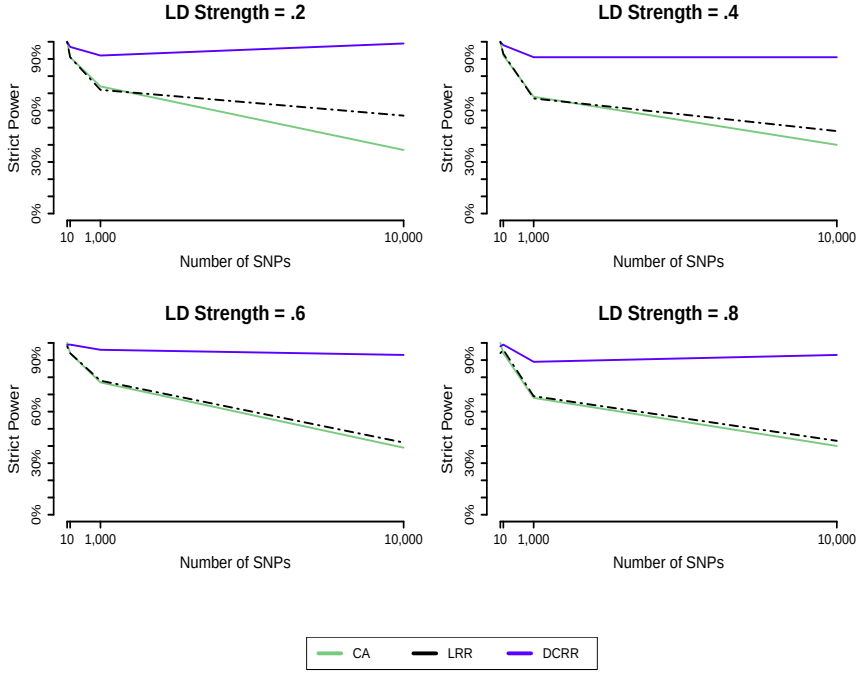

FIGURE 3.2: The changing pattern of strict power of three approaches as increasing p when MAF $= 0.1$ for each LD strength.

We generated 216,100 SNPs on 84 individuals (56 cases and 28 controls). We made three causative r.SNPs in Chromosome (Chr) 1, 3, and 5. The mean of r1.SNP is assigned to be 5 (Chr 3), r2.SNP be 3 (Chr 1), and r3.SNP be 2 (Chr 5). For each r.SNP, we simulated a block with 20 LD.rSNPs with 10 flanking each side. We continued to use the AR(1) covariance structure. The mean of these 60 LD.rSNPs are randomly selected

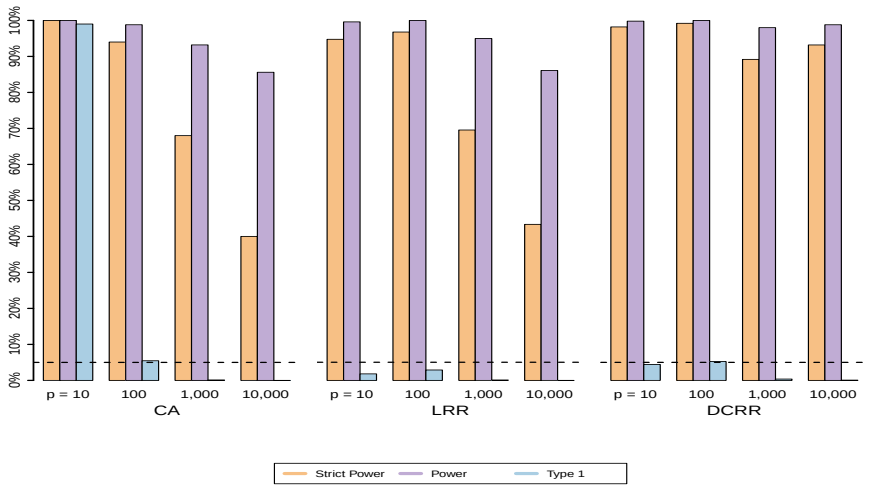

FIGURE 3.3: The simultaneous changing pattern of strict power, power, and type I of three approaches as increasing p when MAF=0.1 and $\rho = .8$.

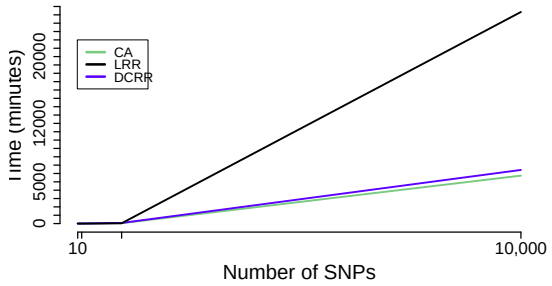

FIGURE 3.4: The changing pattern of computational time (in minutes) of three approaches as increasing p.

from $Unif(0, 1.5)$. The ρ of r1.SNP and r3.SNP blocks were randomly selected from $Unif(0.6, 0.9)$, and the ρ of r2.SNP block was fixed to be 0.8. By setting $\beta = 3$ for the three r.SNPs and 0 for all other 60 LD.rSNPs, we connected the phenotype with these three causative SNPs using the same approach as that of the Simulation 1 design. Then, we randomly generated all remaining 216,037 n.SNPs through a binomial distribution, with a randomly chosen allele frequency q from $Unif(.05, .095)$ and three genotypes from corresponding probabilities $(1-q)^2, 2q(1-q)$, and q^2. Finally, we randomly selected the position number of the three causative SNPs and arranged the r.SNPs, LD.rSNPs, and n.SNPs into the whole genome.

After applying the DCRR approach with $d = 2n$ and $\lambda = 2$ (see Figure 3.5), we successfully detected the three polygenically causative SNPs without being confounded by the other 60 purposely designed strong LD.rSNPs and a large number of n.SNPs

Figure (3.6). The r3.SNP in Chr 5 is only slightly above the threshold because we designed it to have a weak effect.

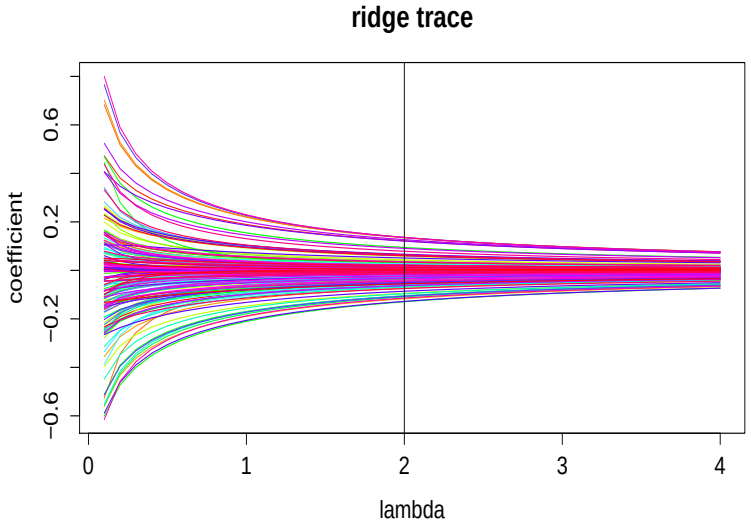

FIGURE 3.5: Ridge trace plot of the 168 important SNPs using LRR for the simulated big data.

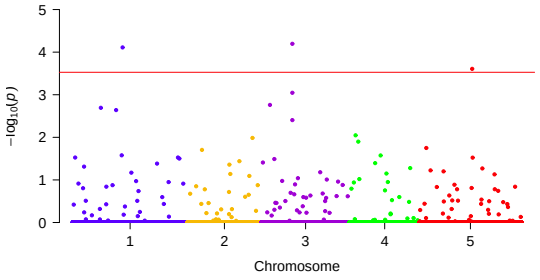

FIGURE 3.6: The Manhattan plot of the simulated big data along the whole genome, based on $-\log_{10}$ of genome-wide simultaneous P values of 216,100 SNPs against its physical chromosomal position. Chromosomes are shown in alternate colors. Three causative SNPs located in Chr 1 (moderate effect), 2 (strong effect), and 5 (slight effect), affecting the phenotype jointly with complex LD structure.

CHAPTER 4

Real Data Analysis

Our DCRR approach was applied to search for significant causative SNPs for a binary trait of the *Arabidopsis thaliana* hypersensitive response to the bacterial elicitor *AvrRpm1*, with 84 inbred lines (28 susceptibilities and 56 resistances) and 216,130 SNPs. This data is publicly available from the link (http://arabidopsis.usc.edu). *A. thaliana* has a genome of approximately 120 megabases and SNP density of one SNP per 500 base pairs [3]. Five statistical models have been tested on this same data, and reported that this *AvrRpm1* trait was monogenically regulated by the gene *RPM1*. i.e. the bacterial avirulence gene *AvrRpm1* directly identified the corresponding resistance gene *RISISTANCE TO P.SYRINGAW PV MACULICOLA 1 (PRM1)* [35] . Atwell et al. compared two single-SNP approaches: Fisher's exact test without correcting for background confounding SNPs and a mixed model implemented in EMMA to correct for confounding SNPs (Supplementary Figure 36 in page 52 of [3]). Shen et al. proposed a Heteroscedastic effects model (HEM), determined 5% genome-wide significance thresholds via permuation test, and claimed that the HEM approach successfully eliminated many spurious associations and improved the traditional ridge regression (SNP-BLUP) approach (Figure w of [70]). Our DCRR model effectively identified the *RPM1* gene in exactly the same position (Chr 3, 2227823 bp), with a significance level 10^{-12} on the highest peak. Figure 4.1 demonstrates the manhattan plot of the AVRRpm1 trait along the whole genome, based on $-log_{10}$ of genome-wide simultaneous p-values of 216,130 SNPs against its physical chromosomal position. The blue horizontal line correspond to a 5% genome-wide simultaneous significance threshold with Bonferroni correction for 250,000 tests. The red horizontal line represents the proposed multiple correction threshold for 5% genome-wide simultaneous threshold with Bonferroni correction for only $d = 189$ tests.

TABLE 4.1: Significant SNPs detected by DCRR

Rank	Chromosome	Base Pair Position (bp)	Gene	Dcorr	P-value
1	3	2227823	RPM1	0.5846	1.01×10^{-11}
2	3	2225899	...	0.5075	2.75×10^{-9}
3	3	2225040	...	0.5075	7.94×10^{-9}
22	3	2231452	...	0.3450	5.35×10^{-8}

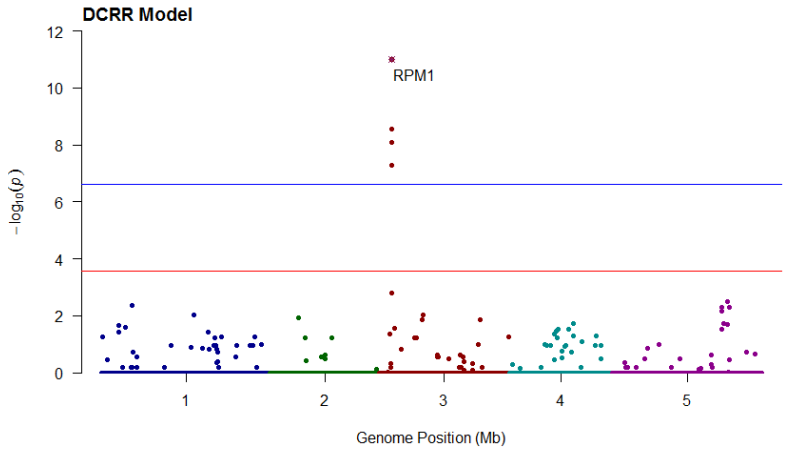

FIGURE 4.1: The Manhattan plot of the *AvrRpm1* along the whole genome, based on $-\log_{10}$ of genome-wide simultaneous P values of 216,130 SNPs against its physical chromosomal position. Chromosomes are shown in alternate colors.

The four significant causative polymorphism that passed the DCRR threshold (in red) also passed the thresholds of other approaches (in blue), are summarized in Table 4.1. Using the Arabidopsis Genome Initiative (AGI) genetic map and the Arabidopsis information resource (TAIR.org, verified on 5/7/2015) GBrowse database, we matched our significant findings with three genes. The rank 1 SNP lied within the single large exon of RPM1 (2229024-2225952). The rank 2 SNP lied approximately 50bp past the 3' end of the RPM1 region. The rank 3 SNP lied within an intron in the neighboring alba DNA/RNA binding protein (2225254-2223001), and the rank 22 SNP lied within exon4 of the neighboring NSN1 gene (nucleostemin-like 1, 2232361-2229590). Additionally, the DCRR eliminated many nominally significant associations. Indeed, the shrinkage effect of DCRR was much stronger than that of any of the other four approaches. We noticed a reduction in the number of moderate associations in the whole genome, and those with significance levels from 10^{-3} to 10^{-6} in EMMA and Fisher disappear from DCRR. Additionally, one slightly significant SNP in Chr 5 in EMMA and some highly significant SNPs closely neighboring *RPM1* in EMMA and Fisher were all eliminated in DCRR.

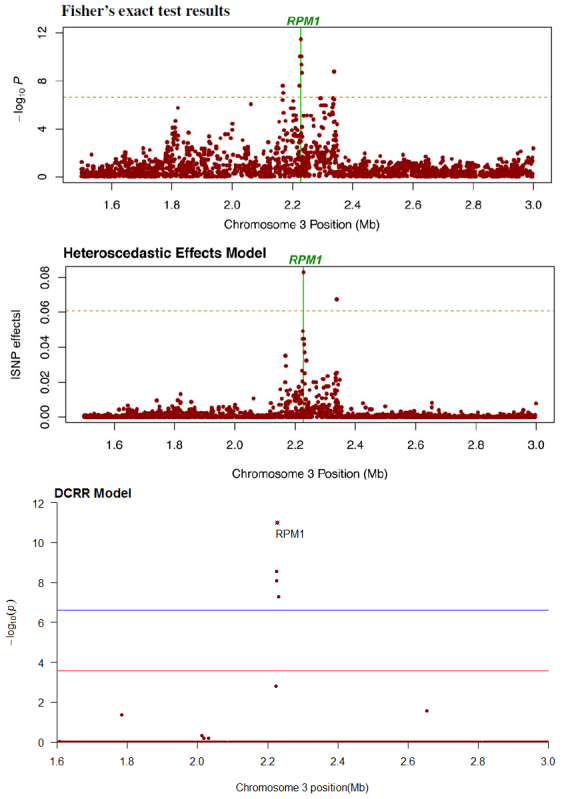

FIGURE 4.2: Magnification of the genome region surrounding *RPM1*. The current findings for the same region using three different approaches are compared.

We noticed a second peak (0.1 Mb away from *RPM1*) that was detected as highly significant by both Fisher and HEM models judging from Figure 4.2 [3, 70]. However, DCRR results indicated it is a spurious signal confounded by strong background LD. If the process was limited to ranking by DC, that SNP indeed ranked high with a similar pattern as Fisher and HEM. However, the iterative DC that adjusts residuals to break down the original correlation structures reduced that SNP to an extremely low rank, 156997^{th} among all candidates with a dCorr value of just 0.0444. Therefore, it was highly unlikely that this SNP (Chr 3, 2337844 bp) was associated with the phenotype. To further verify this conclusion we examined the LD of this SNP with several surrounding SNPs. After a χ^2 test using Equation (2.1), we found this SNP was in strong LD with over 50 other polymorphisms (Table 4.2). As observed from Table 4.2, it was highly correlated with all four significant SNPS (denoted with an asterisk) reported in Table 4.1, especially having P value of 10^{-11} with *RPM1*. It was also highly correlated with many other non-causative SNPs, for example it showed a P value 10^{-16} with position 2334985 and P value 10^{-15} with position 2335305.

We further visually examined the genetic patterns for the region surrounding gene

TABLE 4.2: The pairwise LD strength of the point located in Chr 3 with position number 2337844bp with several surrounding SNPs. The Pvalue is obtained from χ^2 test with 1 degree of freedom

Chromosome	Base Pair Position (bp)	χ^2	P-value
3	2227823*	41.9792	9.22×10^{-11}
3	2225899*	29.9614	4.41×10^{-8}
3	2231452*	24.9712	5.81×10^{-7}
3	2225040*	18.9063	1.37×10^{-5}
3	2334985	64.3782	9.99×10^{-16}
3	2335305	60.2751	8.21×10^{-15}
3	2332822	46.5432	8.96×10^{-12}
3	2333137	49.6274	1.85×10^{-12}
3	2332597	49.6274	1.85×10^{-12}
3	2334723	38.4016	5.75×10^{-10}
3	2336637	28.7376	8.28×10^{-08}
3	2336926	31.2202	2.30×10^{-08}
3	2336966	28.7376	8.28×10^{-08}
3	2334909	31.7913	1.71×10^{-08}
3	2291826	28.7225	8.35×10^{-08}
3	2295084	28.7225	8.35×10^{-08}
3	2320691	28.7225	8.35×10^{-08}
3	2294447	26.2953	2.92×10^{-07}
3	2331847	27.2956	1.74×10^{-07}
3	2336077	27.2956	1.74×10^{-07}
3	2302458	26.2953	2.92×10^{-07}
3	2302750	26.2953	2.92×10^{-07}
3	2304433	23.9354	9.96×10^{-07}
3	2304563	26.2953	2.92×10^{-07}
3	2305255	26.2953	2.92×10^{-07}
3	2306492	26.2953	2.92×10^{-07}
3	2308001	26.2953	2.92×10^{-07}
3	2310061	26.2953	2.92×10^{-07}
3	2325609	21.7285	3.14×10^{-06}
3	2261331	20.7359	5.27×10^{-06}
3	2318129	18.5587	1.64×10^{-05}
3	2326014	17.2805	3.22×10^{-05}
3	2327593	18.6292	1.58×10^{-05}

RPM1 using a haploview heatmap, with short-range of 7.3 kb and medium-range of 28.1 kb (see Figure 4.3). All pariwise r^2 among SNPs in the region were computed, with nine color schemes representing the varied level of LD strengths (red denotes strong LD, yellow for medium LD, and white for negligible LD). The LD patterns among the closest SNPs to the right side of the causative SNP were very strong (> 0.9), while the majority of SNPs were in medium LD (r^2 from 0.4 to 0.7). A close inspection of the 20 closest surrounding SNPs highlighted that the LD pattern in the neighborhood of *RPM1* varied

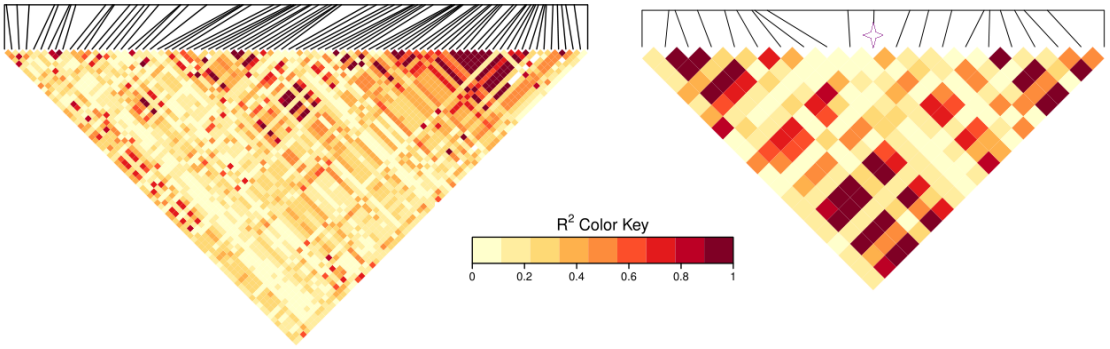

FIGURE 4.3: Haploview heatmap plot of the surrounding SNPs in the *RPM1* gene region. Left panel: medium range of 28.1 kb involving 100 neighbored SNPs; Right panel: short range of 7.3 kb involving 20 neighbored SNPs.

substantially, with 8 SNPs showing strong, and 6 SNPs unlinked (i.e. 70% closest SNPs having medium to strong LD with it).

The total computation time for this data comprised 6 hours on a windows operating system with a 2.10 Ghz Intel Xeon processor and 32GB of RAM. The top $d = 189$ important SNPs were selected by the iterative DC procedure, after which all noise SNPs whose Dcorr values fell below 0.25 were filtered (Figure 4.4). We choose $\lambda = 2$ for our analysis (Figure 4.5).The results were relatively stable, and negligible differences were observed when we changed λ to any value from 1 to 3.

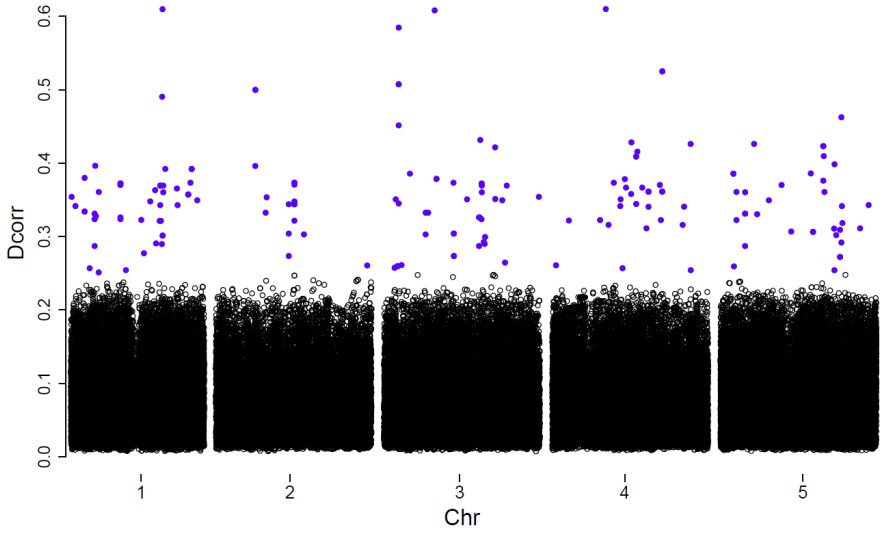

FIGURE 4.4: Dcorr value and location of the top $d = 189$ important SNPs selected by the iterative DC procedure *AvrRpm1*.

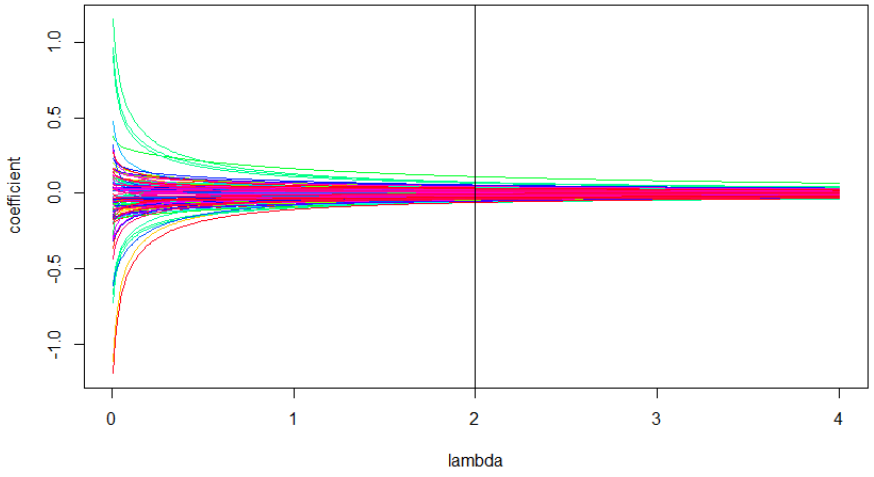

FIGURE 4.5: Ridge trace plot of the 189 important SNPs using LRR for the *AvrRpm1* data.

CHAPTER 5

Discussion

High-throughput genotyping techniques and large data repositories of case-control sample consortium provide opportunites for GWAS to unravel the genetic etiology of complex traits. With the number of SNPs per DNA array growing from 10,000 to 1 million [1], the ultra-high dimension of datasets is one of the grand challenges in GWAS.

We proposed a novel DCRR approach to address the complex LD, multiple joint genetic effects, and ultra high dimension problems in the whole genome data. We considered an *A. thaliana* whole genome data set that Atwell et al.reported as carrying several challenges: False positive rates or spurious significant association were present due to confounding effects of high population structure. The true positive signal was difficult to identify because the a priori candidates were over-represented by surrounding SNPs in the vicinity through complex diffuse 'mountain range' like peaks covering a broad and complex region without a clear center. Sometimes the true causal polymorphism did not have stronger signal than the spurious ones, which could have occurred when r.SNPs were positively correlated with other r.SNPs or with genomic background SNPs. The sample size was relatively small ($n = 84$), which may have limited the power of statistical significance. The natural selection on each locus may have been strong, such that the allele frequency distributions of the causative loci were very different from those of the background noise loci. Those distributions may have further disabled many statistical approaches that address genome-wide associations. Finally, a single-SNP model may have caused model misspecification. As was stated by Atwell et al., "At least for complex traits, the problem is better thought of as model misspecificaiton: when we carry out GWA analysis using a single SNP at a time (as was done here and in most other previous GWA studies), we are in effect modeling a multifactorial trait as if it were due to a single locus. The polygenic background of the trait is ignored, as are other unobserved variables."

Our approach solved the challenges mentioned by Atwell et al. By breaking down the complex LDs among causative and non-causal SNPs, the causative effects were reinforced while the nominally spurious signal shrunk towards zero. The shrinkage effect of the DCRR approach presented herein was much more robust and accurate than previous approaches (Figures 4.1 and 4.2), and the false positive rates were decreased dramatically while the true positive rates (power) increased. After filtering out the majority of noise and reducing the SNPs from millions to hundreds, the problems caused by ultra high dimension were removed. After generating the MAF of all loci randomly from a $Unif(0.05, 0.95)$ distribution, which imitated strong natural selection effects and also considered the effects of rare alleles, The DCRR approach still successfully detected the causative SNPs. By considering multiple joint effects with complex correlation structures that were neglected by the single-SNP model, the DCRR approach is superior to the CA approach in both power and type I error control.

Malo et al. applied ridge regression to handle the LD among genetic associations. Their work mainly focused on continuous phenotypes and a moderate dimension ($p > n$ but not $p >> n$) [58]. Cule et al. proposed the asymptotic significance test approaches in ridge regression for both binary and continuous phenotype, but their approach is also mainly focused on moderate dimension [13]. The advantages of DCRR were assessed extensively in previous Section 3.1, and the DCRR approach can be easily extended to continuous phenotypes. Since a binary response tends to have fewer statistical properties, i.e. the prediction errors tend to be much higher for binary than continuous outcomes, we expect that the performance of DCRR for continuous traits will only improve.

Methods to increase the signal to noise ratio are critical for successful GWAS and the challenges of GWAS are not specific to the dataset from Atwell et al. The monogenetic control with one causative locus in the *AvrRpm1* dataset may not fully highlight the power of the DCRR approach. As future work, we will apply the DCRR approach to polygenic traits such as human diseases or traits in organisms with agricultural importance. For organisms under artificial selection for trait improvement, such as agricultural crops, spurious or extraneous SNPs in a marker-assisted selection scheme could add cost and time in genotyping as well as possibly misdirect selection priorities. Therefore, DCRR approach has the potential to provide improved efficiency and accuracy to researchers to design their experiments with applied outcomes wisely.

BIBLIOGRAPHY

[1] David Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *science*, 322(5903):881–888, 2008.

[2] Peter Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[3] Susanna Atwell, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, Tina T Hu, et al. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, 2010.

[4] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[5] AHD Brown. Sample sizes required to detect linkage disequilibrium between two or three loci. *Theoretical population biology*, 8(2):184–201, 1975.

[6] Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[7] Lon R Cardon and John I Bell. Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91–99, 2001.

[8] Rui Chen, George I Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo YK Lam, Rong Chen, Elana Miriami, Konrad J Karczewski, Manoj Hariharan, Frederick E Dewey, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307, 2012.

[9] Jonathan C Cohen, Robert S Kiss, Alexander Pertsemlidis, Yves L Marcel, Ruth McPherson, and Helen H Hobbs. Multiple rare alleles contribute to low plasma levels of hdl cholesterol. *Science*, 305(5685):869–872, 2004.

[10] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[11] Dana C Crawford, Christopher S Carlson, Mark J Rieder, Dana P Carrington, Qian Yi, Joshua D Smith, Michael A Eberle, Leonid Kruglyak, and Deborah A Nickerson. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *The American Journal of Human Genetics*, 74(4):610–622, 2004.

[12] Erika Cule. *ridge: Ridge Regression with automatic selection of the penalty parameter*, 2014. R package version 2.1-3.

[13] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC bioinformatics*, 12(1):372, 2011.

[14] Mark J Daly, John D Rioux, Stephen F Schaffner, Thomas J Hudson, and Eric S Lander. High-resolution haplotype structure in the human genome. *Nature genetics*, 29(2):229–232, 2001.

[15] Elisabeth Dawson, Gonçalo R Abecasis, Suzannah Bumpstead, Yuan Chen, Sarah Hunt, David M Beare, Jagjit Pabial, Thomas Dibling, Emma Tinsley, Susan Kirby, et al. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418(6897):544–548, 2002.

[16] B Devlin and Neil Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322, 1995.

[17] Linda M Dong, John D Potter, Emily White, Cornelia M Ulrich, Lon R Cardon, and Ulrike Peters. Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. *Jama*, 299(20):2423–2436, 2008.

[18] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.

[19] Norman Richard Draper, Harry Smith, and Elizabeth Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1966.

[20] Douglas F Easton, Karen A Pooley, Alison M Dunning, Paul DP Pharoah, Deborah Thompson, Dennis G Ballinger, Jeffery P Struewing, Jonathan Morrison, Helen Field, Robert Luben, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093, 2007.

[21] Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

[22] Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494), 2011.

[23] Jianqing Fan, Shaojun Guo, and Ning Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

[24] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[25] Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*, 2006.

[26] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[27] Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038, 2009.

[28] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[29] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[30] Stacey B Gabriel, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002.

[31] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[32] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[33] Anna González-Neira, Francesc Calafell, Arcadi Navarro, Oscar Lao, Howard Cann, David Comas, and Jaume Bertranpetit. Geographic stratification of linkage disequilibrium: a worldwide population study in a region of chromosome 22. *Hum Genomics*, 1(6):399–409, 2004.

[34] Benjamin J Grady, Eric Torstenson, and Marylyn D Ritchie. The effects of linkage disequilibrium in large scale snp datasets for mdr. *BioData mining*, 4(1), 2011.

[35] Murray R Grant, Laurence Godiard, Esther Straube, Tom Ashfield, Jürgen Lewald, Airlie Sattler, Roger W Innes, and Jeffrey L Dangl. Structure of the arabidopsis rpm1 gene enabling dual specificity disease resistance. *Science*, 269(5225):843–846, 1995.

[36] Marvin Gruber. *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*, volume 156. CRC Press, 1998.

[37] Julius Gudmundsson, Patrick Sulem, Andrei Manolescu, Laufey T Amundadottir, Daniel Gudbjartsson, Agnar Helgason, Thorunn Rafnar, Jon T Bergthorsson, Bjarni A Agnarsson, Adam Baker, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature genetics*, 39(5):631–637, 2007.

[38] Christopher A Haiman, Nick Patterson, Matthew L Freedman, Simon R Myers, Malcolm C Pike, Alicja Waliszewska, Julie Neubauer, Arti Tandon, Christine Schirmer, Gavin J McDonald, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature genetics*, 39(5):638–644, 2007.

[39] AM Halawa and MY El Bassiouni. Tests of regression coefficients under ridge regression models. *Journal of Statistical Computation and Simulation*, 65(1-4):341–356, 2000.

[40] Peter Hall and Hugh Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3), 2009.

[41] Peter Hall, Yvonne Pittelkow, and Malay Ghosh. Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):159–173, 2008.

[42] Trevor Hastie and Robert Tibshirani. Efficient quadratic regularization for expression arrays. *Biostatistics*, 5(3):329–340, 2004.

[43] Douglas M Hawkins and Xiangrong Yin. A faster algorithm for ridge regression of reduced rank data. *Computational statistics & data analysis*, 40(2):253–262, 2002.

[44] Qianchuan He and Dan-Yu Lin. A variable selection method for genome-wide association studies. *Bioinformatics*, 27(1):1–8, 2011.

[45] Arthur E Hoerl, Robert W Kannard, and Kent F Baldwin. Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2):105–123, 1975.

[46] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[47] Sarah M Hook, Amanda J Phipps-Green, Fathimath Faiz, Les McNoe, Cushla McKinney, Jade E Hollis-Moffatt, and Tony R Merriman. Smad2: A candidate gene for the murine autoimmune diabetes locus idd21. 1. *The Journal of Clinical Endocrinology & Metabolism*, 96(12):E2072–E2077, 2011.

[48] Richard S Houlston and Julian Peto. The search for low-penetrance cancer susceptibility alleles. *Oncogene*, 23(38):6471–6476, 2004.

[49] Uk H Jo, Sle GL Han, Jae H Seo, Kyong H Park, Jae W Lee, Hyo J Lee, Jeong S Ryu, and Yeul H Kim. The genetic polymorphisms of her-2 and the risk of lung cancer in a korean population. *BMC cancer*, 8(1):359, 2008.

[50] LB Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome research*, 10(10):1435–1444, 2000.

[51] E Kulinskaya and A Lewin. Testing for linkage and hardy-weinberg disequilibrium. *Annals of human genetics*, 73(2):253–262, 2009.

[52] Jerald F Lawless and P Wang. A simulation study of ridge and other regression estimators. *Communications in Statistics-Theory and Methods*, 5(4), 1976.

[53] Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992.

[54] RC Lewontin. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49, 1964.

[55] Gaorong Li, Heng Peng, Jun Zhang, Lixing Zhu, et al. Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877, 2012.

[56] Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

[57] Yi Li, Wing-Kin Sung, and Jian Jun Liu. Association mapping via regularized regression analysis of single-nucleotide–polymorphism haplotypes in variable-sized sliding windows. *The American Journal of Human Genetics*, 80(4):705–715, 2007.

[58] Nathalie Malo, Ondrej Libiger, and Nicholas J Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, 82(2):375–385, 2008.

[59] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[60] Jonathan Marchini, Peter Donnelly, and Lon R Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature genetics*, 37(4):413–417, 2005.

[61] Gilean AT McVean, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584, 2004.

[62] Yan A Meng, Yi Yu, L Adrienne Cupples, Lindsay A Farrer, and Kathryn L Lunetta. Performance of random forest when snps are in linkage disequilibrium. *BMC bioinformatics*, 10(1):78, 2009.

[63] Annette M Molinaro, Nicholas Carriero, Robert Bjornson, Patricia Hartge, Nathaniel Rothman, and Nilanjan Chatterjee. Power of data mining methods to detect genetic associations and interactions. *Human heredity*, 72(2):85–97, 2011.

[64] Jakob C Mueller. Linkage disequilibrium for different scales and applications. *Briefings in bioinformatics*, 5(4):355–364, 2004.

[65] Benjamin H Mullin, Cyril Mamotte, Richard L Prince, Tim D Spector, Frank Dudbridge, and Scott G Wilson. Conditional testing of multiple variants associated with bone mineral density in the flnb gene region suggests that they represent a single association signal. *BMC genetics*, 14(1):107, 2013.

[66] Nila Patil, Anthony J Berno, David A Hinds, Wade A Barrett, Jigna M Doshi, Coleen R Hacker, Curtis R Kautzer, Danny H Lee, Claire Marjoribanks, David P McDonough, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, 2001.

[67] Jonathan K Pritchard and Molly Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

[68] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[69] David E Reich, Michele Cargill, Stacey Bolk, James Ireland, Pardis C Sabeti, Daniel J Richter, Thomas Lavery, Rose Kouyoumjian, Shelli F Farhadian, Ryk Ward, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001.

[70] Xia Shen, Moudud Alam, Freddy Fikse, and Lars Rönnegård. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, 2013.

[71] J.-H. Shin, S. Blay, B. McNeney, and J. Graham. Ldheatmap: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Soft*, 16:Code Snippet 3, 2006.

[72] Montgomery Slatkin. Linkage disequilibriumunderstanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[73] Lucia Sobrin, Todd Green, Xueling Sim, Richard A Jensen, E Shyong Tai, Wan Ting Tay, Jie Jin Wang, Paul Mitchell, Niina Sandholm, Yiyuan Liu, et al. Candidate gene association study for diabetic retinopathy in persons with type 2 diabetes: the candidate gene association resource (care). *Investigative ophthalmology & visual science*, 52(10):7593–7602, 2011.

[74] Lincoln D Stein et al. The case for cloud computing in genome informatics. *Genome Biol*, 11(5):207, 2010.

[75] Yan V Sun, Kerby A Shedden, Ji Zhu, Nam-Hee Choi, and Sharon LR Kardia. Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression. In *BMC proceedings*, volume 3, page S67. BioMed Central Ltd, 2009.

[76] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[77] Stephen Turner. *qqman: Q-Q and manhattan plots for GWAS data*, 2014. R package version 0.1.2.

[78] E Vago and S Kemeny. Logistic ridge regression for clinical data analysis (a case study). *Appl Ecol Env Res*, 4(2):171–179, 2006.

[79] Kristina M Visscher and Daniel H Weissman. Would the field of cognitive neuroscience be advanced by sharing functional mri data? *BMC medicine*, 9(1):34, 2011.

[80] Jeffrey D Wall and Jonathan K Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587–597, 2003.

[81] Tao Wang, Xiaofeng Zhu, and Robert C Elston. Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *The American Journal of Human Genetics*, 80(5):911–920, 2007.

[82] William YS Wang, Bryan J Barratt, David G Clayton, and John A Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.

[83] Bruce S Weir et al. *Genetic data analysis. Methods for discrete population genetic data.* Sinauer Associates, Inc. Publishers, 1990.

[84] JOHN C WHITTAKER, ROBIN THOMPSON, and MIKE C DENHAM. Marker-assisted selection using ridge regression. *Genetical research*, 75(02):249–252, 2000.

[85] Elizabeth A Worthey, Alan N Mayer, Grant D Syverson, Daniel Helbling, Benedetta B Bonacci, Brennan Decker, Jaime M Serpe, Trivikram Dasu, Michael R Tschannen, Regan L Veith, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine*, 13(3):255–262, 2011.

[86] Minzhu Xie, Jing Li, and Tao Jiang. Detecting genome-wide epistases based on the clustering of relatively frequent items. *Bioinformatics*, 28(1):5–12, 2012.

[87] Xiang-Hong Xu, Shan-Shan Dong, Yan Guo, Tie-Lin Yang, Shu-Feng Lei, Christopher J Papasian, Ming Zhao, and Hong-Wen Deng. Molecular genetic studies of gene identification for osteoporosis: the 2009 update. *Endocrine reviews*, 31(4):447–505, 2010.

[88] Wonsuk Yoo, Brian A Ference, Michele L Cote, and Ann Schwartz. A comparison of logistic regression, logic regression, classification tree, and random forests to identify effective gene-gene and gene-environmental interactions. *International journal of applied science and technology*, 2(7):268, 2012.

[89] Patrizia Zavattari, Rosanna Lampis, Costantino Motzo, Miriam Loddo, Annapaola Mulargia, Michael Whalen, Mario Maioli, Efisio Angius, John A Todd, and Francesco Cucca. Conditional linkage disequilibrium analysis of a complex disease superlocus, iddm1 in the hla region, reveals the presence of independent modifying gene effects influencing the type 1 diabetes risk encoded by the major hla-dqb1,-drb1 disease loci. *Human molecular genetics*, 10(8):881–889, 2001.

[90] Dmitri V Zaykin, Alexander Pudovkin, and Bruce S Weir. Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics*, 180(1):533–545, 2008.

[91] Eleftheria Zeggini, Michael N Weedon, Cecilia M Lindgren, Timothy M Frayling, Katherine S Elliott, Hana Lango, Nicholas J Timpson, JR Perry, Nigel W Rayner, Rachel M Freathy, et al. Wellcome trust case control consortium (wtccc), mccarthy mi, hattersley at: Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341, 2007.

[92] Sihai Dave Zhao and Yi Li. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis*, 105(1):397–411, 2012.

[93] Wei Zhong and Liping Zhu. An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation*, (ahead-of-print):1–15, 2014.

[94] Manuela Zucknick, Sylvia Richardson, and Euan A Stronach. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical applications in genetics and molecular biology*, 7(1), 2008.