OPEN ACCESS Freely available online

PLOS ONE

Adaptive Ridge Regression for Rare Variant Detection

Haimao Zhan, Shizhong Xu*

Department of Botany and Plant Sciences, University of California Riverside, Riverside, California, United States of America

Abstract

It is widely believed that both common and rare variants contribute to the risks of common diseases or complex traits and the cumulative effects of multiple rare variants can explain a significant proportion of trait variances. Advances in high-throughput DNA sequencing technologies allow us to genotype rare causal variants and investigate the effects of such rare variants on complex traits. We developed an adaptive ridge regression method to analyze the collective effects of multiple variants in the same gene or the same functional unit. Our model focuses on continuous trait and incorporates covariate factors to remove potential confounding effects. The proposed method estimates and tests multiple rare variants collectively but does not depend on the assumption of same direction of each rare variant effect. Compared with the Bayesian hierarchical generalized linear model approach, the state-of-the-art method of rare variant detection, the proposed new method is easy to implement, yet it has higher statistical power. Application of the new method is demonstrated using the well-known data from the Dallas Heart Study.

Citation: Zhan H, Xu S (2012) Adaptive Ridge Regression for Rare Variant Detection. PLoS ONE 7(8): e44173. doi:10.1371/journal.pone.0044173

Editor: Momiao Xiong, University of Texas School of Public Health, United States of America

Received May 28, 2012; **Accepted** July 30, 2012; **Published** August 28, 2012

Funding: The project was supported by the USDA National Institute of Food and Agriculture Grant 2007–02784 to SX. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: shizhong.xu@ucr.edu

Introduction

Over the past decade, many causal polymorphic variants for common diseases have been successfully identified by genome-wide association studies (GWAS) [1–3] which are based on the common-disease-common-variant (CDCV) assumption. The associated variants greatly facilitate understanding of the genetic basis underlying common diseases. However, most association studies are used to identify common variants which have minor allele frequency (MAF) greater than 1%. This is mainly because traditional SNP genotyping arrays only capture variants with relatively high MAF. Although many genetic variants have been identified for common diseases, large proportion of the heritability of a trait cannot be explained by the detected variants. Many factors can lead to the missing heritability phenomenon: (1) underestimation of the effects of common variants, (2) undetectable common variants with small effects, and (3) rare variants [4,5]. The advance in sequencing technologies makes it possible to sequence some important candidate genes [6,7] and even the whole genome [8]. The available sequence information allows us to find out most variants across the genome and identify associated variants with different allele frequencies.

Meanwhile, it is found that genetic variants with low MAF, often called rare variants, may substantially contribute to phenotypic expression [9–14]. The missing heritability phenomenon in GWAS may be due to rare variants which are not captured by the traditional SNP genotyping arrays [5]. Therefore, identifying rare variants would help understand the genetic basis and disease etiology. Rare variants, which have lower minor allele frequencies compared to common variants, tend to have larger effects than common variants [9]. Many GWAS studies indicate that most identified common variants have odds ratio ranging from 1.1 to 1.3 with a mean odds ratio of 1.36. However, the mean odds ratio of rare variants is 3.74 and most rare variants have much greater odds ratio than common variants [9]. In addition, many non-synonymous rare mutations from exon sequencing are functional variants for some common diseases [9]. Many studies have been carried out to investigate the effects of rare variants by sequencing exons of candidate genes [6,7,11,12,15] and several rare variants have been found to be associated with common diseases. For example, rare variants in the IFIH1 gene were found to be strongly associated with Type I diabetes [15]. Some rare variants in ANGPTL3 and ANGPTL5 are much more common in the lowest quartile of plasma triglyceride level [7].

Statistical power of genetic variant identification depends on the sample size, the effect of the variant and the minor allele frequency [13,16]. Since the minor allele frequencies (MAF) of rare variants are extremely low (less than 1%∼5%), it is extremely challenging to identify the causal rare variants by using methods of traditional association studies [16–20]. The univariate tests (e.g. Chi-Squared test, Fisher's exact test, linear regression) have to take into account multiple test corrections to control family-wise error rate (FWER) and false discovery rate (FDR). Multiple-marker tests (e.g. multiple regression, Hotelling's T^2 test) increase the degree of freedom in hypothesis testing. Univariate test and multivariate test both loose power when the allele frequencies are very low [18].

Until recently, much effort has been placed on the development of new statistical methods for detecting rare variants. Most of the existing methods pool variants in the same group into one variant, which collectively combines the information from multiple variants and tries to increase the power of rare variant identification. For example, cohort allelic sum test (CAST) [21] combines the rare variants in the same region into a single "variant". The frequency of the pooled single variant can be compared between the case and control populations. The combined multivariate and collapsing (CMC) method [18]

collapses rare variants in the same group into one single "variant", and then counts the number of individuals carrying this marker in the case and control populations. The Hotelling's T^2 test is used to analyze the collapsed genotype data. The authors proved that the CMC method is much more powerful than the traditional single marker and multiple marker tests. It is also robust to the misclassification of rare variants. Morris and Zeggini [22] proposed two likelihood ratio tests (RVT1 and RVT2) based on different linear regression models to analyze rare variants. The first model treats the proportion of rare variants carrying at least one minor allele as predictor variable. In the second model, the indicator of presence or absence of the minor allele at any rare variant for one individual is used as the predictor in the linear model, which is similar to the collapsing method proposed by Li and Leal [18]. Madsen and Browning [19] weighed each rare variant by the minor allele frequency in unaffected individuals and found that this weighted approach can magnify the signal of rare variants. In that analysis, each individual was assigned a genetic score and the scores were ranked to test the significance of the association signal. Price et al. [23] adopted a variable-threshold approach to analyzing rare variants. They calculated a z-score for each reasonable threshold and find the maximum z-score. The significance of the z value was then tested using a permutation analysis.

All these so called burden test methods assume that the effects of rare variants are in the same direction. They collapse rare variants into a single variant and then compare the frequencies in the case and control populations. However, it is well known that genetic variants may not have effects on the phenotype of interests, and some of the variants may have beneficial effects and others have deleterious effects. Therefore, such assumptions seem to be inappropriate and collapsing rare variants in this way will introduce noise and decrease the power. Taking into account different directions of the variants will increase the power of rare variant detection [24,25].

More recently, Yi and Zhi [20] proposed a Bayesian hierarchical generalized linear model (BhGLM) to analyze rare variants without assuming known directions of the variant effects. They introduced two types of parameters to the regression model, a weight parameter for each variant and an overall effect for all variants in the same functional unit. The weights and the overall effect are estimated using the weighted least squares method that incorporates hierarchical prior information. As a result, the weights are estimated based on the contribution of the variants to the phenotype of interest. The association between rare variants and the phenotype of a target trait can be found by testing the significance of a single parameter, the overall effect (or score). Yi et al. [26] eventually proposed a similar method based on the BhGLM. The new method incorporates covariates and divides rare variants into several groups according to the minor allele frequencies and the functions of the variants. They also assigned hierarchical prior distributions to the weights of the variants and the groups. The association between the phenotype and the variants in the same group can be found by testing the group effects. These two methods do not assume known directions of the effects of the rare variants and can identify the collective effects of rare variants in the same group as well as individual variant effects. The BhGLM has a higher power than all the burden test methods. This new method is considered the state-of-the-art method for rare variant detection.

We believe that the high power of the BhGLM is due to (1) appropriate combination of the individual rare variants (the new score) and (2) assignment of the hierarchical priors. After a thorough evaluation of these new methods, we found that there is still some room for improvement. The new score of the BhGLM method is a first moment parameter (shared effect). An alternative score may be a second moment parameter (shared variance). The Yi and Zhi's method requires prior distributions, and thus different priors may produce different results. The hyper-parameters involved in the prior distributions may also affect the results. In this study, we proposed to use a shared variance among rare variants as the new score. The method is originally called ridge regression [27]. It is further modified to discriminate against rare variants with small effects. This modified ridge regression is called the adaptive ridge regression [28]. The adaptive ridge regression (ARR) is performed under the maximum likelihood framework, and thus prior distributions of parameters are not required, equivalent to independent uniform priors for all parameters.

Methods

The key issue in rare variant detection is to combine all rare variants into a single score (shared feature) and perform a single test for the collective effect of all rare variants. Our new method will be developed based on this notion. For the paper to be self-contained, we will briefly introduce ridge regression [27], based on which a new method called adaptive ridge regression will be developed.

Ridge regression

Let y_j be the phenotypic value of a quantitative trait measured from individual j for $j=1,\cdots,n$, where n is the sample size. The following linear model is used to describe the relationship between y_j and the rare variants,

$$y_j = X_j\beta + \sum_{k=1}^{m} Z_{jk}\gamma_k + \varepsilon_j \quad (1)$$

where m is the number of rare variants, X_j is a row vector representing the incidences of some covariates (fixed effects), β is a column vector for the fixed effects, Z_{jk} is a genotype indicator variable for marker k, γ_k is the effect of the kth rare variant, and $\varepsilon_j \sim N(0,\sigma^2)$ is the residual error with an assumed normal distribution. The genotype indicator variable is coded as

$$Z_{jk} = \begin{cases} +1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \\ -1 & \text{for } A_2A_2 \end{cases} \quad (2)$$

where A_1 is the rare allele and A_2 is the common allele of locus k. This coding system is adopted from that of QTL mapping [29]. An alternative coding system is

$$Z_{jk} = \begin{cases} +1 & \text{for } A_1A_1 \text{ or } A_1A_2 \\ -1 & \text{for } A_2A_2 \end{cases} \quad (3)$$

where a genotype carrying one or two rare alleles is coded as 1 and the common allele homozygote is coded as -1. Because of the rare frequency of allele A_1, the probability of A_1A_1 is negligible and thus the population virtually consists of only the heterozygote and the homozygote of the common allele. The second coding system has an advantage of saving computer storage. In this study, we took the first coding system, i.e., equation (2).

We assume that the effect of each rare variant is sampled from a normal distribution with mean zero and a common variance, i.e., $\gamma_k \sim N(0,\phi^2)$ for $k=1,\cdots,m$. By doing this, we can evaluate the

shared nature of the rare variants, i.e., they all come from the same distribution with the same mean and the same variance, so that a test statistic can be derived to test this single variance. If $\phi^2=0$, none of the rare variants are associated with the trait of interest. This particular formulation treats the cofactors as fixed effects and the rare variants as random effects. If ϕ^2 is a predetermined constant, the method is called ridge regression analysis with a ridge factor [27] of $\lambda=\sigma^2/\phi^2$. However, we can estimate ϕ^2 from the data under the mixed model framework [30], in which the expectation is $\mathrm{E}(y)=X\beta$ and the variance covariance matrix of the phenotypic values is

$$\mathrm{var}(y)=V=ZZ^T\phi^2+I\sigma^2 \tag{4}$$

where $Z=\{Z_{jk}\}$ is an $n\times m$ matrix for the genotype indicators of all subjects for all rare variants. The maximum likelihood or restricted maximum likelihood methods can be used to estimate the parameters $\theta=\left\{\beta,\phi^2,\sigma^2\right\}$ and test $H_0:\phi^2=0$. Rejection of H_0 leads to a conclusion that these rare variants are collectively associated with the trait. The likelihood ratio test statistic may be used to test H_0,

$$\xi=-2\left[L(\tilde{\beta},\tilde{\sigma}^2)-L(\hat{\beta},\hat{\phi}^2,\hat{\sigma}^2)\right] \tag{5}$$

where $\hat{\theta}=\left\{\hat{\beta},\hat{\phi}^2,\hat{\sigma}^2\right\}$ represents the ML estimate of θ under the full model and $\tilde{\theta}=\left\{\tilde{\beta},\tilde{\sigma}^2\right\}$ represents the ML estimate of θ under the null model. When the sample size is sufficiently large, ξ follows approximately a chi-square distribution with one degree of freedom (χ_1^2). Crainceanu and Ruppert [31] stated that ξ follows asymptotically a mixture of two chi-square distributions, denoted by $0.5\chi_0^2+0.5\chi_1^2$. We will use both distributions to draw the critical values for the test statistic. In addition, we will also use permutation tests [32] to find the empirical distribution of the likelihood ratio test statistic and thus the empirical threshold for the controlled Type I error rate. Therefore, the asymptotic chi-square distributions may not be required in real data analysis.

The whole purpose of rare variant analysis is to test all the rare variants collectively using the likelihood ratio test statistic for $H_0:\phi^2=0$. However, each individual rare variant can also be estimated and tested using the mixed model equation [30] with θ substituted by $\hat{\theta}$. The mixed model equation is

$$\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\end{bmatrix}=\begin{bmatrix}X^TX & X^TX\\ Z^TX & Z^TZ+\hat{\lambda}I\end{bmatrix}^{-1}\begin{bmatrix}X^Ty\\ Z^Ty\end{bmatrix} \tag{6}$$

where $\hat{\lambda}=\hat{\sigma}^2/\hat{\phi}^2$. Let us define a C matrix by

$$\begin{aligned}C&=\mathrm{var}\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\end{bmatrix}\\ &=\hat{\sigma}^2\begin{bmatrix}X^TX & X^TX\\ Z^TX & Z^TZ+\hat{\lambda}I\end{bmatrix}^{-1}\\ &=\begin{bmatrix}C_{\beta\beta} & C_{\beta\gamma}\\ C_{\gamma\beta} & C_{\gamma\gamma}\end{bmatrix}\end{aligned} \tag{7}$$

The sub matrix $C_{\gamma\gamma}$ is the variance matrix for all the rare variants. Given the estimated γ_k and its estimation error $s_k=\sqrt{\mathrm{var}(\hat{\gamma}_k)}$, a test statistic can be drawn,

$$W_k=\frac{\hat{\gamma}_k^2}{\mathrm{var}(\hat{\gamma}_k)} \tag{8}$$

from which a p-value can be found, assuming that (under the null model) W_k is approximately distributed as a chi-square variable with one degree of freedom. Again, in rare variant detection, our main purpose is to test $\phi^2=0$, and thus testing an individual rare variant is only a by-product of the analysis.

Under the ridge regression method, each rare variant is treated as a random variable. The shared feature is the common variance denoted by ϕ^2, This particular treatment has eliminated the assumption of same direction of rare variant effects in all the burden test methods described in the introduction.

Adaptive ridge regression

If only a few variants are associated with the trait, then ϕ^2 will be "diluted" by those non-associated variants. This means that the assumption of a common variance is violated. We now propose an adaptive ridge regression to selectively weigh each rare variant. The modified model is formulated as

$$y_j=X_j\beta+\sum_{k=1}^{m}Z_{jk}c_k\gamma_k+\varepsilon_j \tag{9}$$

where c_k is another effect for variant k. It seems redundant to define two effects for each rare variant. However, the two effects, c_k and γ_k, have different meanings. The first effect c_k may be defined as a random effect with a marker specific variance, i.e., $c_k\sim N(0,\tau_k^2)$ for $k=1,\cdots,m$. The second effect γ_k is defined as a random effect with a common (shared) variance, i.e., $\gamma_k\sim N(0,\phi^2)$ for $k=1,\cdots,m$. Because c_k and γ_k have different variances, they can be estimated separately. This partitioning of the rare variant effect appears to be similar to the model proposed by Yi and Zhi [20] but they differ in a fundamental way. Using our notation, their model can be expressed as

$$y_j=X_j\beta+\sum_{k=1}^{m}(Z_{jk}c_k)\alpha+\varepsilon_j \tag{10}$$

in which the authors proposed a common effect α (a single first moment parameter) for all variants. We proposed marker specific effect γ_k but with a common variance ϕ^2. Our model is more like the polygenic model for quantitative traits, where each γ_k is a polygenic effect with a common genetic variance ϕ^2.

We now discuss parameter estimation for the new adaptive ridge regression method. The method is based on the classical mixed model methodology. Given the value of each c_k, we can rewrite the adaptive ridge regression model as

$$y_j=X_j\beta+\sum_{k=1}^{m}Z_{jk}^*\gamma_k+\varepsilon_j \tag{11}$$

where $Z_{jk}^*=Z_{jk}c_k$ is a weighted independent variable. We can estimate ϕ^2 and predict γ_k under the original ridge regression procedure (mixed model methodology) described in the previous section with Z_{jk} substituted by Z_{jk}^*. Hypothesis test for $H_0:\phi^2=0$ can also be performed under the mixed model framework. Each

individual marker effect is actually redefined as $\gamma_k^* = c_k\gamma_k$ and the Wald test statistic remains the same as described before,

$$W_k = \frac{\hat{\gamma}_k^{*2}}{\mathrm{var}(\hat{\gamma}_k^*)} = \frac{c_k^2\hat{\gamma}_k^2}{c_k^2\mathrm{var}(\hat{\gamma}_k)} = \frac{\hat{\gamma}_k^2}{\mathrm{var}(\hat{\gamma}_k)} \quad (12)$$

The question left is how to find c_k so that the adaptive ridge regression can selectively adjust Z_{jk} to prevent ϕ^2 from being diluted by the non-associated rare variants. There are many different ways to estimate c_k. For example, we may use an iterative approach to estimating c_k given γ_k by reformulating the model as

$$y_j = X_j\beta + \sum_{k=1}^{m} Z_{jk}^* c_k + \varepsilon_j \quad (13)$$

where $Z_{jk}^* = Z_{jk}\gamma_k$ is a newly weighted independent variable. Given c_k to estimate γ_k and given γ_k to estimate c_k require iterations. The iteration process continues until the sequence converges. Estimating c_k given γ_k using this approach may be complicated because each c_k has its own variance. For m markers, we need to estimate m marker specific effects c_k and m marker-specific variances τ_k^2 for $k = 1, \cdots, m$. We will leave this approach as a next project and pursue a simple method to estimate c_k.

In this study, we adopted a different method for estimating c_k. This approach leads to the Lasso [33] estimate of γ_k if c_k is restricted in a special way. Grandvalet [28] demonstrated that if we let $c_k \geq 0$ and enforce the following constraint

$$\sum_{k=1}^{m} c_k^2 = m \quad (14)$$

the solution for γ_k is the Lasso (least absolute shrinkage and selection operator) estimate of γ_k, assuming that β is removed from the model by centralization of the data and $\lambda = \sigma^2/\phi^2$ is a constant provided by the investigator prior to the analysis. In our problem, we also estimate $\theta = \left\{\beta, \phi^2, \sigma^2\right\}$ using the mixed model methodology. The Lasso parameter is estimated as a by-product because $\lambda = \sigma^2/\phi^2$ is estimated from the data, not predetermined by the investigator. Whether such a solution of γ_k is still Lasso or not is unknown. Any way, we adopted the approach of Grandvalet [28] to find c_k given γ_k, which is

$$c_k^2 = \frac{m\gamma_k^2}{\sum_{k=1}^{m}\gamma_k^2} \quad (15)$$

During the iteration process, if any $c_k^2 \leq 10^{-5}$ happens, c_k is set to zero and the corresponding Z_{jk} will be deleted from the model permanently so that the dimension of the model will be quickly reduced to the number of non-zero elements of vector $c = \{c_k\}$. The method is not sensitive to the threshold. We tested several other thresholds, e.g., 10^{-3} and 10^{-8}. The results are much the same (data not shown). In fact, we do not need to set this threshold. The purpose of using this threshold is to improve the computational speed, because once an effect is set to zero, this effect will no longer be evaluated in the subsequent iterations.

Adaptive ridge for multiple groups of rare variants

The model can be extended to handle multiple groups of rare variants. The notation become complicated so we have to use a compact matrix notation for the model. For a single group, we may use

$$y = X\beta + Z\gamma + \varepsilon \quad (16)$$

where $Z = \{Z_{jk}\}$ is an $n \times m$ matrix and $\gamma = \{\gamma_k\}$ is an $m \times 1$ vector. Assume that we now have g groups and the number of markers in the lth group is m_l, where $\sum_{l=1}^{g} m_l = m$ is the total number of markers. We now label the appropriate matrices by a subscript l to indicate the group, the extended model becomes

$$y = X\beta + \sum_{l=1}^{g} Z_l\gamma_l + \varepsilon \quad (17)$$

where Z_l is an $n \times m_l$ matrix and γ_l is an $m_l \times 1$ vector. The group specific γ_l is assumed to be $N(0, \phi_l^2 I)$ distributed. The variance matrix is

$$V = \sum_{l=1}^{g} Z_l Z_l^T \phi_l^2 + I\sigma^2 \quad (18)$$

The parameter vector is $\theta = \left\{\beta, \phi_1^2, \cdots, \phi_g^2, \sigma^2\right\}$, which can be estimated using the maximum likelihood method [34,35]. Once θ is estimated, the effects of rare variants are estimated using the mixed model equations. When $g = 2$, for example, the mixed model equations are

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{bmatrix} = \begin{bmatrix} X^TX & X^TZ_1 & X^TZ_2 \\ Z_1^TX & Z_1^TZ_1 + \hat{\lambda}_1 I & Z_1^TZ_2 \\ Z_2^TX & Z_2^TZ_1 & Z_2^TZ_2 + \hat{\lambda}_2 I \end{bmatrix}^{-1} \begin{bmatrix} X^Ty \\ Z_1^Ty \\ Z_2^Ty \end{bmatrix} \quad (19)$$

where $\hat{\lambda}_l = \hat{\sigma}^2/\hat{\phi}_l^2$ for $l = 1,2$. We now have g different weight systems, one for each group. Let c_l be an $m_l \times 1$ vector of weights for group l. It is defined by

$$c_l = \sqrt{\frac{m_l}{\gamma_l^T\gamma_l}}|\gamma_l| \quad (20)$$

where $|\gamma_l|$ is an $m_l \times 1$ vector of the absolute values of γ_l because of the constraint $c_l \geq 0$. The weighted Z matrix is defined as $Z_l^* = Z_l\mathrm{diag}(c_l)$, where $\mathrm{diag}(c_l)$ is a diagonal matrix with the diagonal elements filled by the values of vector c_l. The adaptive ridge regression for multiple group rare variants is performed simply with Z_l substituted by Z_l^*.

Many different hypotheses can be tested for the multiple group variant analysis. An overall hypothesis is $H_0 : \phi_1^2 = \ldots = \phi_g^2 = 0$, which can be tested using the likelihood ratio test statistics. Under the null hypothesis, the test statistic follows asymptotically a chi-square distribution with g degrees of freedom. Each individual group can also be tested. For example, to test the lth group, one needs to evaluate a reduced model by excluding ϕ_l from the model and defining a likelihood ratio test statistic. Under the null hypothesis $H_l : \phi_l^2 = 0$, the test statistic follows asymptotically a chi-square distribution with one degree of freedom. In this study, we used the permutation test to draw the empirical threshold values of the likelihood ratio test statistics for a given Type I error rate. As a result, the asymptotic chi-squares critical values are not used in real data analysis.

Table 1. Parameters of three genes of the Dallas Heart Study estimated separately using the ARR method proposed in this study.

Parameter	*ANGPTL3*	*ANGPTL4*	*ANGPTL5*
Intercept (β_1)	4.201 ± 0.112[1]	4.152 ± 0.175	4.304 ± 0.069
Age (β_2)	0.009 ± 0.001	0.009 ± 0.001	0.009 ± 0.001
Gender (β_3)	-0.088 ± 0.009	-0.089 ± 0.009	-0.087 ± 0.009
Race 1 (β_4)	0.142 ± 0.024	0.135 ± 0.024	0.139 ± 0.024
Race 2 (β_5)	-0.214 ± 0.021	-0.206 ± 0.021	-0.230 ± 0.021
Race 3 (β_6)	0.020 ± 0.022	0.012 ± 0.022	0.027 ± 0.022
Residual variance (σ^2)	0.310	0.310	0.311
Genetic variance (ϕ^2)	0.026	0.077	0.010
Likelihood ratio test (ξ)	7.248	9.935	0.000
Theoretical *p*-value (χ_1^2)[2]	7.10×10^{-3}	1.62×10^{-3}	1.000
Theoretical *p*-value ($0.5\chi_0^2 + 0.5\chi_1^2$)[3]	3.39×10^{-3}	6.80×10^{-4}	1.000
Empirical *p*-value (permutation)[4]	0.029	0.001	1.000

[1]The numbers after $\pm$ for the six fixed effects are the standard errors.
[2]The theoretical *p*-value (χ_1^2) for each gene was calculated using a threshold of 3.84 for the test statistic.
[3]Theoretical *p*-value ($0.5\chi_0^2 + 0.5\chi_1^2$) for each gene was calculated using a threshold of 2.71 for the test statistic.
[4]The empirical *p*-value (permutation) was calculated using a threshold drawn from the permutation study.
doi:10.1371/journal.pone.0044173.t001

Results

Application to the Dallas Heart Study Data

Angiopoietin-like proteins (*ANGPTLs*) [36–40] can regulate triglyceride metabolism by inhibiting the activity of lipoprotein lipase. Romeo et al. [6,7] resequenced the exons and some intronic regions of *ANGPTL3* (MIM 604774), *ANGPTL4* (MIM 605910) and *ANGPTL5* (MIM 607666) genes in 3,551 individuals from a multiethnic population (601 Hispanic, 1,830 African American, 1,045 European American and 75 others). They wanted to find the sequence variants which have effects on the regulation of plasma triglyceride level. For the three genes, *ANGPTL3*, *ANGPTL4* and *ANGPTL5*, a total of 282 sequence variants (SNP) were genotyped (88 variants in *ANGPTL3*, 94 variants in *ANGPTL4* and 100 variants in *ANGPTL5*). In addition to triglyceride level and race, gender and age were also recorded for each individual. To test the effects of sequence variants, age, gender and race were treated as covariates in the adaptive ridge regression analysis. Since there are some missing data in age, all missing values for age were replaced by the mean age of all subjects. The original phenotypic value (triglyceride level) was log transformed prior to the analysis, as did by the original authors.

In the population of the Dallas Heart Study, the minor allele frequency of rare variants ranges from 0.014% to 37.9%. Most of the sequence variants have MAF less than 1%. Therefore, the data contain many rare variants as well as a few common variants.

Since there are three genes, variants within the same gene are considered in one group. The way of grouping variants can be arbitrary. Variants may be grouped based on the minor allele frequencies or predicted biological functions [26]. We may also define groups based on physical locations of the rare variants. It is reasonable to analyze the variants in the same gene as one group because variants in the same gene may work systematically as a unit.

First, we analyzed the rare variants one group (gene) at a time. The adaptive ridge regression tests the group effect of variants in the same gene. The single group model is

$$y_j = \sum_{i=1}^{6} X_{ji}\beta_i + \sum_{k=1}^{m} Z_{jk}c_k\gamma_k + \varepsilon_j \qquad (21)$$

where the six covariates represent the intercept (β_1), age effect (β_2), gender effect (β_3), and three effects for the race (β_4,β_5 and β_6). Note that there are four ethnic groups, but only three estimable effects, which explains why we have three fixed effects for the race factor alone. The number of markers m takes 88, 94 and 100, respectively, for the three genes. We reported the empirical *p*-value for each gene (group) drawn from permutation analysis (1000 permuted samples) along with the theoretical *p*-values from χ_1^2 and $0.5\chi_0^2 + 0.5\chi_1^2$ distributions. In the permutation analyses, we kept the marker genotype data intact but reshuffled the phenotype along with the six covariates (fixed effects) across all the subjects. This permutation analysis only destroyed the association between the phenotype and the markers and did not destroy the association between the phenotype and the covariates. The estimated parameters along with the test statistic and the *p*-values are listed in Table 1 for the adaptive ridge regression method. Genes *ANGPTL3* and *ANGPTL4* had *p*-values smaller than 0.05, and thus rare variants of these two genes are collectively associated with triglyceride level. Variants of gene *ANGPTL5* are not associated with the trait at all because the *p*-value is 1.00. Estimated individual marker effects will be reported later when the joint analysis of three genes are reported. We also analyzed the three genes separately (one gene at a time) using BhGLM [26]. The results of BhGLM are listed in Table 2 and they are similar to our ARR analysis. Genes *ANGPTL3* and *ANGPTL4* are strongly associated with the triglyceride level but gene *ANGPTL5* is not.

We now report results of joint analysis for the three genes in the Dallas Heart Study. First, we used the adaptive ridge regression method to analyze the three genes jointly. This time, we have four hypotheses to test. The overall test for all three genes, $H_0 : \phi_1^2 = \phi_2^2 = \phi_3^2 = 0$, and a test for each gene, i.e., $H_1 : \phi_1^2 = 0$, $H_2 : \phi_2^2 = 0$ and $H_3 : \phi_3^2 = 0$. The *p*-value of each test was calculated using the permutation generated empirical threshold

Table 2. Parameters of three genes of the Dallas Heart Study estimated separately using the BhGLM method.

Parameter	*ANGPTL3*	*ANGPTL4*	*ANGPTL5*
Intercept (β_1)	6.051 ± 0.523[1]	7.211 ± 0.528	3.677 ± 0.512
Age (β_2)	0.009 ± 0.001	0.009 ± 0.001	0.009 ± 0.001
Gender (β_3)	-0.088 ± 0.009	-0.088 ± 0.009	-0.088 ± 0.009
Race 1 (β_4)	0.144 ± 0.024	0.133 ± 0.024	0.137 ± 0.024
Race 2 (β_5)	-0.220 ± 0.020	-0.214 ± 0.020	-0.224 ± 0.020
Race 3 (β_6)	0.024 ± 0.022	0.016 ± 0.022	0.022 ± 0.022
Residual variance (σ^2)	0.311	0.309	0.312
Overall score (α)	0.020 ± 0.006	0.038 ± 0.007	-0.007 ± 0.005
Wald test (W)	11.062	30.250	1.588
Theoretical p-value (χ^2_1)[2]	8.81×10^{-4}	3.80×10^{-8}	0.208
Empirical p-value (permutation)[3]	0.033	0.000	0.355

[1]The numbers after $\pm$ for the six fixed effects are the standard errors.
[2]The theoretical p-value (χ^2_1) for each gene was calculated using a threshold of 3.84 for the test statistic.
[3]The empirical p-value (permutation) was calculated using a threshold drawn from the permutation study.
doi:10.1371/journal.pone.0044173.t002

value of the corresponding test statistic (1000 permuted samples) and the theoretical p-values from χ^2 and mixture χ^2 distributions. The results are listed in Table 3. First, the estimated parameters of the joint analysis regarding the model (e.g., fixed effects and the residual error variance) are similar to the separate analyses shown in Table 1. The estimated gene specific parameters and the p-values showed that genes *ANGTPL3* and *ANGTPL4* are associated with triglyceride level but *ANGTPL5* is not. The overall test for the three genes is also significant. This time, we also report the p-value for each individual rare variant, as shown in Figure 1 (the top panels). In fact, it is the $-\log_{10}(p)$ value that is plotted against the markers. One marker (8357_non_coding) in gene *ANGPTL3* is significant ($p<0.05$ and $-\log_{10}(p)>1.301$). Two rare variants (1313_E40K and 8191_R278Q) in *ANGPTL4* are significant. No variants are significant in gene *ANGPTL5*.

We also analyzed the three genes jointly using BhGLM. The results of this analysis are listed in Table 4. The general conclusions are the same as the ARR analysis, *ANGPTL3* and *ANGPTL4* are associated with triglyceride level but *ANGPTL5* is not. The plot of $-\log_{10}(p)$ against the markers for the BhGLM analysis is presented in Figure 1 (bottom panels). The same variant (8357_non_coding) in gene *ANGPTL3* is also significant here. The two significant rare variants (1313_E40K and 8191_R278Q) detected for gene *ANGPTL4* with the ARR analysis are also significant in the BhGLM analysis. In addition, two more rare variants (1175_Intronic and 8279_P307P) are detected by the BhGLM analysis. Each of the two additional rare variants is closely linked to one of the previously detected ones by the ARR analysis. The rare variant named 1313_E40K, detected by both the ARR and BhGLM methods, was also found to be significantly associated with plasma triglyceride level in different population-based studies (Dallas Heart Study, Atherosclerosis Risk in Communities Study and Copenhagen City Heart Study) [6]. The 1313_E40K carriers had lower triglyceride level than the non-carriers.

The analyses of Dallas Heart Data showed that the method is not sensitive to the number of groups, because results of separate and joint analyses are much the same. Because ridge regression is a random model approach, the method is also not sensitive to the number of markers within each group. In fact, the number of markers of each group can be more than the sample size. This is the beauty of the random model approach, which the fixed model lacks.

Simulation Studies

Instead of using population genetics models to generate the genotype data, we used the sequence data from the Dallas Heart Study for the simulation studies without making any assumption about the rare variants. The real sequence variants are believed to be more appropriate in the simulations [16,26]. The covariates such as age, gender and race were included in the model. The effects of the fixed effects (including the intercept) and the residual error variance used to generate the data took the estimated values obtained from the ARR analysis for gene *ANGPTL4*. The marker genotypes took the genotypes of the 94 variants of this gene only. The purpose of the simulation study is to evaluate the empirical Type I error and statistical power. Therefore, we only used the single gene model to evaluate the Type I errors and powers of the ARR method and BhGLM for comparison.

Simulation of Type I Error Rate. The phenotypic value for each of the $n=3551$ subjects was generated using the following model,

$$y_j = \sum_{i=1}^{6} X_{ji}\beta_i + \varepsilon_j \qquad (22)$$

where the X_{ji}'s are covariates of the Dallas Heart Study, the β_i's are fixed effects estimated in the analysis of *ANGPTL4* and ε_j was simulated from a $N(0,\sigma^2)$ distribution with σ^2 taking the value estimated from the analysis of gene *ANGPTL4*. This model assumes zero effects for all the 94 variants and thus any association of the variants is a false positive. The simulated data were then subject to the same analysis as the real data described early with all the 94 variants included in the model. For the ARR analysis, we used two criteria to evaluate the Type I error rate. In the first criterion, we choose $\chi^2_{0.05,1}=3.84$ as the threshold value for the likelihood ratio statistic, above which the gene (group of variants) was declared as significance. The threshold value of the likelihood test statistic for the second criterion was $0.5\chi^2_{\alpha_0,0}+0.5\chi^2_{\alpha_1,1}=2.71$, where $\alpha_0+\alpha_1=0.05$. The simulation was replicated 1000 times. Under each criterion, the number of replicates whose test statistics

reached the corresponding threshold was counted as the number of false positives. This number divided by 1000 is the actual (observed) Type I error rate. If the observed Type I error rate is below 0.05, we then conclude that the Type I error rate is under control (may be conservative). The criterion that is closer to 0.05 should be recommended.

We suspected that the actual Type I error rate is related to the number of rare variants included in the model. Therefore, we evaluated the Type I errors under different numbers of variants included, starting from $m=3$ and progressively increased to $m=94$. Under each model size (m), 1000 replicated simulations were conducted and the actual Type I error rate was observed. Figure 2 shows the observed Type I error rate plotted against the number of rare variants included in the model for the ARR method along with the BhGLM method [26]. For the BhGLM method, the test statistic was the Wald test statistic and the threshold value for the Wald test is also approximated by $\chi^2_{0.05,1}=3.84$. From Figure 2, we can see that the actual Type I error rate is indeed related to the model size, large models tend to give higher Type I error rates. However, according to $\chi^2_{0.05,1}=3.84$, the Type I error rate for the ARR method is under control (all below the expected 0.05 probability). Therefore, this criterion may be too conservative. For the $0.5\chi^2_{\alpha_0,0}+0.5\chi^2_{\alpha_1,1}=2.71$ criterion, the observed Type I error rate is much closer to the expected 0.05 probability. When $m=60$ and 80, the Type I error rates are exactly 0.05. However, when m reaches 94, the observed Type I error rate is slightly higher than the 0.05 probability. The observed Type I error rate for the BhGLM method, however, is out of control for all model sizes examined except when only $m=3$ variants were included in the model.

Our conclusions from this simulation experiment are (1) the $0.5\chi^2_{\alpha_0,0}+0.5\chi^2_{\alpha_1,1}=2.71$ criterion is recommended for the likelihood ratio test (in the ARR analysis), (2) the $\chi^2_{0.05,1}=3.84$ criterion for the likelihood ratio test is over conservative and may be more preferable to some investigators (in the ARR analysis) and (3) the $\chi^2_{0.05,1}=3.84$ criterion for the Wald test statistics is too liberal (out of control for the Type I error rate) in the BhGLM analysis. To control the Type I error rate under 0.05 for the BhGLM, the threshold level should be further increased. The Type I error analysis is useful if permutation analysis is not performed. We realized that BhGLM allows users to choose their own prior distribution. The program also provides a set of default priors. We simply used the default priors, which may partly explain the high Type I error rates.

Simulation of Empirical Power. Again, we used the 94 variants of *ANGPTL4* as the true genotype data for the power analysis. The estimated fixed effects and effects of the 94 variants for *ANGPTL4* from the ARR analysis were used as the true values for the power analysis. The residual variance will determine the

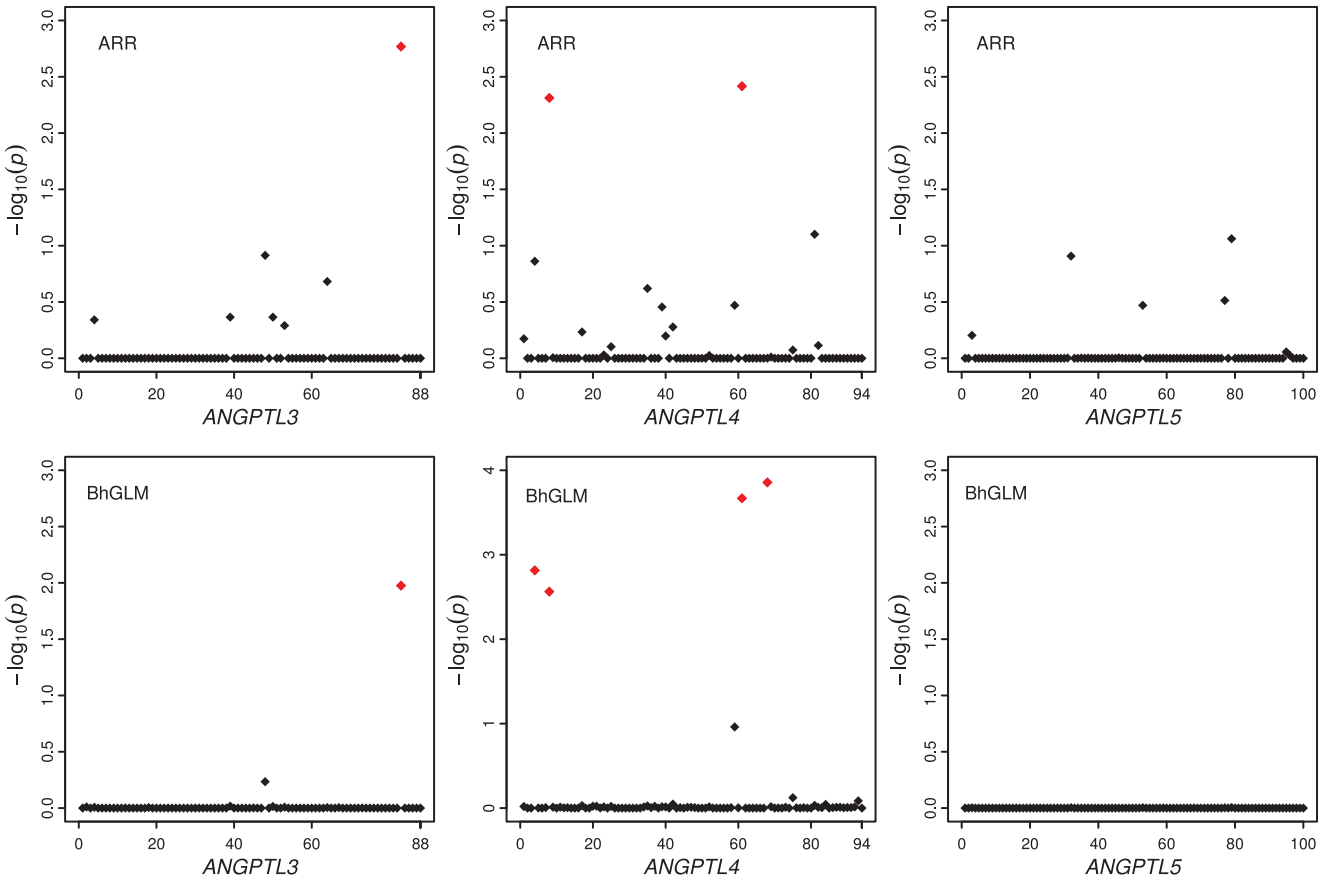

Figure 1. Significance level for each marker in *ANGPTL3*, *ANGPTL4* and *ANGPTL5* generated from the joint analyses. *P* value is shown on the $-\log_{10}$ scale. The top panels show the result of the adaptive ridge regression (ARR) analysis and the bottom panels show the results of the Bayesian hierarchical generalized linear model (BhGLM) analysis. The red dots represent variants with *p*-values smaller than 0.05, i.e., $-\log_{10}(p)>1.301$.
doi:10.1371/journal.pone.0044173.g001

Table 3. Parameters of three genes of the Dallas Heart Study estimated jointly using the ARR method proposed in this study.

Parameter	*ANGPTL3*	*ANGPTL4*	*ANGPTL5*
Gene specific information			
(ϕ_l^2)Variance component	0.023	0.059	0.009
Likelihood ratio test (ξ_l)	11.051	13.733	0.979
Theoretical *p*-value (χ_1^2)	8.86×10^{-4}	2.11×10^{-4}	0.322
Theoretical *p*-value ($0.5\chi_0^2+0.5\chi_1^2$)	4.60×10^{-4}	1.30×10^{-4}	0.162
Empirical *p*-value (permutation)	0.027	0.010	0.558
Model information			
Intercept (β_1)	4.032 ± 0.181		
Age (β_2)	0.009 ± 0.001		
Gender (β_3)	-0.088 ± 0.009		
Race 1 (β_4)	0.138 ± 0.024		
Race 2 (β_5)	-0.208 ± 0.022		
Race 3 (β_6)	0.016 ± 0.022		
Residual variance (σ^2)	0.307		
Likelihood ratio test (ξ)	21.269		
Theoretical *p*-value (χ_3^2)	9.26×10^{-5}		
Theoretical *p*-value ($0.5\chi_2^2+0.5\chi_3^2$)	8.00×10^{-5}		
Empirical *p*-value (permutation)	0.014		

doi:10.1371/journal.pone.0044173.t003

proportion of the phenotypic variance explained by the rare variants. Recall that the linear model for *ANGPTL4* is

$$y_j = X_j\beta + \sum_{k=1}^{94} Z_{jk}\gamma_k + \varepsilon_j \qquad (23)$$

where X_j and Z_{jk} were chosen from gene *ANGPTL4* from the Dallas Heart Study, β and γ_k took values estimated from the ARR method, and $\varepsilon_j\tilde{N}(0,\sigma^2)$ is the residual error with variance σ^2. The genetic value for individual j is defined as

$$a_j = \sum_{k=1}^{94} Z_{jk}\gamma_k \qquad (24)$$

Table 4. Parameters of three genes of the Dallas Heart Study estimated jointly using the BhGLM method.

Parameter	*ANGPTL3*	*ANGPTL4*	*ANGPTL5*
Gene specific information			
(α_l)Group effect	0.017 ± 0.005	0.049 ± 0.009	-0.006 ± 0.005
Wald test (W_l)	9.716	30.692	1.266
Theoretical *p*-value (χ_1^2)	1.83×10^{-3}	3.02×10^{-8}	0.261
Empirical *p*-value (permutation)	0.031	0.000	0.390
Model information			
Intercept (β_1)	9.066 ± 0.987		
Age (β_2)	0.009 ± 0.001		
Gender (β_3)	-0.088 ± 0.009		
Race 1 (β_4)	0.139 ± 0.024		
Race 2 (β_5)	-0.218 ± 0.020		
Race 3 (β_6)	0.018 ± 0.022		
Residual variance (σ^2)	0.308		
Wald test (W)	41.673		
Theoretical *p*-value (χ_3^2)	4.71×10^{-9}		
Empirical *p*-value (permutation)	0.000		

doi:10.1371/journal.pone.0044173.t004

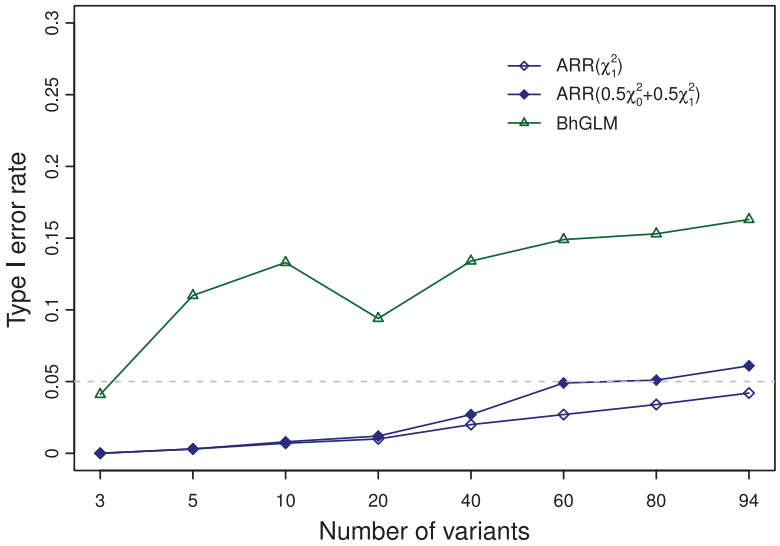

Figure 2. Type I error rates of the ARR and BhGLM methods obtained from the simulation studies. The Type I error rate of the Bayesian hierarchical generalized linear model (BhGLM) method was calculated using a threshold of 3.84 for the Wald test statistic. The Type I error rates of the adaptive ridge regression (ARR) method were calculated using thresholds of 3.84 and 2.71, respectively, corresponding to the $\chi^2_{0.05,1}$ and $0.5\chi^2_{\alpha_0,0}+0.5\chi^2_{\alpha_1,1}$ criteria ($\alpha_0+\alpha_1=0.05$).
doi:10.1371/journal.pone.0044173.g002

The genetic variance is defined as the variance of a_j across all individuals, as shown below,

$$\sigma^2_G=\mathrm{var}(a)=\frac{1}{3551}\sum_{j=1}^{3551}(a_j-\bar{a})^2=0.0021071 \qquad (25)$$

The total phenotypic variance is

$$\sigma^2_P=\sigma^2_G+\sigma^2=0.0021071+\sigma^2 \qquad (26)$$

in which the variance due to the covariates (fixed effects) has been removed. The heritability of the trait is

$$h^2=\frac{\sigma^2_G}{\sigma^2_P}=\frac{\sigma^2_G}{\sigma^2_G+\sigma^2}=\frac{0.0021071}{0.0021071+\sigma^2} \qquad (27)$$

We choose several different values of σ^2 to control the heritability at the following levels, 0.6%, 0.8%, 1%, 2% and 3%. At each level of the heritability, we calculated σ^2 and used σ^2 to simulate a random residual error to add to the fixed effect and the genetic effect to generate a phenotypic value y_j. At each level of the heritability, the simulation was replicated 1000 times. The empirical statistical power was then obtained by counting the proportion of the replicated samples that are significant over the 1000 replicates. For the ARR method, three criteria were used to determine the threshold values for the likelihood ratio test statistic. They are $\chi^2_{0.05,1}=3.84$, $0.5\chi^2_{\alpha_0,0}+0.5\chi^2_{\alpha_1,1}=2.71$ and $\xi_{0.05}=3.45$. The last one, $\xi_{0.05}=3.45$, was obtained from the simulation experiment in the section of Type I error rate study. To compare the power of the ARR analysis with that of the BhGLM analysis, the same datasets were also analyzed using the BhGLM program. The power of BhGLM was determined using two criteria, $\chi^2_{0.05,1}=3.84$ for the Wald test and $W_{0.05}=9.78$ obtained from the null model simulation study (Type I error rate study). We knew that using the $\chi^2_{0.05,1}=3.84$ criterion would overestimate the power for the BhGLM method because the actual Type I error rate for the BhGLM analysis was much higher than 0.05. The power analysis showed that using the theoretical threshold $\chi^2_{0.05,1}=3.84$, the BhGLM method appears to be more powerful than the ARR method (see Figure 3, panel a). However, the ARR analysis based on $0.5\chi^2_{\alpha_0,0}+0.5\chi^2_{\alpha_1,1}=2.71$ has almost the same power as BhGLM at different levels of heritability (see Figure 3, panel b). When the empirical thresholds are used (drawn from the Type I error rate study), the ARR method is more powerful than the BhGLM method (see Figure 3, panel c). A permutation generated threshold for the BhGLM method should be used in real data analysis because the Type I error rate cannot be controlled using the theoretical threshold.

Discussion

The adaptive ridge regression method was developed based on the original ridge regression [27]. The purpose of the adaptation is to selectively weigh each rare variant based on its size, denoted by c_k for the kth rare variant, so that the overall genetic variance ϕ^2 is not "diluted" by the non-associated variants. The adaptive ridge regression requires just a few iterations to converge. Figure 4 shows the iteration process of the p-values calculated from the $0.5\chi^2_{\alpha_0,0}+0.5\chi^2_{\alpha_1,1}=2.71$ criterion for the three genes (*ANGPTL3*, *ANGPTL4* and *ANGPTL5*) analyzed separately by the ARR method. For gene ANGPTL3, the p-value of the initial step (ridge regression without adaptation) is greater than 0.05. Just one step of adaption, the p-value has dropped to the 0.05 significance level. For gene *ANGPTL4*, the p-value of the initial step (ridge regression without adaptation) is already lower than the 0.05 probability. Further iterations continue to drop the p-value. For gene *ANGPTL5*, the p-value is very high and remains high after iterations. This figure clearly shows the necessity of the adaptive steps for rare variant detection.

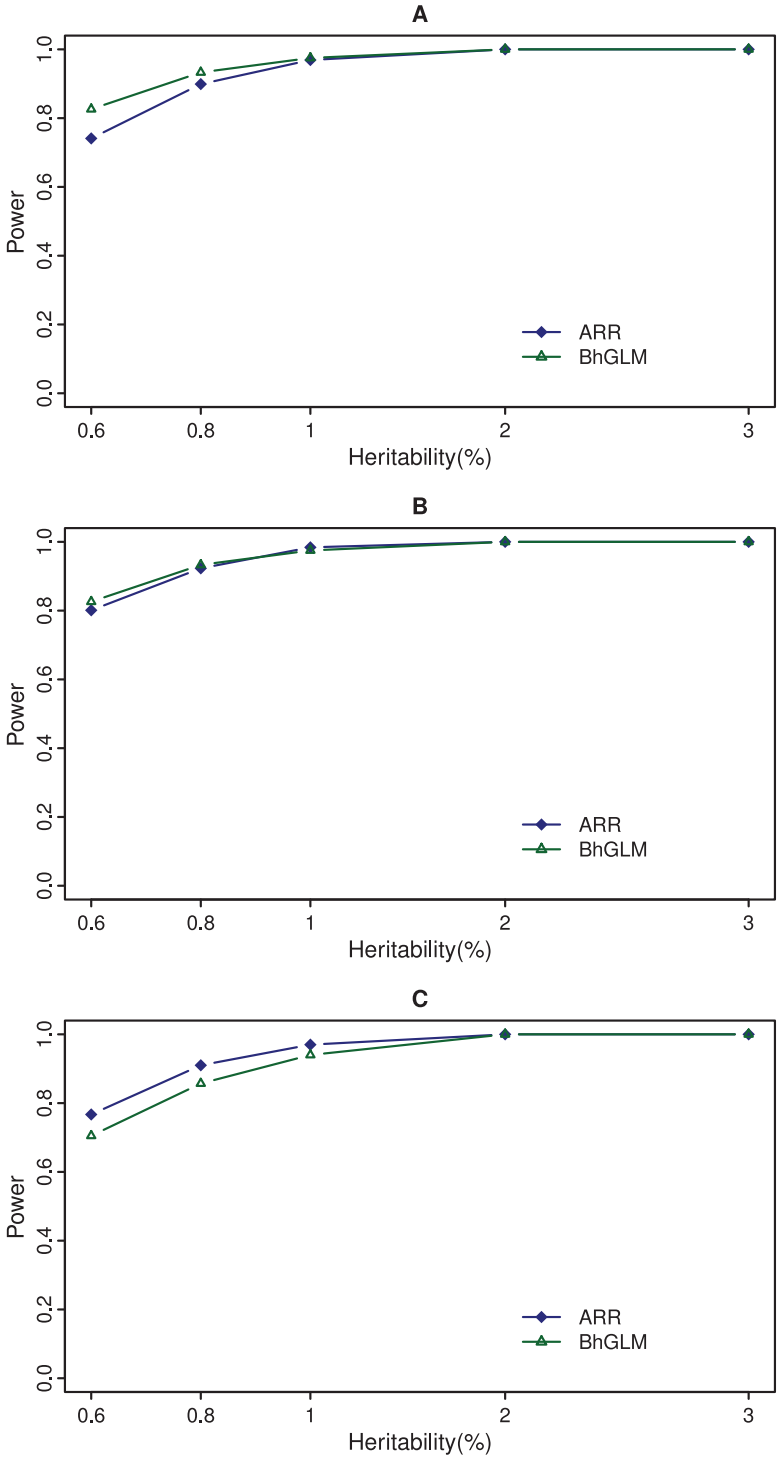

Figure 3. Power comparison between ARR and BhGLM at significance level of 0.05. The top panel (A) gives the powers of the adaptive ridge regression (ARR) and the Bayesian hierarchical generalized linear model (BhGLM) evaluated at the threshold of 3.84. The panel in the middle (B) shows the powers of ARR and BhGLM evaluated at the threshold 2.71 for ARR and 3.84 for BhGLM. The bottom panel (C) shows the powers of ARR and BhGLM using thresholds of 3.45 and 9.78, respectively, to control the 0.05 Type I error rate.
doi:10.1371/journal.pone.0044173.g003

One reviewer brought a recent publication to our attention [41]. The method is called sequence kernel association test (SKAT). After reading this paper, we agreed that our approach is similar to SKAT. However, SKAT only gives the score test and no parameter estimation is provided. This explains why SKAT is fast computationally. There are three major advantages of the adaptive ridge regression. First, a high score test does not mean the effects are large. It may be caused by small effects but large sample size. The score test cannot tell the difference. Our method not only provides a test but also an estimate of the group variance.

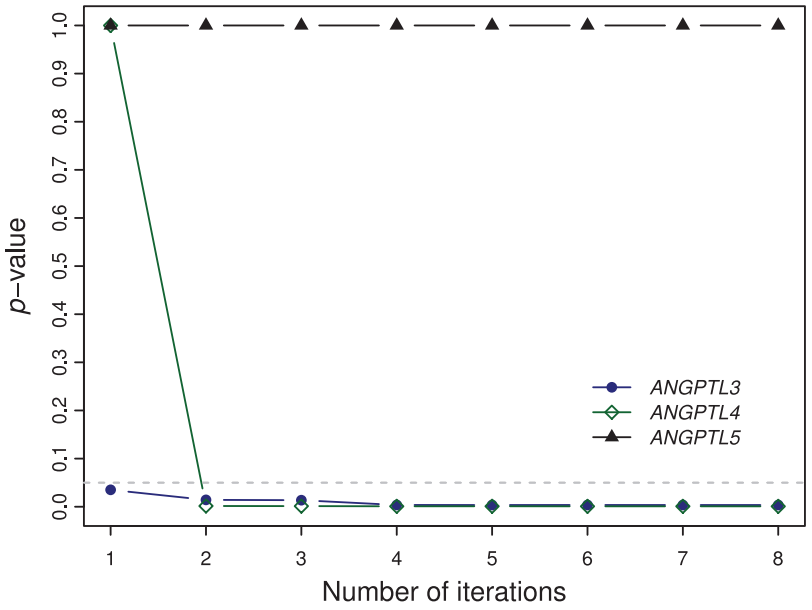

Figure 4. Changes of *p*-values of the three genes during the iteration process in the separate analysis. The *p*-values of the three genes in separate analysis using adaptive ridge regression (ARR) are plotted against the iteration process. P-values are calculated based on $0.5\chi_0^2+0.5\chi_1^2$ distribution.
doi:10.1371/journal.pone.0044173.g004

We can provide a total proportion of the phenotypic variance contributed by the rare variants. Secondly, we introduced an adaptive step to the original ridge regression. This step plays the role of "weighting" of the SKAT method but it can "homogenize" the effect of each rare variant within a group. The ridge regression performs better under the "homogenized" rare variant effect assumption. Thirdly, our method works for both rare and common variants. However, the SKAT method was particularly designed for rare variants because the "weights" for the common variants will be almost zero (excluded from the model), according to the authors of that paper. There is a possibility to use the score test under our adaptive ridge regression framework. The estimation procedure will remain the same, but we may simply replace the likelihood ratio test by the score test. The "weights" obtained from the adaptive ridge regression will be used in the score test. This needs to be further investigated.

We did not compare the ARR method with other rare variant detection methods other than the BhGLM method. The reason for this is that Yi and Zhi [20] already compared BhGLM with many other methods and showed that BhGLM outcompeted all of them. Given the fact that our method is more powerful than BhGLM (simulation study), we concluded that the ARR method is also more powerful than the other methods. The BhGLM program provides a set of default priors, which were used in this study. Users do have the option to choose their own priors. If different priors were chosen, the power of the BhGLM may change slightly (in either direction). The default priors provided by Yi and Zhi (2011) were drawn from extensive simulation studies and should be quite robust. It is difficulty to choose the optimal set of priors in simulation studies. However, it is easy to choose the optimal prior set in real data analysis. We need a criterion to evaluate the priors. Statistical power is not a viable criterion in real data analysis because the true rare variant effects are not known. The mean squared error (MSE) via cross validation may be a viable choice for the criterion. This requires further investigation. Our ARR method is a maximum likelihood approach, equivalent to uniform priors for all variances. In theory, we can also assign the variances to other priors to improve the power. This deserves further investigation.

The adaptive ridge regression method has been shown to be the Lasso [33] estimation if the Lassos parameter $\lambda=\sigma^2/\phi^2$ is predetermined by the investigator. Our new contribution is to estimate $\lambda=\sigma^2/\phi^2$ using estimated variance components. This approach has provided a new way to select the shrinkage factor λ based on data. In the original Lasso method, the author used cross validation to determine the shrinkage factor. With the new method, the Lasso parameter is estimated from the data and thus has eliminated the cross validation step. The extension of the ARR to multiple groups of rare variant detection is conceptually similar to the group Lasso method [42,43] in which different groups have different Lasso parameters, as given by $\lambda_l=\sigma^2/\phi_l^2$. This idea has a general application to detection of multiple groups of variants as well as their interactions (epistatic effects). The current methods of rare variant detection have not been able to detect interactions between two groups of rare variants. The Dallas Heart Study dataset contains three genes (groups). Our next project will be analyzing the three pairs of group interactions among the three genes. Gene *ANGPTL5* has no effect on triglyceride level. However, it may interact with other two genes. The full model will include three group variances plus three variance components of the interaction.

The Lasso method itself may not be perfect for all data. It may work for some data but not work for other data. Using the adaptive ridge regression approach, we may modify the shrinkage factor through different choice of the constraint. For example, the constraint of c_k given by Grandvalet [28] is $\sum_{k=1}^{m} c_k^2=m$. This constraint determines the level of shrinkage. An obvious extension may be $\sum_{k=1}^{m} c_k^2=\rho m$, where $0<\rho<\infty$ is another factor we can use to control the strength of the shrinkage. Our adaptive ridge regression is equivalent to $\rho=1$, a special case of the general method.

The new method is developed for continuous traits under the linear mixed model framework [30]. In many situations, the trait

of interest may be a binary trait. The generalized linear mixed model (GLMM), which is an extension of the linear mixed model, can be used to analyze the association of multiple rare variants and a binary trait. This extension is very straight forward because the methodology of GLMM has been well established. The simple extension includes the adaptive steps.

Finally, we performed all the analyses using an R program. The R package is called Adaptive Ridge Regression (ARR) which can be downloaded from the authors' personal website: www.statgen.ucr.edu.

Author Contributions

Conceived and designed the experiments: SX. Performed the experiments: HZ. Analyzed the data: HZ. Contributed reagents/materials/analysis tools: HZ. Wrote the paper: HZ SX.

References

1. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41: 47–55.
2. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40: 955–962.
3. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40: 189–197.
4. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322: 881–888.
5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
6. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39: 513–516.
7. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, et al. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119: 70–79.
8. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
9. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695–701.
10. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11: 415–425.
11. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869–872.
12. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, et al. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A 103: 1810–1815.
13. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82: 100–112.
14. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124–137.
15. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387–389.
16. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773–785.
17. Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44: 293–308.
18. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311–321.
19. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.
20. Yi N, Zhi D (2010) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35: 57–69.
21. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615: 28–56.
22. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188–193.
23. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.
24. Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70: 42–54.
25. Pan W, Shen X (2011) Adaptive tests for association analysis of rare variants. Genet Epidemiol 35: 381–388.
26. Yi N, Liu N, Zhi D, Li J (2011) Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLoS Genet 7: e1002382.
27. Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12: 55–67.
28. Grandvalet Y (1998) Least absolute shrinkage is equivalent to quadratic penalization. International Conference on Artificial Neural Networks (ICANN'98). Skövde, Sweden.
29. Xu S (1998) Further investigation on the regression method of mapping quantitative trait loci. Heredity 80: 364–373.
30. Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447.
31. Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society Series B 66: 165–185.
32. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.
33. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B 58: 267–288.
34. Harville DA (1977) Maximum-likelihood approaches to variance component estimation and to related problems. J Amer Stat Assoc 72: 320–340.
35. Hartley HO, Rao JNK (1967) Maximum likelihood estimation for the mixed analysis of variance model. Biometrika 54: 93–108.
36. Koishi R, Ando Y, Ono M, Shimamura M, Yasumo H, et al. (2002) Angptl3 regulates lipid metabolism in mice. Nat Genet 30: 151–157.
37. Koster A, Chao YB, Mosior M, Ford A, Gonzalez-DeWhitt PA, et al. (2005) Transgenic angiopoietin-like (angptl)4 overexpression and targeted disruption of angptl4 and angptl3: regulation of triglyceride metabolism. Endocrinology 146: 4943–4950.
38. Li C (2006) Genetics and regulation of angiopoietin-like proteins 3 and 4. Curr Opin Lipidol 17: 152–156.
39. Ono M, Shimizugawa T, Shimamura M, Yoshida K, Noji-Sakikawa C, et al. (2003) Protein region important for regulation of lipid metabolism in angiopoietin-like 3 (ANGPTL3): ANGPTL3 is cleaved and activated in vivo. J Biol Chem 278: 41804–41809.
40. Yoshida K, Shimizugawa T, Ono M, Furukawa H (2002) Angiopoietin-like protein 4 is a potent hyperlipidemia-inducing factor in mice and inhibitor of lipoprotein lipase. J Lipid Res 43: 1770–1772.
41. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89: 82–93.
42. Antoniadis A, Fan J (2001) Regularization of wavelet approximations (with discussion). Journal of the American Statistical Association 96: 939–967.
43. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Statist Soc B 68: 49–67.