Cherlin *et al. BMC Proceedings* 2018, **12**(Suppl 9):38
https://doi.org/10.1186/s12919-018-0149-2

BMC Proceedings

PROCEEDINGS Open Access

Using penalized regression to predict phenotype from SNP data

Svetlana Cherlin*, Richard A. J. Howey and Heather J. Cordell

From Genetic Analysis Workshop 20
San Diego, CA, USA. 4 - 8 March 2017

Abstract

Background: In a typical genome-enabled prediction problem there are many more predictor variables than response variables. This prohibits the application of multiple linear regression, because the unique ordinary least squares estimators of the regression coefficients are not defined. To overcome this problem, penalized regression methods have been proposed, aiming at shrinking the coefficients toward zero.

Methods: We explore prediction of phenotype from single nucleotide polymorphism (SNP) data in the GAW20 data set using a penalized regression approach (LASSO [least absolute shrinkage and selection operator] regression). We use 10-fold cross-validation to assess predictive performance and 10-fold nested cross-validation to specify a penalty parameter.

Results: By analyzing approximately 600,000 SNPs we find that, when the sample size comprises a few hundred individuals, SNP effects are heavily penalized, resulting in a poor predictive performance. Increasing the sample size to a few thousand individuals results in a much smaller penalization of the true effects, thus greatly improving the prediction.

Conclusions: LASSO regression results in a heavy shrinkage of the regression coefficients, and also requires large sample sizes (several thousand individuals) to achieve good prediction.

Background

In a typical genome-wide association study (GWAS), several thousands to several millions of single nucleotide polymorphism (SNP) markers are genotyped in a sample size of several hundred to several thousand individuals, thus leading to many more predictor variables than response variables. In this case, multiple linear regression cannot be used because the unique ordinary least squares estimators of the regression coefficients are not defined. Methods that allow for more predictors than observations [1] may cause model overfitting. Overfitted models are likely to demonstrate poor predictive ability when applied to new data. To overcome these problems, penalized regression methods have been proposed [2–6], aiming at shrinking the regression coefficients toward zero. Depending on the form of the penalty function, some methods (eg, ridge regression [3]) only shrink the coefficients without setting them to zero, whereas other methods (eg, the least absolute shrinkage and selection operator [LASSO] regression [2]) allow shrinkage of the coefficients down to exactly zero, thus performing variable selection.

The strength of the penalty is controlled by a regularization parameter that determines the amount of shrinkage imposed. One challenge of penalized approaches is choosing an optimal value of the regularization parameter. This is often done by *k*-fold cross-validation to find the parameter value in the training folds that minimizes the average mean squared error in the test folds. Assessing the predictive performance can also be done using *k*-fold cross-validation. In this case, the two cross-validation experiments are combined into one so-called nested cross-validation. In nested cross-validation, an outer cross-validation loop is used to assess the predictive performance, while, within each outer fold, an inner cross-validation loop is used to find the regularization parameter [7]. The most commonly

* Correspondence: svetlana.cherlin@ncl.ac.uk
Institute of Genetic Medicine, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK

used number of inner and outer folds is 10 because it provides a reasonable classification accuracy [8].

Here, we focus on LASSO linear regression, which has the property of variable selection. We apply 10-fold cross-validation to assess the out-of-sample predictive performance, using 10-fold nested cross-validation to specify the penalty parameter. We explore the effect of the sample size on the predictive ability of LASSO regression.

Methods

The GAW20 data are based on the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study data set [9] that investigated the epigenetic determinants of triglyceride (TG) response. TGs are major blood lipids [10] that constitute an important biomarker of cardiovascular disease risk [11]. Previous GWAS studies found a number of loci associated with TG levels [12]. We focus on predicting the TG response by analyzing the GWAS (SNP) data and four measures of TG. The first two measures (at visits 1 and 2) were taken before the lipid-lowering drug treatment; the second two measures (at visits 3 and 4) were taken after the treatment.

We performed quality control (QC) on the GAW20 GWAS (SNP) data using standard procedures outlined in Turner et al. [13]. SNP-level QC removes 63,907 SNPs with low minor allele frequency (< 0.01), and 2694 SNPs for failing a test of Hardy-Weinberg equilibrium ($p \leq 0.00001$). Of 822 individuals for whom we have SNP data, the simulated phenotype is available for 680 individuals and the real phenotype is available for 778 individuals. Working on a log scale, we take the mean of the TG measures for visits 1 and 2 as a baseline measure, and the mean of visits 3 and 4 as a follow-up measure. If the measure for either visit 1 or 2 is not available, we take the only available measure as a baseline measure. Similarly, if the measure for either visit 3 or 4 is not available, we take the only available measure as a follow-up measure. We adjust the follow-up measure for the baseline measure, age, center, smoking status, and first 20 principal components (PCs) of SNP effects using a linear regression. The number of PCs was defined by examining the quantile–

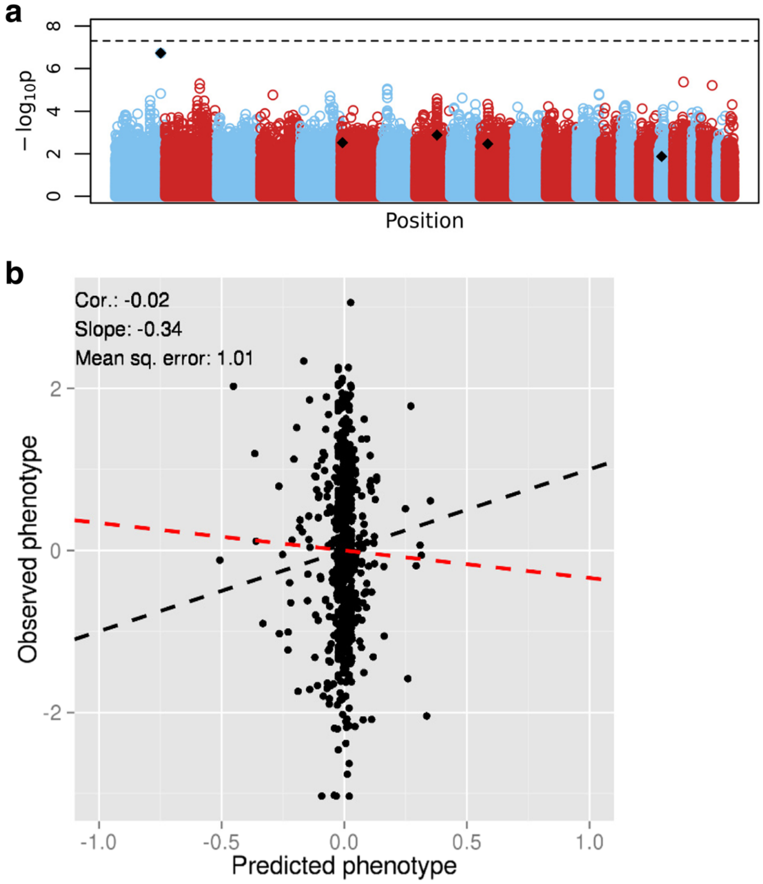

Fig. 1 Analysis of the GAW20 simulated data, replicate 84. **a**, Manhattan plot of *p* values from tests of association between SNP and phenotype. The *black dots* represent causal SNPs. The *dashed line* represents genome-wide significance. **b**, Prediction results. The *black dashed line* is the equality line; the *red dashed line* is the best-fit line

quantile (Q-Q) plot of the p values from the ordinary linear regression. When 20 PCs were incorporated into the linear regression, the Q-Q plot showed no inflated p values, which suggests that the relatedness and population stratification are accounted for. We take the standardized residuals as our final phenotype.

Lasso

Consider a standard multiple linear regression, $\boldsymbol{y} = \beta_0 + \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is a vector of response variables; $\boldsymbol{X}$ is a $n \times p$ matrix of predictor variables; β_0 is an intercept; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a vector of regression coefficients; and $\boldsymbol{\epsilon}$ is a vector of the error terms, $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2)$. For $n > p$ the estimated values of the coefficients are found by minimizing the residual sum of squares:

$$\hat{\beta}_o, \hat{\boldsymbol{\beta}} = \operatorname{argmin}\left[\sum\nolimits_{i=1}^{n}\left(y_i - \beta_o - \sum\nolimits_{j=1}^{p}\beta_j X_{ij}\right)^2\right]$$

However, in a typical GWAS, $n < p$. In this case, penalized regression is often used, where the estimators of $\boldsymbol{\beta}$ are found by minimizing the sum of the residual sum of squares and a penalty function:

$$\hat{\beta}_o, \hat{\boldsymbol{\beta}} = \operatorname{argmin}\left[\sum\nolimits_{i=1}^{n}\left(y_i - \beta_o - \sum\nolimits_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda P(\lambda, \boldsymbol{\beta})\right]$$

Here $P(\lambda, \boldsymbol{\beta})$ is the penalty function with a regularization parameter λ which controls the amount of shrinkage. The LASSO penalty [2] utilizes an ℓ_1-norm penalty, that is, $P(\lambda, \boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{\ell_1}$; consequently, the estimators of the coefficients take the form:

$$\hat{\beta}_o, \hat{\boldsymbol{\beta}} = \operatorname{argmin}\left[\sum\nolimits_{i=1}^{n}\left(y_i - \beta_o - \sum\nolimits_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda \sum\nolimits_{j=1}^{p}|\beta_j|\right]$$

One important property of the LASSO penalty is that it allows the coefficients to be set to exactly zero, thus performing variable selection.

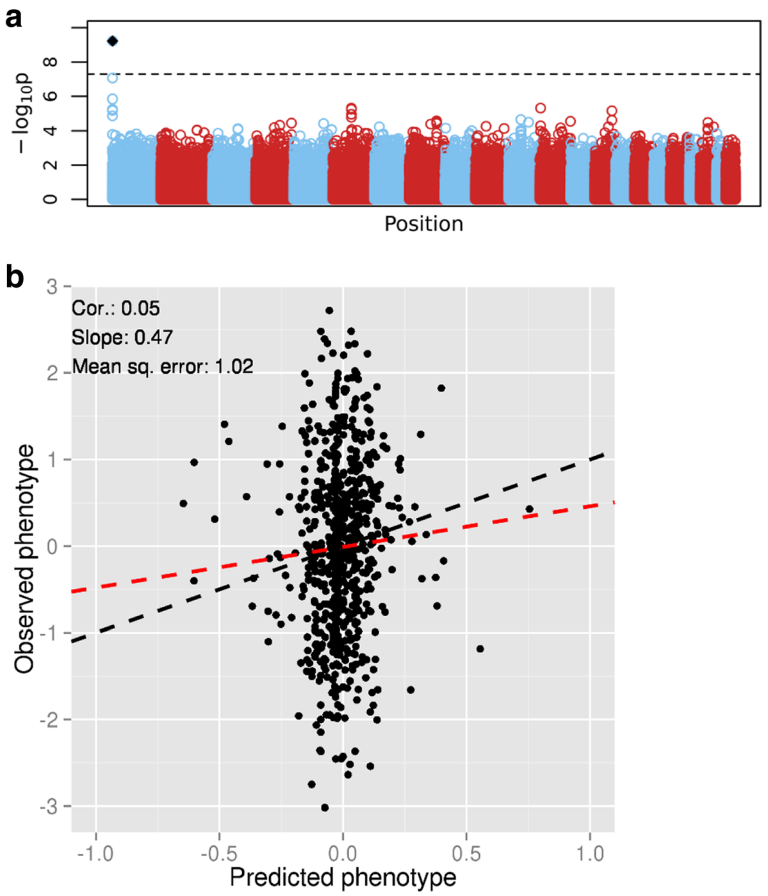

Fig. 2 Analysis of the subset of the illustrative simulated data set (700 individuals). **a**, Manhattan plot of p values from tests of association between SNP and phenotype. The *black dot* represents the causal SNP. The *dashed line* represents genome-wide significance. **b**, Prediction results. The *black dashed line* is the equality line; the *red dashed line* is the best-fit line

Results

Simulated data

We analyzed replicate 84 of the simulated data, consulting the "answers" before performing the analysis. The Manhattan plot from standard linear regression shows no genome-wide significant associations, however, one of the known causal SNPs is nearly genome-wide significant (Fig. 1a). We applied LASSO regression on these data and assessed the predictive performance through 10-fold nested cross-validation, after reducing the number of SNPs to approximately 56,000 using a linkage disequilibrium (LD)–based clumping procedure implemented in the PLINK software [14]. This procedure groups correlated SNPs together and chooses one representative SNP per group, thus reducing dimensionality and eliminating the redundancy of information in the data. We assessed the prediction accuracy through the Pearson correlation coefficient (the square root of heritability), the mean squared error (lower values indicate better fit), and the slope of the best-fit line (a slope of 1 suggests perfect prediction). Figure 1b shows that the predicted phenotypes are heavily shrunk toward zero, resulting in a very low correlation between observed and predicted values and a slope that is far from 1. Poor prediction can be explained by the fact that the effect (regression coefficient) of the most significant true causal SNP is heavily shrunk (a mean estimate of – 0.051 and a SD of 0.062 over the 10 folds, compared to a mean estimate of – 0.443 and a SD of 0.02 over the 10 folds in standard linear regression).

To investigate whether increasing the sample size can improve the prediction, we simulated phenotypes for 7753 individuals for whom we had SNP genotype data available from previous studies. The phenotypes were simulated using only 1 causal SNP with an effect size of – 0.44 (similar to the effect of the nearly genome-wide significant causal SNP from the GAW20 simulated data) and heritability of 0.05, which induces a correlation of approximately 0.23 between the observed and the predicted phenotype. First, we analyzed a subset of these data comprising 700 individuals (similar to the sample size of the

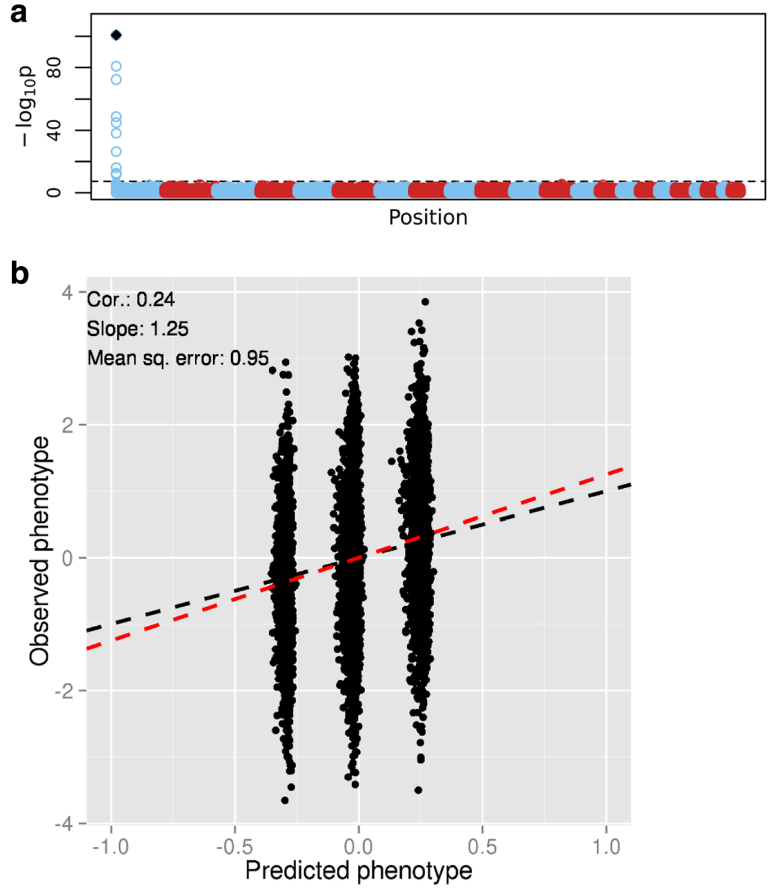

Fig. 3 Analysis of the illustrative simulated data set (7753 individuals). **a**, Manhattan plot of *p* values from tests of association between SNP and phenotype. The *black dot* represents the causal SNP. The *dashed line* represents genome-wide significance. **b**, Prediction results. The *black dashed line* is the equality line; the *red dashed line* is the best-fit line

GAW20 simulated data). The results resemble those seen in the GAW20 simulated data (Fig. 2). Even though the causal SNP is consistently picked up by LASSO, its effect is poorly estimated (the mean estimate is – 0.096 and the SD is 0.019 over the 10 folds) and has the same order of magnitude as the SNPs (false positive) chosen by LASSO, resulting in poor prediction. We then repeat the analysis using the full sample (7753 individuals). The Manhattan plot shows that the significance of the causal SNP has greatly increased (Fig. 3a). The prediction plots show three distinct clusters representing the estimated effects within the three genotype categories of the causal SNP (Fig. 3b). The effect of the causal SNP is much better estimated (the mean estimate is – 0.247 and the SD is 0.009 over the 10 folds) and the effect size is 10 to 1000 times greater than that of the non-causal SNPs falsely chosen by LASSO. The range of the regularization parameter across folds for the full data set (0.046 to 0.056) is smaller than that for the subset (0.081 to 0.183). This suggests that the regression coefficients for the full data set are much less penalized, greatly improving the prediction.

Real data

We also applied LASSO regression to the real GAW20 data and assessed the predictive performance through 10-fold nested cross-validation, after again reducing the number of SNPs to approximately 56,000 using an LD-based clumping procedure. The Manhattan plot shows no genome-wide significant associations (Fig. 4a) and it resembles the Manhattan plot of the GAW20 simulated data. The prediction results are very similar to the results seen in the GAW20 simulated data (Fig. 4b). In a sense, these results are not surprising given the similar sample sizes of the simulated and the real data sets.

Discussion

The prediction ability of LASSO was assessed on the simulated and the real GAW20 data sets through 10-fold nested cross-validation. Both data sets demonstrated poor predictive performance and exhibited a noticeable shrinkage of the predicted phenotype toward zero. By examining the effect of the most significant causal SNP in the simulated data, we found that it was heavily penalized. To

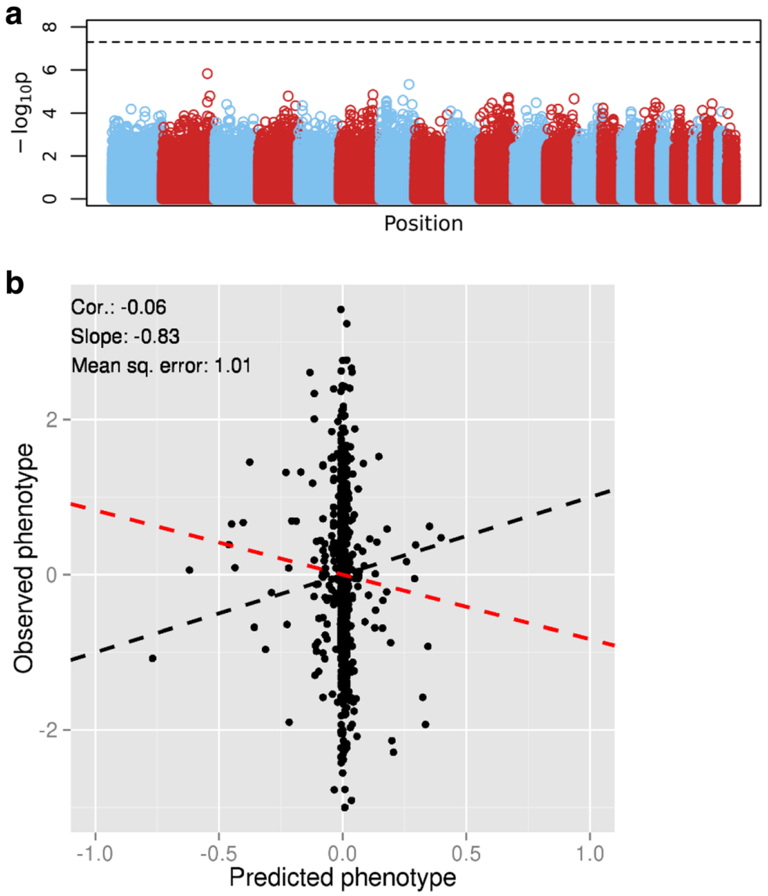

Fig. 4 Analysis of the GAW20 real data set. **a**, Manhattan plot of *p* values from tests of association between SNP and phenotype. The *dashed line* represents genome-wide significance. **b**, Prediction results. The *black dashed line* is the equality line; the *red dashed line* is the best-fit line

investigate this issue, we analyzed a much larger data set (approximately 7000 individuals). We found that the effect of the causal SNP was much less penalized, thus enabling the best prediction possible with that SNP.

Conclusions

The prediction ability of LASSO was assessed on the GAW20 data sets and on the much larger data set available from previous studies. Poor predictive performance is achieved for data sets of a few hundred individuals with a weak signal. This can be explained by the fact that the LASSO regression coefficients are substantially shrunk. Other regularized methods that do not result in such a heavy shrinkage of the regression coefficients might be of use. For example, with hyper-LASSO [6] the extent of the shrinkage depends on the size of the coefficients, and adaptive LASSO [5] uses different adaptive weights for penalizing different coefficients. Both of these can potentially lead to a moderate shrinkage. However, with LASSO, increasing the sample size from a few hundred to a few thousand individuals increased the strength of the signal and reduced the amount of shrinkage of the regression coefficients, thus improving the prediction. We conclude that LASSO regression requires large sample sizes (several thousands of individuals) to achieve good prediction.

Funding

Publication of this article was supported by NIH R01 GM031575. Support for this work was provided by the Wellcome Trust (Grant 102858/Z/13/Z) and an MRC/Arthritis Research UK award: Maximizing Therapeutic Utility in RA (MATURA) (Grant MR-K015346).

Availability of data and materials

The data that support the findings of this study are available from the Genetic Analysis Workshop (GAW), but restrictions apply to the availability of these data, which were used under license for the current study. Qualified researchers may request these data directly from GAW.

About this supplement

This article has been published as part of BMC Proceedings Volume 12 Supplement 9, 2018: Genetic Analysis Workshop 20: envisioning the future of statistical genetics by exploring methods for epigenetic and pharmacogenomic data. The full contents of the supplement are available online at https://bmcproc.biomedcentral.com/articles/supplements/volume-12-supplement-9.

Authors' contributions

SC conducted the statistical analysis and drafted the manuscript. RAJH conducted the transformation of the data to a PLINK-readable format and performed QC on the data. HJC conceived the overall study and critically revised the manuscript. All the authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 17 September 2018

References

1. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2004;32(2):407–99.
2. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267–88.
3. Cessie SL, Houwelingen JCV. Ridge estimator in logistic regression. J R Stat Soc Ser C Appl Stat. 1992;41(1):191–201.
4. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67(2):301–20.
5. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–29.
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(7):e1000130.
7. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7(1):91.
8. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI-95. 1995:1137–43.
9. Irvin MR, Zhi D, Joehanes R, Mendelson M, Aslibekyan S, Claas SA, Thibeault KS, Patel N, Day K, Jones LW, et al. Epigenome-wide association study of fasting blood lipids in the genetics of lipid- lowering drugs and diet network study. Circulation. 2014;130(7):565–72.
10. Kwiterovich PO Jr. The metabolic pathways of high-density lipoprotein, low-density lipoprotein, and triglycerides: a current review. Am J Cardiol. 2000;86(12A):5L–10L.
11. Miller M, Stone NJ, Ballantyne C, Bittner V, Criqui MH, Ginsberg HN, Goldberg AC, Howard WJ, Jacobson MS, Kris-Etherton PM, et al. Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. Circulation. 2011;123(20):2292–333.
12. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274–83.
13. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, et al.: Quality control procedures for genome wide association studies. Curr Protoc Hum Genet 2011; Chapter 1: Unit 1.19.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007;81(3):559–75.