BMC Bioinformatics

RESEARCH ARTICLE Open Access

Significance testing in ridge regression for genetic data

Erika Cule*, Paolo Vineis and Maria De Iorio

Abstract

Background: Technological developments have increased the feasibility of large scale genetic association studies. Densely typed genetic markers are obtained using SNP arrays, next-generation sequencing technologies and imputation. However, SNPs typed using these methods can be highly correlated due to linkage disequilibrium among them, and standard multiple regression techniques fail with these data sets due to their high dimensionality and correlation structure. There has been increasing interest in using penalised regression in the analysis of high dimensional data. Ridge regression is one such penalised regression technique which does not perform variable selection, instead estimating a regression coefficient for each predictor variable. It is therefore desirable to obtain an estimate of the significance of each ridge regression coefficient.

Results: We develop and evaluate a test of significance for ridge regression coefficients. Using simulation studies, we demonstrate that the performance of the test is comparable to that of a permutation test, with the advantage of a much-reduced computational cost. We introduce the *p*-value trace, a plot of the negative logarithm of the *p*-values of ridge regression coefficients with increasing shrinkage parameter, which enables the visualisation of the change in *p*-value of the regression coefficients with increasing penalisation. We apply the proposed method to a lung cancer case-control data set from EPIC, the European Prospective Investigation into Cancer and Nutrition.

Conclusions: The proposed test is a useful alternative to a permutation test for the estimation of the significance of ridge regression coefficients, at a much-reduced computational cost. The *p*-value trace is an informative graphical tool for evaluating the results of a test of significance of ridge regression coefficients as the shrinkage parameter increases, and the proposed test makes its production computationally feasible.

Background

Genetic data collected in case-control or cohort studies of well-defined disease phenotypes can be used to identify genetic variants, typically single nucleotide polymorphisms (SNPs), associated with disease status. In a genetic association study, the data consist of genotypes and corresponding phenotypes from large numbers of individuals with the disease phenotype of interest (cases) and disease-free controls. A significant difference in the frequency of a genetic variant between the case and control groups is taken to be indicative of an association between that variant (or, more probably, a nearby causal variant correlated with the typed variant) and the phenotype of interest. A framework also exists for analysing a continuous phenotypes. Once one or more associated variant have been identified, they can be included in predictive models for the estimation of disease risk in individuals for whom the (potentially future) disease status is unknown. Technological developments, including commercially available chips for typing millions of genetic variants simultaneously, and next-generation sequencing technologies such as those used in the 1000 Genomes Project [1], are enabling the collection of large amounts of genetic data from large numbers of individuals which means that analysis of contemporary genetic data sets involves the study of high-dimensional data.

A number of statistical approaches have successfully been used to investigate the strength of association between genetic variants and a phenotype of interest. These methods include testing the significance of the association of each variant with the phenotype independently using standard univariate tests (such as the Cochran-Armitage test for trend [2] or Fisher's exact

* Correspondence: erika.cule05@imperial.ac.uk
Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

test). However, there are disadvantages to relying on univariate methods for the analysis of genetic data. The data from a genome-wide association study typically consists of thousands or millions of SNPs, and this large number of predictors introduces a multiple testing problem. Univariate methods therefore require stringent significance thresholds due to the large number of tests being undertaken to prevent a high false-positive rate [3]. Further, univariate methods fail to take into account the combined effect of multiple SNPs. It is plausible that many genetic variants together contribute to the phenotype being studied [4], and one advantage of using multivariate methods is that they allow for the study of the combined effect of multiple SNPs. Multivariate methods allow for the control of confounding variables, as in the lung cancer replication study in this paper, where gender, smoking status and age were included as unpenalised covariates in the model.

Using multivariate methods, the regression model can be extended to include, for example, interaction or higher order terms, and in such a case a penalised regression approach would be appropriate [5]. Penalised regression methods have been applied to genetic data [6,7]. Among a number of regression approaches used for prediction in high-dimensional data, ridge regression has been shown to perform best in terms of prediction error [8]. Ridge regression has successfully been used to analyse genetic data where SNPs were in high LD [9]; it is the test of significance used by Malo, Libiger & Schork that we evaluate here.

To begin, we consider two regression models commonly used in the analysis of genetic data - the linear and the logistic regression models, as follows.

The standard linear regression model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{Y}$ is a $(n \times 1)$ vector of dependent variables, Y_i, $i = (1, ..., n)$ and $\mathbf{X}$ is a $(n \times m)$ matrix of predictors. $\boldsymbol{\beta}$ is a $(m \times 1)$ vector of regression coefficients β_j, $j = (1, ..., m)$ and $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of normally distributed random errors, with $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. An example would be a model of the relationship between a continuous phenotype (such as blood pressure or plasma lipid concentration) measured in n individuals, and the genotype of these n individuals at m SNPs.

The ordinary least squares estimator for $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2}$$

The significance of individual OLS regression coefficients $\hat{\beta}_j$ in a multiple regression model can be estimated using a Wald test. The test statistic is

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

where $se(\hat{\beta}_j)$ is an estimate of the standard error of the j^{th} regression coefficient. Under the null hypothesis $H_0 : \hat{\beta}_j = 0$, T_0 follows a Student t distribution with n - m degrees of freedom.

Binary outcomes commonly arise in biomedical data where they may represent, for example, cases and controls. In the logistic regression model, $\mathbf{Y}$ is an n-dimensional vector of response variables taking values 0 (controls) or 1 (cases), and $\mathbf{X}$ the $n \times m$ matrix of explanatory variables, as before. For the i^{th} individual we denote

$$\mathbf{x}_i = (X_{i1}, \ldots, X_{im})$$

The i^{th} response Y_i is a Bernoulli variable with probability of success equal to p_i. The logistic regression model relates the probability p_i that the i^{th} observation is a case to the predictor variables as

$$P(Y_i = 1|\mathbf{x}_i) = p_i = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}} \tag{3}$$

where $\boldsymbol{\beta}$ is a vector of parameters to be estimated.

The significance of individual logistic regression coefficients, $\hat{\beta}_j$, can be estimated using the test statistic

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

where $se(\hat{\beta}_j)$ is an estimate of the standard error of the j^{th} regression coefficient. Under the null hypothesis that $\beta_j = 0$, T_0 asymptotically follows a standard normal distribution.

Genetic data often comprises more predictor variables, m, than observations, n. In such a situation, unique maximum likelihood estimates of regression parameters do not exist. Further, collinearity in the predictors, due to linkage disequilibrium (LD) in genetic data, which typically increases with the increasing density of available markers, results in unstable maximum likelihood estimates of regression coefficients.

An extensive literature exists on the application of modified regression techniques to the analysis of high-dimensional data. Penalised regression constrains the magnitude of the estimated regression coefficients, allowing their estimation when ordinary least squares (OLS) estimates cannot be obtained. In a Bayesian context, these techniques are equivalent to the specification of a particular prior distribution on the coefficients. For example, Lasso regression [10] constrains the sum of the absolute value of the regression coefficients to be less than a constant. This is equivalent to imposing a double exponential prior centred at zero on the coefficients. Lasso regression can estimate some coefficients to be exactly zero, permitting dimension reduction in the model. Hoggart, Whittaker, De Iorio & Balding [6] considered a modified Lasso regression approach for the identification of causal SNPs in genome-wide or resequenced data, with the aim of identifying regions of association whilst considering all SNPs simultaneously. When SNPs are in high LD, their method offers improvement over both single-SNP analysis and Lasso regression in terms of the power to detect causal variants, and a notable improvement over single-SNP analysis in terms of false-positive rate.

Ridge regression [11] is a another penalised regression approach, in which a penalty is applied to the sum of the squared parameter estimates. Ridge regression has been used in a number of large-scale data analysis scenarios, including marker-assisted selection [12], expression data analysis [13], and genetic association studies when SNPs are in high LD [14]. From a Bayesian viewpoint, ridge regression can be considered as standard multiple regression with the coefficients estimates having a prior distribution that is normal with mean zero and known variance [15]. In genetic epidemiology, it is desirable to estimate the strength of the association between a variant and a phenotype. This is problematic when using ridge regression which, unlike other penalised regression approaches, does not reduce the number of parameters in the model, nor estimate the significance of each fitted coefficient.

A test of significance for coefficients estimated using ridge regression, based on an approximation of their distribution under the null hypothesis, was proposed by Halawa & El Bassiouni [16]. The test was originally developed and evaluated for data with continuous outcomes, when different methods were used to compute the shrinkage parameter, λ, which controls the degree of shrinkage of the regression coefficients and hence their distribution under the null hypothesis. Malo, Libiger & Schork [9] used the same test in an evaluation of the applicability of ridge regression as a means of accommodating LD in association studies. They used the test in a comparison of the performance of ridge regression, multiple regression and single-SNP analysis when SNPs are in varying degrees of LD. They found that ridge regression identified different SNPs as associated with phenotype compared to single-SNP analysis or multiple regression. However, they did not consider the performance of the test itself.

In this paper, we evaluate the performance of a test of significance for ridge regression coefficients. Our test is based on the test proposed by Halawa & El Bassiouni [16]. We extend the test, making it applicable in the $m > n$ scenario that is common in contemporary genetic data sets. We evaluate the performance of the test in simulation studies, using scenarios representative of realistic high-desnity genetic marker data, considering a range of data set dimensions and degrees of shrinkage.

Ridge regression has also been applied in the logistic regression framework [17]. We extend the test proposed by Halawa & El Bassiouni [16] to the logistic ridge regression model, and again evaluate the test in a range of simulation scenarios at different values of λ.

In both linear and logistic ridge regression, we compare the approximate test of significance to a permutation test. We view the permutation test as a benchmark as it produces an estimate of the null distribution of the parameter estimates. However, the permutation test is computationally intensive and becomes more so when data are high-dimensional. The test we propose makes it feasible to estimate significance with a much lower computational burden.

We introduce the *p*-value trace, a plot of the negative logarithm of the *p*-values of the ridge regression coefficients with increasing shrinkage parameter. This plot enables the visualisation of the relative change in significance of each coefficient, and facilitates the identification of predictors most affected by increased penalisation in terms of significance.

We apply the approximate test of significance for logistic ridge regression coefficients to a lung cancer data set, demonstrating the utility of the test when correlation exists among the predictors.

This paper is organised as follows. We first describe the approximate test of significance and the permutation test to which it is compared. We then describe the simulation studies used in this paper. In the Results section we evaluate the performance of the proposed test in a range of simulation scenarios. Further, we apply the test to a lung cancer case control data set. In the Discussion we discuss the results and potential applications for the test.

Methods

Significance testing in linear ridge regression

Ridge regression replaces the OLS estimator $\hat{\boldsymbol{\beta}}$ (equation (2)) with the ridge regression estimator [11]:

$$\hat{\boldsymbol{\beta}}^{\lambda} = (\mathbf{X'X} + \lambda\mathbf{I})^{-1}\mathbf{X'Y} \tag{4}$$

Where collinearity exists in $\mathbf{X}$, the OLS estimates of $\boldsymbol{\beta}$ can be unstable, having large variance. Hoerl & Kennard [11] demonstrate that there exists a value of λ for which the ridge regression estimates $\hat{\boldsymbol{\beta}}^{\lambda}$ have smaller mean square error (MSE) than the OLS estimates. Where $m > n$, OLS estimates of $\boldsymbol{\beta}$ cannot be obtained because the matrix $\mathbf{X'X}$ is singular. The addition of the constant λ

to the diagonal of the **X′X** matrix makes it invertible, so ridge regression estimates can be obtained.

Halawa & El Bassiouni [16] use simulation studies to investigate applications of a 'non-exact' *t*-type test for the individual coefficients of a linear regression model fitted using ridge regression, based on the *t*-test above. The test statistic is

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

where $\hat{\beta}_j^\lambda$ is the estimate of the j^{th} regression coefficient under the ridge regression model, and $se(\hat{\beta}_j^\lambda)$ is an estimate of the standard error.

Estimates of the standard error of the j^{th} element of $\hat{\boldsymbol{\beta}}^\lambda$ are obtained as the square root of the j^{th} element of the diagonal of the covariance matrix

$$\text{Var}\ (\hat{\boldsymbol{\beta}}^\lambda) = \sigma^2(\mathbf{X'X} + \lambda\mathbf{I})^{-1}\mathbf{X'X}(\mathbf{X'X} + \lambda\mathbf{I})^{-1}$$

In practise, σ^2 is replaced by its estimate, given by the residual mean square of the ridge model:

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\nu} \tag{5}$$

ν is the residual effective degrees of freedom. Halawa & El Bassiouni [16] use $\nu = n - m$. However, when $m > n$ this gives a negative estimate of the residual mean square. Instead, we use the definition of residual effective degrees of freedom given in Hastie & Tibshirani [18], which makes use of the "hat matrix", **H**:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X'X} + \lambda\mathbf{I})^{-1}\mathbf{X'Y} \tag{6}$$

$$= \mathbf{HY} \tag{7}$$

H is termed the 'hat matrix', because it 'puts the hat on' **Y**, transforming it to $\hat{\mathbf{Y}}$. Degrees of freedom for error are defined as

$$\nu = n - \text{tr}(2\mathbf{H} - \mathbf{HH'}) \tag{8}$$

In linear regression, the hat matrix reduces to $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ and n - tr (2**H** - **HH′**) reduces to $n - m$. The test statistic T_λ is assumed to follow a Student *t* distribution as in standard multiple linear regression. However, the effective number of parameters of the penalised regression fit is smaller than m. Hastie and Tibshirani define tr (**H**) as the degrees of freedom taken up by the penalised regression fit [18].tr (**H**) reduces to $n - m$ in ordinary linear regression. Then, T_λ is assumed to follow a Student *t* distribution with $n - tr$ (**H**) degrees of freedom.

In the case of large sample size, as is typically the case in genetic data, the distribution of the test statistic is asymptotically normal, as noted by Halawa & El Bassiouni [16]. We compared the significance levels of the approximate test assuming both a normal and a Student *t*-distribution of the test statistic and found that the results were substantially identical. Therefore we assume that under H_0, $T_\lambda \sim \mathcal{N}(0, 1)$ and use the normal distribution to test the significance of ridge regression coefficients. The results from the corresponding tests assuming that under H_0, $T_\lambda \sim t_{n\ -\ \text{tr}(\mathbf{H})}$ are provided in an Additional File. See Additional File 1, Tables S1 and S2.

Significance testing in ridge logistic regression

Ridge regression has been applied to the logistic regression model [17,19]. Cessie & van Houwelingen [17] show how ridge regression can be used to improve the parameter estimates in logistic regression when the number of predictors is relatively large or highly correlated. They discuss different ways of choosing the shrinkage parameter to minimize prediction error. Vago & Kennedy [19] apply ridge logistic regression to a clinical data set.

In logistic ridge regression, the log-likelihood function is penalised with the penalty applied to the L_2 norm of $\boldsymbol{\beta}$ [19]. Maximum likelihood estimates of $\boldsymbol{\beta}$ are obtained by maximising the logarithm of the likelihood function [19], typically using the Newton-Raphson algorithm. The approximate test statistic is

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

Standard errors of the coefficient estimates are obtained as the square roots of the j^{th} element of the diagonal of the covariance matrix. This matrix is estimated from the final Newton-Raphson iteration:

$$\begin{aligned}
\text{Var}\left(\hat{\boldsymbol{\beta}}^\lambda\right) &= \text{Var}\left[\left(\mathbf{X'WX} + 2\lambda\mathbf{I}\right)^{-1}\mathbf{X'Wz}\right] \\
&= \left(\frac{\partial^2\ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)^{-1} I(\boldsymbol{\beta}) \left(\frac{\partial^2\ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)^{-1} \\
&= \left(\mathbf{X'WX} + 2\lambda\mathbf{I}\right)^{-1}\left(\mathbf{X'WX}\right)\left(\mathbf{X'WX} + 2\lambda\mathbf{I}\right)^{-1}
\end{aligned}$$

where $I(\boldsymbol{\beta})$ is the observed information matrix, ℓ is the (penalised) log-likelihood, and **W** is the weight matrix:

$$\mathbf{W} = \text{diag}\left[\hat{p}_i(1 - \hat{p}_i)\right]$$

$$\hat{p}_i = \frac{e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}^\lambda}}{1 + e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}^\lambda}}$$

$\mathbf{z}$ is an $n \times 1$ column vector with elements

$$z_i = \text{logit}\left[\hat{p}_i\right] + \frac{Y_i - \hat{p}_i}{\hat{p}\left(1 - \hat{p}_i\right)}$$

Again we assume that under H_0, $T_\lambda \sim \mathcal{N}(0, 1)$ and use the normal distribution to test the significance of ridge regression coefficients.

Permutation test

To evaluate the proposed test, we compare its performance to that of a permutation test, which we view as a benchmark. In the permutation test, to obtain a null distribution of the regression coefficients, the elements of the outcome vector $\mathbf{Y}$ are randomly permuted. The ridge regression model is fitted using the permuted observations, to obtain ridge regression coefficients. By performing 1000 such permutations, a null distribution of the regression coefficients is generated. The permutation test p-value is calculated as the proportion of regression coefficients from the null distribution greater than or equal in absolute value to the absolute value of the coefficient fitted to the true (non-permuted) data.

Choice of shrinkage parameter

Hoerl & Kennard [11] present an existence theorem for ridge regression. They demonstrate the existence of a value of the shrinkage parameter λ in equation (4) which will give estimates $\hat{\boldsymbol{\beta}}^{\lambda}$ with smaller mean squared error than the OLS estimates $\hat{\boldsymbol{\beta}}$ given in (2).

However, to date no analytical method to find the 'best' value of λ in terms of minimising MSE has been determined. A number of data-driven methods have been proposed. These methods aim to determine a value of λ based on the data that will result in estimates of $\hat{\boldsymbol{\beta}}^{\lambda}$ with improved mean squared error properties. For example, Hoerl, Kennard & Baldwin [20] propose the following as an estimate of λ, with $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$ taken from the OLS estimates:

$$\lambda_{\text{HKB}} = \frac{m\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}$$

An alternative estimate was suggested by Lawless & Wang [21], based on the adoption of the Bayesian perspective mentioned above. Cross-validation based methods have also been proposed in the literature [22].

Hoerl & Kennard [11] introduce the ridge trace, a plot of the estimates $\hat{\boldsymbol{\beta}}^{\lambda}$ as λ increases from zero - see for example Figure 1. They propose choosing λ corresponding to the region on the ridge trace at which estimates of $\hat{\boldsymbol{\beta}}^{\lambda}$ no longer change significantly as λ increases further.

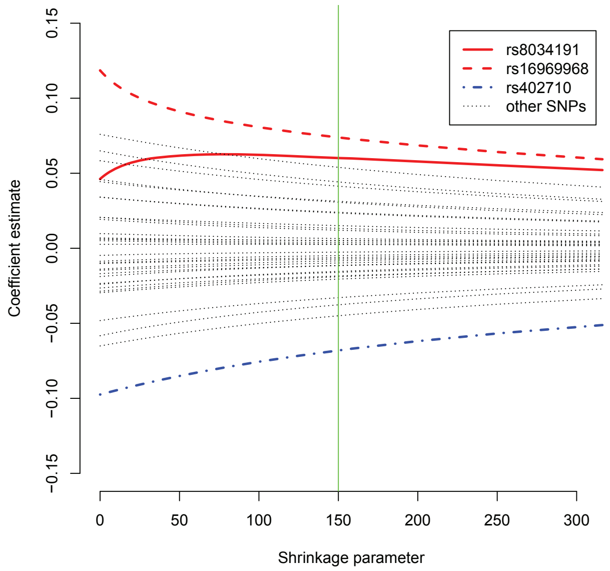

Figure 1 Ridge trace - Lung cancer data. Change in ridge regression coefficients with increasing shrinkage parameter. The three SNPs that have previously been shown to be associated with lung cancer risk, rs8034191, rs16969968 and rs402710, are shown in bold. Other SNPs are represented by dotted lines. The vertical line is at λ = 150, where the ridge estimates stabilise.

Following the ridge trace of Hoerl & Kennard [11], we introduce a plot of p-values of the regression coefficients against λ as λ increases from zero (Figure 2). We refer to this plot as a 'p-value trace'. This p-value trace enables the visualisation of the change in p-values of the regression coefficients with increasing shrinkage.

Simulation study

The proposed test was evaluated using simulated genotype data. FREGENE software [23,24] was used to simulate a population of haplotypes. FREGENE simulates the forwards-in-time evolution of sequence-like genetic data. The forward-in-time simulation allows demographic and selection scenarios to be implemented and recombination to be modelled. Haplotypes used in this study are taken from a simulation representing the neutral evolution of 10,500 individuals over 200,000 generations, with recombination parameters that realistically model recombination in humans. This results in a population of 21,000 haplotypes. The simulated haplotypes, together with details of the simulation, are available to download from http://www.ebi.ac.uk/projects/BARGEN/ [25].

Based on these haplotypes, genotypes and corresponding phenotypes were generated as follows. A range of scenarios were considered, comprising n individuals at m SNPs.

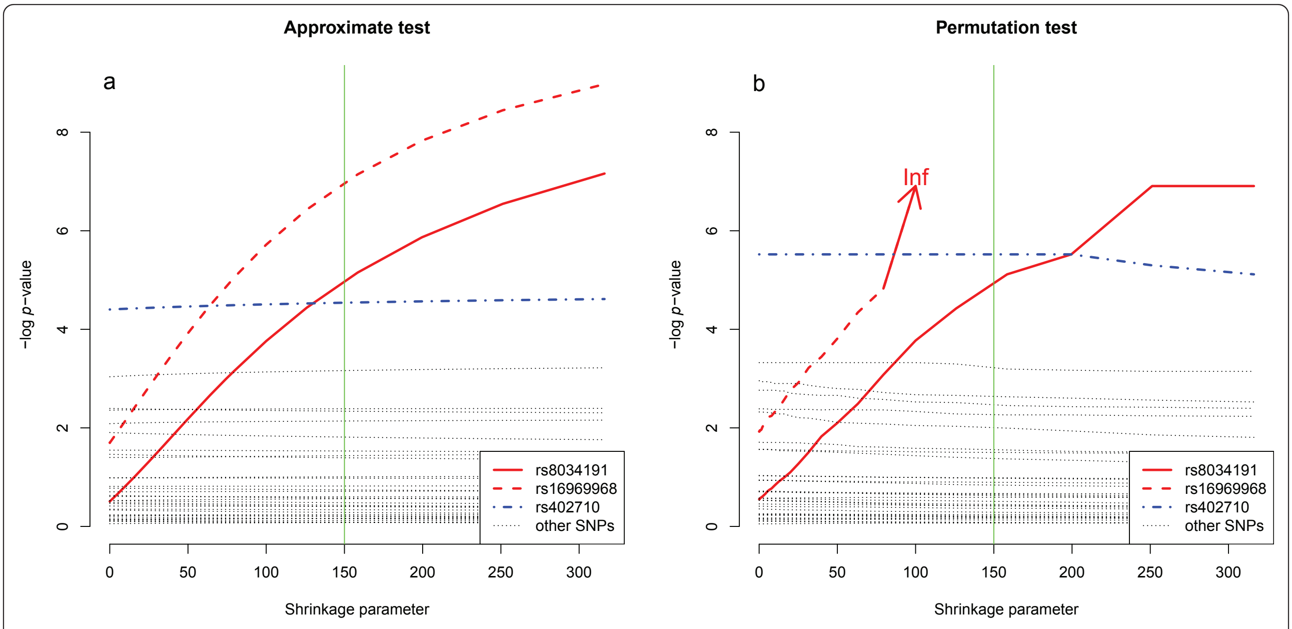

Figure 2 ***p*-value traces - Lung cancer data**. Change in *p*-values of ridge regression coefficients in lung cancer data with increasing shrinkage parameter. *p*-values are plotted on the - log scale. The three SNPs that have previously been shown to be associated with lung cancer risk, rs8034191, rs16969968 and rs402710, are shown in bold. Other SNPs are represented by dotted lines. The vertical line is at λ = 150, where the ridge estimates stabilise. **a** approximate *p*-values, **b** permutation test *p*-values.

1. Designate one SNP as the causal SNP, selected at random from those with population minor allele frequency in the range 0.10 - 0.15. A subregion of m SNPs from the 1 Mb region, containing this causal SNP, is used for the genotype simulation. The subregion is selected at random to be of length m and to contain the designated causal SNP. Thus the correlation structure among the subregion depends on the correlation structure of the haplotype region at that point. In the resultant sample, correlation with the causal SNP ranges from low to perfect (r^2 = 1).
2. Sample two haplotypes (with replacement) from the population of 21,000 haplotypes. Sum the minor allele count at each SNP to form a genotype.
3. Simulate the phenotype for this individual.
Continuous phenotypes were generated as $Y_i \sim \mathcal{N}(\mu, \sigma^2)$.
Case-control phenotypes were generated following the liability model used by [26]. The penetrance function, f_k, is the probability of being a case, Pr (Y_i = 1) given having k copies of the minor allele at the causal SNP. The genotype relative risk, r, is f_1 /f_0 and K is the population prevalence. Then, with the population frequency of the minor allele of the causal SNP as p, under an additive genetic model, f_0 = K/(1 - 2p + 2pr), f_1 = rf_0 and f_2 = 2r f_0 - f_0. A sample of n/2 cases and n/2 controls was generated by generating an individual genotype as described above, then assigning the individual to be a case with probability f_k and a control otherwise. This process is repeated until n/2 cases and n/2 controls are obtained.
4. Record the minor allele count (0, 1, 2) at the m SNPs for the i^{th} individual, giving rise to an $n \times m$ matrix of minor allele counts.

Ridge regression coefficients were fitted to data with continuous outcomes using `lm.ridge` from the package `MASS` in R [27] for both the simulated data and the permutation test.

Estimates of regression coefficients under logistic ridge regression models were computed using the Newton-Raphson algorithm.

In the case of both continuous and binary outcomes, SNPs that were invariant in the sampled genotypes were removed from the data, and genotypes were standardised, prior to analysis.

The two tests were evaluated using the true positive rate (TPR) and false positive rate (FPR), averaged over all the replicates for each simulation scenario. We define TPR to be the proportion of causal SNPs, as designated in the data simulation, significantly associated with phenotype at the nominal threshold α = 0.05. TPR is not reported for the null simulations, as there is no causal SNP associated with phenotype in these data. We define FPR to be the proportion of non-causal SNPs significantly associated with phenotype at the same significance threshold.

Results and Discussion

Null Simulation

Genotypes and corresponding phenotypes were generated in two different sized data sets: (1) n = 500, m = 20 and (2) n = 1000, m = 1000. In generating the null data, no SNP was designated the causal SNP. Continuous phenotypes were generated as $Y_i \sim \mathcal{N}(0, 1)$; binary phenotypes were generated as Y_i ~ Binom(1, 0.5). False positive rates are reported at the nominal significance threshold α = 0.05. Four values of the ridge parameter λ were used.

Results of the null simulations are shown in Table 1. We find that approximate test gives similar results to the permutation test in terms of false positive rate, especially in the case of continuous outcomes.

Continuous Phenotypes

Genotypes and corresponding phenotypes were generated as described above, for a range of data sets dimensions: n = 500, 1000, 5000 and m = 20, 100, 1000 and all SNPs in the 1 Mb region (approximately 10,000 SNPs). Phenotypes were generated as $Y_i \sim \mathcal{N}(1 + 2k, 1)$ with k being the minor allele count at the causal SNP. A range of values of the shrinkage parameter λ were used: λ = 0.1, 1, 10, 100. In Figure 3, the left column shows the null distributions, generated in a permutation test, used to estimate the significance of a ridge regression coefficient for a significant SNP (top row) and for a SNP that is not associated with phenotype (bottom row). The coefficient fitted to the original data is indicated. In the right column, the null distribution of the test statistic used in the approximate test is shown, with the test statistic of the fitted coefficient indicated. Ridge regression models were fitted using the shrinkage parameter λ = 1. These results are examples taken from a single simulation, and above each plot the p-value according to the permutation test (left) or the approximate test (right) is shown.

Table 2 compares the performance of the approximate and permutation tests in different simulation scenarios and at different values of the shrinkage parameter. We see that the approximate test performs well compared to the permutation test in terms of power (true positive rate) and that it has a slightly higher false positive rate.

When ranking the SNPs in order of significance, the approximate test and the permutation test ranked the SNPs identically or nearly so (results not shown).

Figure 4 shows a Bland-Altman plot [28] of difference (permutation test p-value - z-type test p-value) against mean for the p-values of 1000 SNPs in 5000 individuals. p-values are plotted on the -log scale. We see that the bias is towards smaller p-values from the approximate test, which is congruous with the higher false positive rate for the approximate test shown in Table 2.

Continuous Phenotypes with Multiple Causal SNPs

For complex diseases, multiple causal SNPs are likely to affect the phenotype. We investigated the performance of the test when more than one SNP is associated with phenotype. We simulated data from two different scenarios: n = 500, m = 100 and n = 500, m = 1000. In each region of m simulated genotypes, ten SNPs with minor allele frequency 0.10 - 0.15 were designated causal and given effect size 1; the non-causal SNPs had effect size 0. Phenotypes were simulated as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, σ^2 = 1 where $\boldsymbol{\beta}$ is the vector of effect sizes.

Results are presented in Table 3. We see that the conclusions drawn about the test, of adequate power at the cost of a slightly higher false positive rate, are equally valid when multiple SNPs in the data are associated with phenotype.

Computational performance comparison

Using an example simulation, we compared the computational time required to compute the approximate and the permutation tests. A data set with dimensions n = 1000, m = 1000 and λ = 1 was used. Approximate test and permutation test p - values were computed and the time taken to arrive at the p-values was recorded. Calculations were done using R version 2.12.0 [27] on an iMac running Mac OS X Version 10.6.7, fitted with an 2.8 Ghz Intel Core i7 processor and 16 GB 1067 MHz DDr3 RAM. Computational times are compared in Table 4. We see that the

Table 1 Performance comparison in null simulation

			Shrinkage Parameter							
			Approximate test				Permutation test			
			0.1	1	10	100	0.1	1	10	100
Individuals	**SNPs**	**Outcome**								
500	20	Continuous	0.066	0.066	0.066	0.075	0.066	0.066	0.066	0.075
		Binary	0.021	0.021	0.041	0.067	0.027	0.027	0.033	0.052
1000	1000	Continuous	0.050	0.049	0.050	0.046	0.051	0.052	0.050	0.046
		Binary	0.118	0.092	0.066	0.053	0.054	0.056	0.053	0.051

False positive rates at the nominal significance threshold α = 0.05 in null datasets. In each scenario, results are averaged over ten replicates

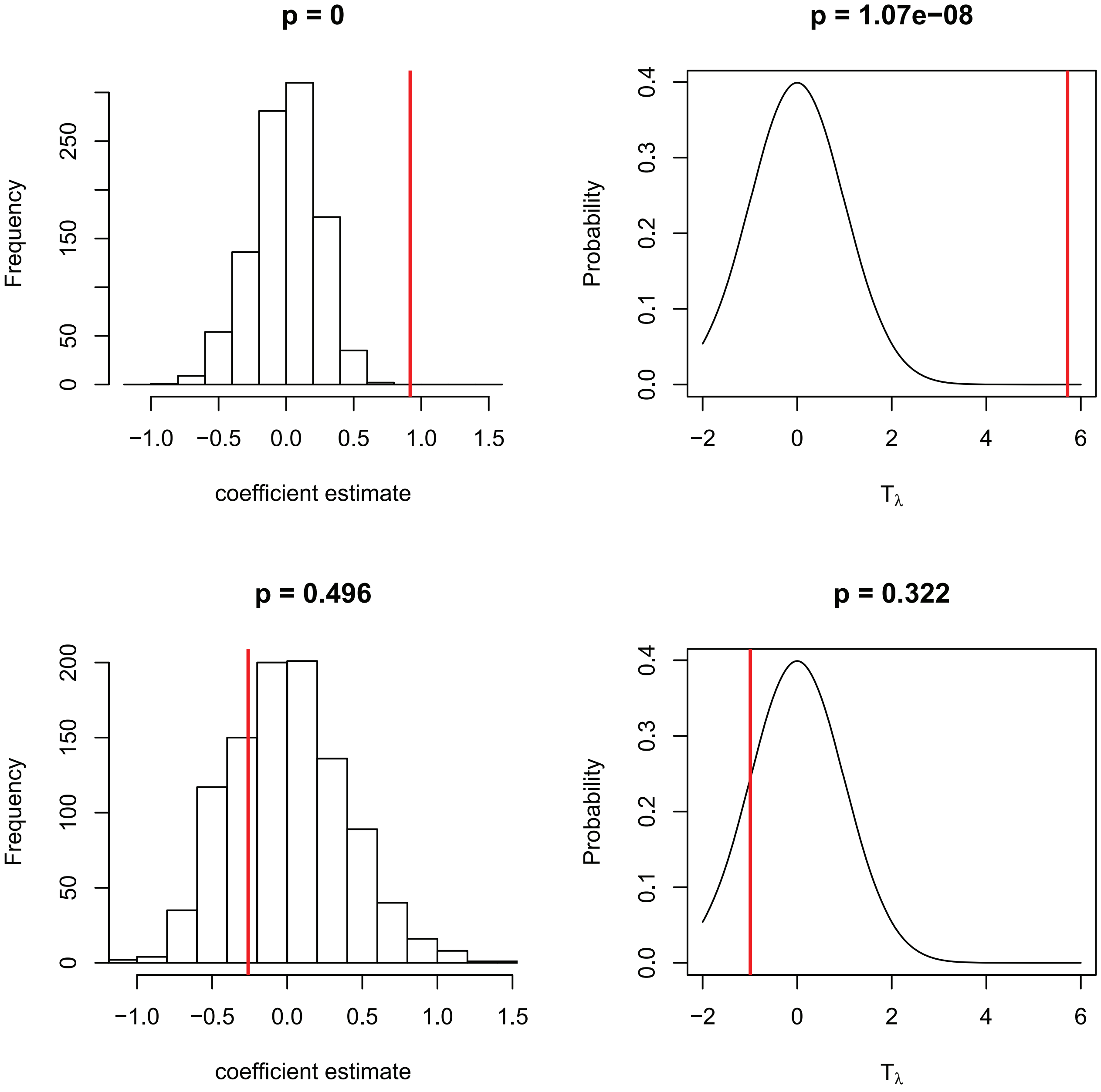

Figure 3 Null distributions of the regression coefficients and tests statistics in the ridge regression model for data with continuous outcomes. Left column - histogram of fitted coefficients obtained from the permutation test, with the coefficient fitted to the original data indicated by a vertical line. Right column - Null distribution of the test statistic used in the approximate test (a standard normal distribution), with the test statistic of the fitted coefficient indicated by a vertical line. Top row, a SNP associated with phenotype; bottom row, a SNP not associated with phenotype. The causal and non-causal SNPs are from the same replicate in the simulation study where n = 500, m = 20, λ = 1. p-values are shown above each plot.

permutation test takes approximately 500 times longer to compute than the approximate test.

Binary phenotypes

Genotypes and corresponding binary phenotypes were generated for nine different data set dimensions: n = 500, 5000 and m = 20, 100, 1000, 2000, and n = 500, m = all SNPs in the 1 Mb region (approximately 10,000 SNPs). The genotype relative risk, r, was specified as 2. For the largest data dimensions (n = 500, $m \approx$ 10, 000), results are not shown for the permutation test, due to the computational time required to fit ridge regression models on data of this size. Similarly, results are not shown when n = 5000 and $m \approx$ 10, 000.

Table 2 Performance comparison in simulated data with continuous outcomes

			Shrinkage Parameter							
			Approximate test				Permutation test			
			0.1	1	10	100	0.1	1	10	100
Individuals	SNPs									
500	20									
		TPR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		FPR	0.045	0.045	0.061	0.133	0.015	0.015	0.017	0.095
	100									
		TPR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		FPR	0.056	0.054	0.071	0.141	0.015	0.018	0.024	0.074
	1000									
		TPR	0.100	0.500	0.900	1.000	0.000	0.200	0.800	1.000
		FPR	0.038	0.045	0.049	0.080	0.007	0.006	0.010	0.029
	ALL									
		TPR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		FPR	0.318	0.071	0.068	0.069	0.019	0.019	0.020	0.020
1000	20									
		TPR	0.900	1.000	1.000	1.000	0.900	1.000	1.000	1.000
		FPR	0.043	0.043	0.087	0.137	0.013	0.013	0.034	0.096
	100									
		TPR	0.900	1.000	1.000	1.000	0.900	0.900	1.000	1.000
		FPR	0.051	0.052	0.060	0.108	0.023	0.023	0.019	0.062
	1000									
		TPR	0.700	0.700	1.000	1.000	0.400	0.500	0.900	1.000
		FPR	0.060	0.058	0.055	0.076	0.007	0.008	0.010	0.020
	ALL									
		TPR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		FPR	0.166	0.155	0.110	0.071	0.015	0.015	0.015	0.017
5000	20									
		TPR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		FPR	0.048	0.048	0.048	0.113	0.006	0.006	0.006	0.053
	100									
		TPR	0.900	0.900	1.000	1.000	0.800	0.900	1.000	1.000
		FPR	0.055	0.052	0.062	0.100	0.003	0.001	0.007	0.055
	1000									
		TPR	0.700	0.700	1.000	1.000	0.700	0.700	0.900	1.000
		FPR	0.046	0.046	0.045	0.060	0.006	0.007	0.008	0.014
	ALL									
		TPR	0.400	0.500	0.900	1.000	0.300	0.900	0.900	1.000
		FPR	0.026	0.027	0.029	0.042	0.007	0.007	0.007	0.009

Performance comparison between a permutation test and the approximate test. Data are simulated genotype data with continuous phenotypes. Reported are proportion of true positive and false positive results at significance threshold $\alpha = 0.05$. TPR = True Positive Rate, FPR = False positive rate. Results for each simulation scenario are averaged over ten replicates.

In Figure 5, the left column shows the null distributions, generated in a permutation test, used to estimate the significance of a ridge regression coefficient for a significant SNP (top row) and for a SNP that is not associated with (bottom row) in data with binary outcomes. The coefficient fitted to the original data is indicated. In the right column, the null distribution of the test statistic used in the approximate test is shown, with the test statistic of the fitted coefficient indicated. Ridge regression models were fitted using the shrinkage parameter $\lambda = 1$. These results are examples taken from a single simulation, and above each plot the p-value according to the permutation test (left) or the approximate test (right) is shown.

Figure 6 compares the ranking of the SNPs from most significant (rank = 1) to least significant. Only twelve SNPs are shown because the SNPs that were invariant in the data were removed before analysis. The SNPs were ranked according to both the approximate test and the

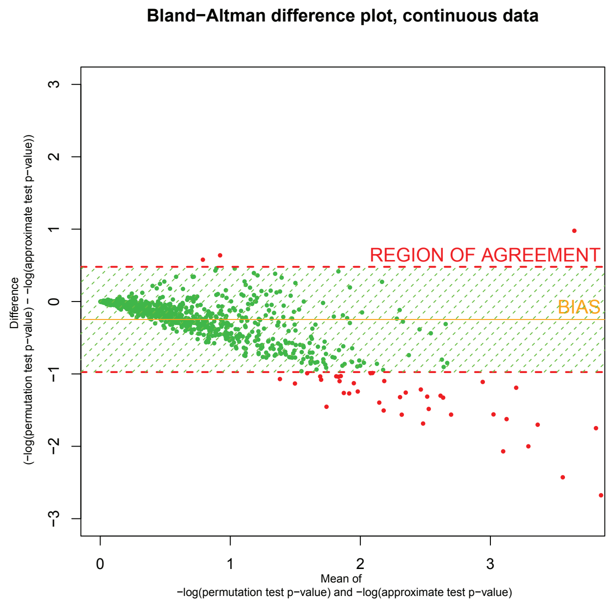

Figure 4 Bland-Altman plot of mean versus difference for *p*-values computed using both tests. *p*-values were computed using both the approximate test and a the permutation test for data with continuous outcomes, and are plotted on the -log scale. n = 5000, m = 1000, λ = 0.1.

permutation test. From Figure 6 we see that whilst the ranking of the SNPs was not identical, the most strongly associated SNPs are ranked as such by both tests.

Table 5 compares the performance of the approximate test and a permutation test of significance for different sized data sets and at different values of the shrinkage parameter: λ = 0.1, 1, 10, 100. For most data set dimensions and values of λ, the z-type test is more conservative than the permutation test, with lower true positive and false positive rates. This is in contrast to the linear regression case.

Comparison with univariate tests of significance

The performance of tests of significance of ridge regression coefficients, in terms of true and false positive rates, was compared to the performance of univariate tests. Comparisons were made in each of the simulation settings: the null simulation, and the simulations with continuous and with binary outcomes. The results are shown in Additional File 1, tables S3, S4 and S5. As would be expected when using a penalised regression approach such as ridge regression, the performance of the corresponding significance test depends on the degree of shrinkage, which is controlled by the shrinkage parameter λ. The performance of the approximate test is comparable to that of a permutation test, with the advantage of a much-reduced computational burden. Further, ridge regression has the advantage over univariate tests of significance that it results in a much lower false positive rate. These advantages of ridge regression compared to univariate methods are further illustrated in the study of lung cancer data which follows.

Table 3 Multiple causal phenotypes

		Approximate test				Permutation Test			
		0.1	**1**	**10**	**100**	**0.1**	**1**	**10**	**100**
SNPs									
100	TPR	0.624	0.717	0.939	1.000	0.252	0.312	0.517	0.910
	FPR	0.064	0.061	0.091	0.250	0.001	0.001	0.004	0.078
1000	TPR	0.210	0.250	0.670	0.970	0.020	0.070	0.170	0.770
	FPR	0.074	0.058	0.060	0.100	0.000	0.000	0.001	0.011

Performance comparison between a permutation test and the approximate test in simulated genotype data with continuous phenotypes and multiple causal SNPs. Reported are proportion of true positive and false positive results at significance threshold α = 0.05. n = 500 and in each scenario ten SNPs are with MAF 0.10 - 0.15 are designated causal with effect size 1; the rest of the SNPs have 0 effect size. TPR = True Positive Rate, FPR = False positive rate. Results for each simulation scenario are averaged over ten replicates.

Table 4 Comparison of computational performance

	Approximate test	Permutation test
time (seconds)	1.936	1043.604

Comparison of computation performance of approximate test and permutation test. n = 1000, m = 1000 and λ = 1. Approximate test and permutation test p - values were computed and the time taken to arrive at the p-values was recorded. Calculations were done using R version 2.12.0 [27] on an iMac running Mac OS X Version 10.6.7, fitted with an 2.8 Ghz Intel Core i7 processor and 16 GB 1067 MHz DDr3 RAM.

Lung cancer data

Genome-wide association studies have identified SNPs associated with lung cancer disease status. SNPs have been identified at chromosomal locations 15q25 [29,30], 5p15 [31] and 6p21 [32]. The associations at 15q25 and 5p15 have been replicated in white populations, but the association at 6p21 has not [33]. Not all studies successfully replicated the associations at 15q25 and 5p15 [34].

Here, we use ridge regression to re-evaluate a set of 35 SNPs for association with lung cancer disease status. Whilst these data are not as high-dimensional as those from a genome-wide study, they allow us to illustrate the features of using ridge regression for genetic data. We show that ridge regression is a useful technique when data are correlated, and illustrate that multivariate methods have advantages over univariate tests of significance.

Data consist of genotypes and non-genetic predictors from approximately 4000 individuals in the European Prospective Investigation into Cancer and Nutrition (EPIC, [35]). Missing genotypes were imputed using mean imputation. Gender, smoking status and age were included as unconstrained parameters in the model.

For the purpose of comparison, univariate (SNP-by-SNP) p-values were calculated. Univariate p-values were calculated by fitting a logistic regression model for each SNP independently, with gender, smoking status and

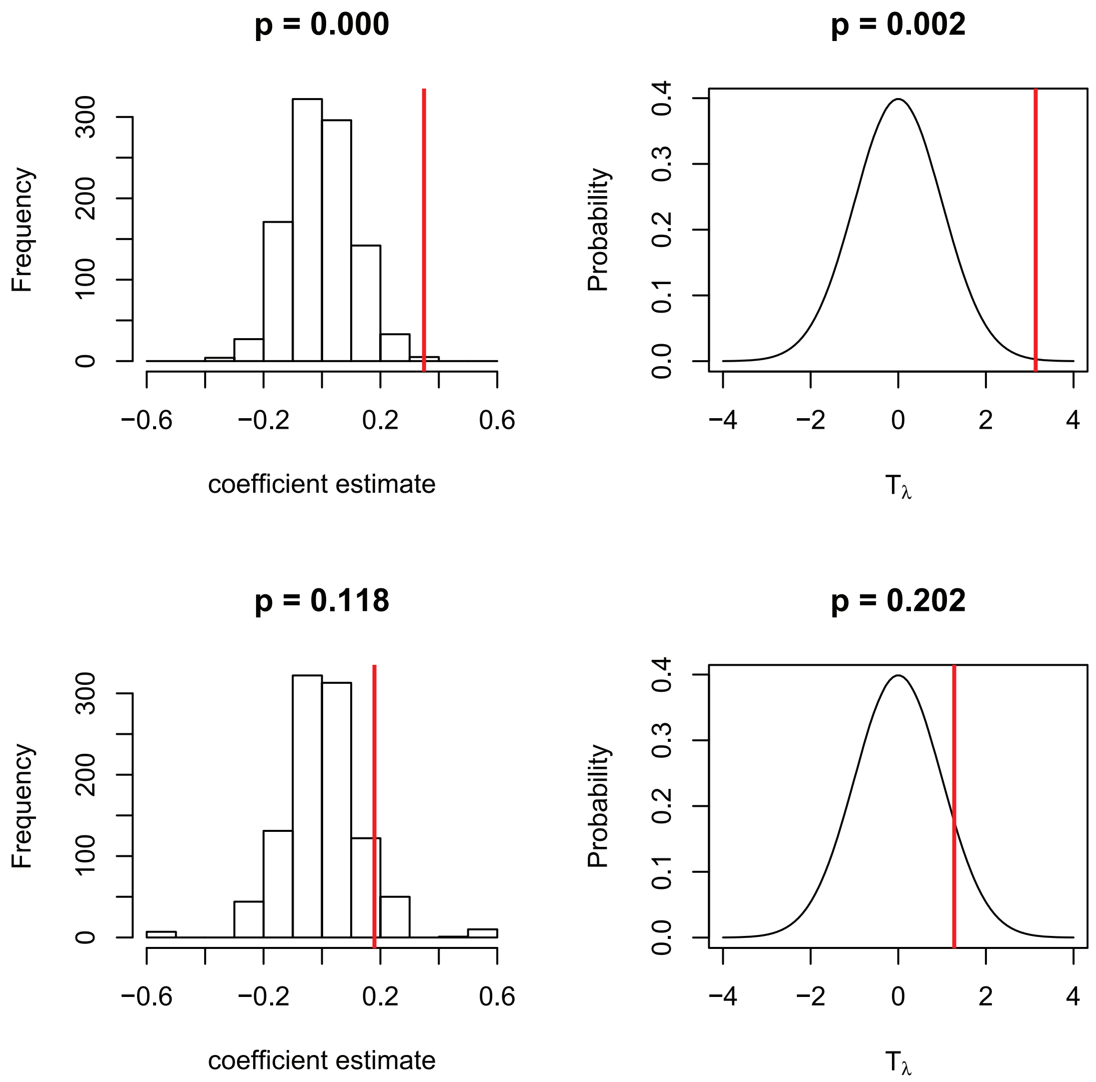

Figure 5 Null distributions of the regression coefficients and tests statistics in the ridge regression model for data with binary outcomes. Left column - histogram of fitted coefficients in the permutation test, with the coefficient fitted to the original data indicated by a vertical line. Right column - Null distribution of the test statistic used in the approximate test (a standard normal distribution) with the test statistic of the fitted coefficient indicated by a vertical line. Top row, a SNP associated with phenotype; bottom row, a SNP not associated with phenotype. The causal and non-causal SNPs are from the same replicate in the simulation study where n = 500, m = 20, λ = 1. p-values are shown above each plot.

age included in the model. A Wald test was then used to estimate the significance of the coefficient.

In this example, $m < n$ and (unpenalised) multivariate regression does give rise to unique parameter estimates. Multivariate p-values (equivalent to p-values from the z-type test for ridge regression coefficients with a ridge penalty of zero) are also reported here.

In a data set of this size, it is computationally feasible to fit a ridge regression model with a range of values of λ. The ridge trace (Figure 1) is a plot of parameter estimates against λ [11]. The ridge trace typically suggests a range of values of λ rather than a single best value. In Figure 1, λ = 150 seems to be the point at which the ridge estimates stabilise, and Table 6 reports the corresponding approximate p-values.

Figure 1 shows the ridge trace, and Figure 2 shows a plot of p-values with increasing λ. Due to low LD between most of the SNPs, most coefficient estimates do not change significantly with increasing λ and nor do estimates of their significance. SNPs rs8034191 and rs16969968 are both located at 15q25 and are in high LD (r^2 = 0.961 in HapMap CEU population, r^2 = 0.81 in our data). In contrast to most of the SNPs, coefficient estimates for these two SNPs do change rapidly with change in the shrinkage parameter.

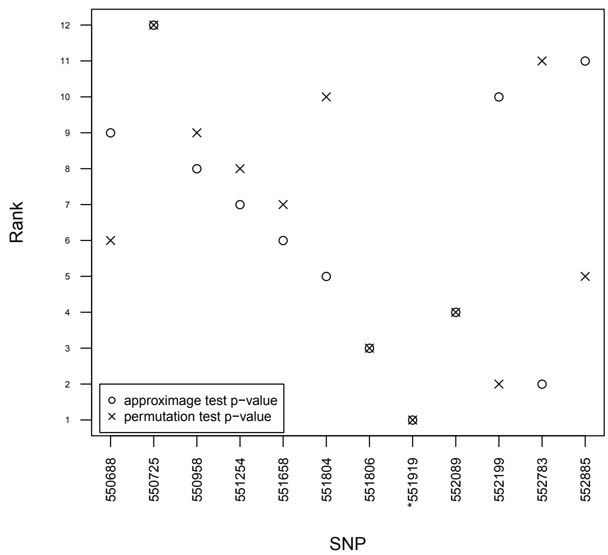

Figure 6 Comparison of the ranking of SNPs. SNPs are ranked by *p*-value, from 1 (most significant) to least significant, by both the approximate test and the permutation test. Data are genotypes with binary outcomes. The SNP name marked with an asterisk * is the causal SNP, and is the most significant SNP according to both tests. $n = 500$, $m = 20$, $\lambda = 1$. Only twelve SNPs are shown because the SNPs that were invariant among the sampled individuals were removed from the data before analysis.

Further, with increasing shrinkage and stabilisation of the estimates, the approximate p-values for these SNPs become significant (Figure 2). rs402710, which is not in LD with other SNPs in the data, is significant in a multiple logistic regression model even when no penalty is included in the model (λ = 0). A further SNP, rs671330, in chromosome 6, has a nominally significant p-value in the approximate test (ranging from 0.048 to 0.040 with increased shrinkage), but again this SNP has not previously been shown to be associated with lung cancer risk.

The p-value trace using permutation test p-values (Figure 2b) shows good agreement with the approximate p-value trace (Figure 2a). Figure 2b is much more computationally expensive to produce than Figure 2a, thus the use of the approximate test makes the plotting of a p-value trace for a range of values of λ more feasible.

Table 6 presents the univariate, multivariate, approximate and permutation test p-values at λ = 150 for four SNPs from the regions which have previously been shown to be associated with lung cancer disease status. Using ridge regression, we replicate the previously found associations at 15q25 and 5p15, but fail to replicate the association at 6p21. Using the univariate test, a further SNP, rs6746834 (on chromosome 2) was nominally significant at the 0.05 level (p = 0.049), but association at this region has not previously been shown. This SNP was not found to be significantly associated with lung cancer disease status by the approximate test. We interpret this as a false positive that arises when univariate tests are used.

The results in Table 6 demonstrate the advantage of multivariate tests, and specifically of ridge regression, over univariate tests of significance. SNP rs402710, which has previously been shown to be associated with lung cancer disease status [31], was not found to be significant using the univariate test but was found as such by the multivariate methods. The two SNPs that are correlated, rs8034191 and rs16969968, were not significant in multiple regression but were significant in ridge regression, demonstrating the advantage of using ridge regression when SNPs are correlated.

These results demonstrate that this approximate test of significance for coefficients fitted using logistic ridge regression reproduces previously ascertained associations, at reduced computational cost compared to a permutation test, even when SNPs are highly correlated.

Conclusions

We present and evaluate the performance of a test of significance for coefficients estimated using ridge regression. We evaluate the test as applied to both linear and logistic ridge regression models. Our evaluation is by means of simulation studies across a range of scenarios representative of genetic data. We evaluate the test by comparing its performance to that of a permutation test.

We evaluate the performance of the test when it is applied to a real data set. The data set comprises lifestyle data and genotypes together with lung cancer case-control status. Using the proposed test, we successfully replicate previously found associations at much reduced computational cost compared to a permutation test. This demonstrates the utility of the test for detecting significant variables when predictor variables are highly correlated, as were two significant SNPs in the lung cancer data.

Stability selection [36] is a method for variable selection that has received attention in the literature in recent years. It combines subsampling of the data with a dimension reduction technique, with the aim of finding consistently significant variables. Penalised regression methods that perform variable selection, such as the Lasso [10] and the Elastic Net [37], have been used in stability selection. To use ridge regression with stability selection, a way of determining which variables to select in each subsample of the data is required. A permutation test would be computationally expensive in this context, rapidly becoming infeasible if for large n, if the guidelines of 100 subsamples of size $n/2$ given by Meinshausen & Bühlmann [36] were followed. The test of significance proposed here, being much less computationally expensive

Table 5 Performance comparison in simulated data with binary outcomes

			Shrinkage Parameter							
			Approximate test				**Permutation test**			
			0.1	**1**	**10**	**100**	**0.1**	**1**	**10**	**100**
Individuals	**SNPs**									
500	20									
		TPR	0.300	0.500	0.900	0.900	0.400	0.600	0.900	0.900
		FPR	0.023	0.036	0.068	0.142	0.078	0.078	0.099	0.174
	100									
		TPR	0.100	0.100	0.500	0.900	0.200	0.200	0.400	0.900
		FPR	0.024	0.037	0.046	0.087	0.050	0.052	0.058	0.115
	1000									
		TPR	0.200	0.300	0.500	0.700	0.100	0.100	0.400	0.700
		FPR	0.103	0.096	0.071	0.054	0.046	0.045	0.047	0.056
	2000									
		TPR	0.000	0.300	0.500	0.700	0.200	0.300	0.300	0.700
		FPR	0.008	0.056	0.081	0.063	0.052	0.049	0.048	0.055
	ALL									
		TPR	0.000	0.000	0.600	0.900	-	-	-	-
		FPR	0.000	0.000	0.014	0.068	-	-	-	-
5000	20									
		TPR	0.700	0.800	1.000	1.000	0.700	0.800	1.000	1.000
		FPR	0.024	0.024	0.030	0.096	0.090	0.083	0.089	0.154
	100									
		TPR	0.400	0.400	0.900	1.000	0.200	0.300	0.900	1.000
		FPR	0.027	0.028	0.041	0.078	0.071	0.067	0.078	0.110
	1000									
		TPR	0.200	0.300	0.600	1.000	0.100	0.200	0.600	1.000
		FPR	0.047	0.046	0.041	0.053	0.053	0.052	0.052	0.062
	2000									
		TPR	0.000	0.200	0.500	1.000	0.000	0.100	0.400	1.000
		FPR	0.074	0.067	0.056	0.057	0.053	0.052	0.053	0.058

Performance comparison between a permutation test and the approximate test in simulated genotype data with case-control phenotypes. Reported are proportion of true positive and false positive results at significance threshold α = 0.05. TPR = True Positive Rate, FPR = False positive rate. Results for each simulation scenario are averaged over ten replicates.

than a permutation test, makes the combination of ridge regression and stability selection a feasible possibility.

The choice of shrinkage parameter in ridge regression is discussed in the literature, but no consensus method provides an universally optimum choice. The proposed test performs well over a range of values of λ. We introduce the p-value trace, a plot of the change in the negative logarithm of the p-value of the regression coefficients as the shrinkage parameter increases from zero. This trace can be used in combination with the ridge trace of [11] to visualise how the significance of the regression parameters, as well as their value, changes with increasing shrinkage. Such a plot could aid the identification of significant coefficients in the ridge model.

Table 6 Lung cancer data

SNP (chromosome)	**Univariate *p*-value**	**Multivariate *p*-value unpenalised**	**Approximate *p*-value λ = 150**	**Permutation *p*-value λ = 150**
rs8034191 (15q25)	0.009	0.603	0.007	0.006
rs16969968 (15q25)	0.001	0.183	0.001	0.000
rs402710 (5p15)	0.213	0.012	0.011	0.004
rs4324798 (6p21)	0.513	0.231	0.248	0.251

Of the 35 SNPs in the lung cancer data set, four have been previously reported to be associated with lung cancer disease status. For these four SNPs, this table reports univariate, multivariate, approximate and permutation test *p*-values.

Additional material

Additional file 1: Table S1 - Performance comparison in null simulation using t-type test. **Table S2** - Performance comparison in simulated data with continuous outcomes using t-type test. **Table S3** - Performance comparison in null simulation including comparison to univariate tests of significance. **Table S4** - Performance comparison with continuous outcomes including comparison to univariate tests of significance. **Table S5** - Performance comparison with binary outcomes including comparison to univariate tests of significance.

Acknowledgements

We thank the European Prospective Investigation into Cancer and Nutrition (EPIC): Elio Riboli, Anne Tjønneland, Kim Overvad, Francoise Clavel-Chapelon, Rudolf Kaaks, Heiner Boeing, Antonia Trichopoulou, Franco Berrino, Domenico Palli, Salvatore Panico, Rosario Tumino, Paolo Vineis, Bas Buneo de Mesquita, Petra Peeters, Eiliv Lund, Carlos A. González, Aurelio Barricarte, Miren Dorronsoro, Maria José Sánchez, Carmen Navarro, J. Ramón Quirós, Jonas Manjer, Goran Hallmans, Nicholas E. Day, Kay-Tee Khaw, Nicholas Wareham and Timothy J. Key.

Authors' contributions

EC ran the simulations, analysed the data, and wrote the paper. PV was involved in the collection of the lung cancer data, supervised the analysis, and read and commented on the draft. MDI came up with the idea, supervised the work, and commented on drafts of the paper. All authors read and approved the final manuscript.

Received: 7 March 2011 Accepted: 19 September 2011
Published: 19 September 2011

References

1. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467(7319)**:1061-1073.
2. Armitage P: **Tests for Linear Trends in Proportions and Frequencies.** *Biometrics* 1955, **11(3)**:375-386.
3. WTCCC: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447(7145)**:661-678.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461(7265)**:747-753.
5. Park M, Hastie T: **Penalized logistic regression for detecting gene interactions.** *Biostatistics* 2008, **9**:30.
6. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ: **Simultaneous Analysis of all SNPs in Genome-Wide and Re-sequencing Association Studies.** *PLoS Genet* 2008, **4**:e100130.
7. Ayers K, Cordell H: **SNP Selection in genome-wide and candidate gene studies via penalized logistic regression.** *Genet Epidemiol* 2010, **34(8)**:879-891.
8. Frank IE, Friedman JH: **A statistical view of some chemometrics regression tools.** *Technometrics* 1993, **35**:109-148.
9. Malo N, Libiger O, Schork NJ: **Accommodating linkage disequilibrium in genetic-association analyses via ridge regression.** *Am J Hum Genet* 2008, **82(2)**:375-385.
10. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J Roy Stat Soc B* 1996, **58**:267-288.
11. Hoerl AE, Kennard RW: **Ridge Regression: Biased Estimation for Nonorthogonal Problems.** *Technometrics* 1970, **12**:55-67.
12. Whittaker JC, Thompson R, Denham MC: **Marker-assisted selection using ridge regression.** *Genetics Research* 2000, **75(02)**:249-252.
13. Zucknick M, Richardson S, Stronach EA: **Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods.** *Stat Appl Genet Mol Biol* 2008, **7**:Article 7.
14. Sun Y, Shedden K, Zhu J, Choi NH, Kardia S: **Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression.** *BMC Proceedings* 2009, **3(Suppl 7)**:S67.
15. Hsiang T: **A Bayesian view on ridge regression.** *The Statistician* 1975, **24(4)**:267-268.
16. Halawa AM, El Bassiouni MY: **Tests of regression coefficients under ridge regression models.** *J Stat Comput Sim* 2000, **65**:341-356.
17. Cessie SL, Houwelingen JCV: **Ridge Estimators in Logistic Regression.** *J Roy Stat Soc C - App* 1992, **41**:191-201.
18. Hastie T, Tibshirani R: *Generalized Additive Models* London: Chapman & Hall/CRC; 1990.
19. Vago E, Kemeny S: **Logistic ridge regression for clinical data analysis (A case study).** *Appl Ecol Env Res* 2006, **4(2)**:171-179.
20. Hoerl AE, Kennard RW, Baldwin KF: **Ridge Regression: some simulations.** *Commun Stat A - Theor* 1975, **4(2)**:105-123.
21. Lawless J, Wang P: **A simulation study of ridge and other regression estimators.** *Commun Stat A -Theor* 1976, **5(4)**:307-323.
22. Golub G, Heath M, Wahba G: **Generalized cross-validation as a method for choosing a good ridge parameter.** *Technometrics* 1979, **21(2)**:215-223.
23. Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ: **Sequence-Level Population Simulations Over Large Genomic Regions.** *Genetics* 2007, **177(3)**:1725-1731.
24. Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ: **FREGENE: Simulation of realistic sequence-level data in populations and ascertained samples.** *BMC Bioinf* 2008, **9**:364.
25. Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ: *BARGEN -Bioinformatics for the Analysis and Exploitation of Resequenced Genomes* 2007 [http://www.ebi.ac.uk/projects/BARGEN/], Accessed 8 December 2008.
26. Tzeng JY: **Evolutionary-based grouping of haplotypes in association analysis.** *Genet Epidemiol* 2005, **28(3)**:220-31.
27. R Development Core Team: *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria* 2009 [http://www.R-project.org].
28. Altman D, Bland J: **Measurement in Medicine: The Analysis of Method Comparison Studies.** *Journal of the Royal Statistical Society. Series D (The Statistician)* 1983, **32(3)**:307-317.
29. Hung RJ, Mckay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, Mclaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-De-Mesquita BH, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcatova I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P: **A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.** *Nature* 2008, **452(7187)**:633-637.
30. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS: **Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1.** *Nat Genet* 2008, **40(5)**:616-622.
31. McKay J, Hung R, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, McLaughlin J, Shepherd F, Montpetit A, Narod S, Krokan H, Skorpen F, Elvestad M, Vatten L, Njølstad I, Axelsson T, Chen C, Goodman G, Barnett M, Loomis M, Lubiñski J, Matyjasik J, Lener M, Oszutowska D, Field J, Liloglou T, Xinarianos G, Cassidy A, Vineis P, Clavel-Chapelon F, Palli D, Tumino R, Krogh V, Panico S, González C, Ramón Quirós J, Martínez C, Navarro C, Ardanaz E, Larrañaga N, Kham K, Key T, Bueno-de Mesquita H, Peeters P, Trichopoulou A, Linseisen J, Boeing H, Hallmans G, Overvad K, Tjønneland A, Kumle M, Riboli E, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P: **Lung cancer susceptibility locus at 5p15.33.** *Nat Genet* 2008, **40(12)**:1404-1406.
32. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, Qureshi M, Dong Q, Gu X, Chen WV, Spitz MR, Eisen T, Amos CI,

Houlston RS: **Common 5p15.33 and 6p21.33 variants influence lung cancer risk.** *Nat Genet* 2008, **40(12)**:1407-1409.

33. Truong T, Hung RJ, Amos CI, Wu X, Bickeboller H, Rosenberger A, Sauter W, Illig T, Wichmann HE, Risch A, Dienemann H, Kaaks R, Yang P, Jiang R, Wiencke JK, Wrensch M, Hansen H, Kelsey KT, Matsuo K, Tajima K, Schwartz AG, Wenzlaff A, Seow A, Ying C, Staratschek-Jox A, Nurnberg P, Stoelben E, Wolf J, Lazarus P, Muscat JE, Gallagher CJ, Zienolddiny S, Haugen A, van der Heijden HFM, Kiemeney LA, Isla D, Mayordomo JI, Rafnar T, Stefansson K, Zhang ZF, Chang SC, Kim JH, Hong YC, Duell EJ, Andrew AS, Lejbkowicz F, Rennert G, Muller H, Brenner H, Le Marchand L, Benhamou S, Bouchardy C, Teare MD, Xue X, McLaughlin J, Liu G, McKay JD, Brennan P, Spitz MR: **Replication of Lung Cancer Susceptibility Loci at Chromosomes 15q25, 5p15, and 6p21: A Pooled Analysis From the International Lung Cancer Consortium.** *J Natl Canc Inst* 2010, **102(13)**:959-971.
34. Yang P, Li Y, Jiang R, Cunningham JM, Zhang F, de Andrade M: **A Rigorous and Comprehensive Validation: Common Genetic Variations and Lung Cancer.** *Canc Epidemiol Biomarkers Prev* 2010, **19**:240-244.
35. Riboli E, Hunt K, Slimani N, Ferrari P, Norat T, Fahey M, Charrondière U, Hémon B, Casagrande C, Vignat J, Overvad K, Tjønneland A, Clavel-Chapelon F, Thièbaut A, Wahrendorf J, Boeing H, Trichopoulos D, Trichopoulou A, Vineis P, Palli D, Bueno-de Mesquita H, Peeters P, Lund E, Engeset D, González C, Barricarte A, Berglund G, Hallmans G, Day N, Key T, Kaaks R, Saracci R: **European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection.** *Public Health Nutrition* 2002, **5(6b)**:1113-1124.
36. Meinshausen N, Bühlmann P: **Stability selection.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010, **72(4)**:417-473.
37. Zou H, Hastie T: **Regularization and Variable Selection via the Elastic Net.** *Journal of the Royal Statistical Society B* 2005, **67(2)**:301-320.

doi:10.1186/1471-2105-12-372
Cite this article as: Cule *et al.*: Significance testing in ridge regression for genetic data. *BMC Bioinformatics* 2011 **12**:372.