Briefings in Bioinformatics Advance Access published June 17, 2013

Forward LASSO analysis for high-order interactions in genome-wide association study

Huijiang Gao, Yang Wu*, Jiahan Li, Hongwang Li, Junya Li and Runqing Yang*

Submitted: 17th January 2013; Received (in revised form): 30th April 2013

Abstract

Previous genome-wide association study (GWAS) focused on low-order interactions between pairwise single-nucleotide polymorphisms (SNPs) with significant main effects. Little is known how high-order interactions effect, especially one among the SNPs without main effects regulates quantitative traits. Within the frameworks of linear model and generalized linear model, the LASSO with coordinate descent step can be used to simultaneously analyze thousands and thousands of SNPs for normal and discrete traits. With consideration of high-order interactions among SNPs, a huge number of genetic effects make the LASSO failing to work under the presented condition of computation. Forward LASSO analysis is, therefore, proposed to shrink most of genetic effects to be zeros stage by stage. Simulation demonstrates that our proposed method could be used instead of the LASSO method for full model in mapping high-order interactions. Application of forward LASSO method is provided to GWAS for carcass traits and meat quality traits in beef cattle.

Keywords: *genome-wide association study; forward LASSO; high-order interaction; quantitative trait; discrete trait*

INTRODUCTION

The interaction of alleles at different loci is ubiquitously observed in plant, animal and biomedical studies of human diseases. Both physiological theories and empirical findings imply that the complexity of genetic regulations and metabolic pathways are results of complex networks of interactions involving multiple genes [1–3]. Therefore, gene interaction plays an essential role for understanding the genetic architecture and its dynamics underlying the quantitative trait and complex disease of interest [4, 5]. Because of the modeling challenges and computational costs of investigating higher-order gene–gene interactions, however, current genome-wide linkage analyses or association studies have been focused on identifying main genetic effects [6–13] and pairwise interaction effects [6, 14–23]. Although numerous evidences suggest that genes not only interact with each other in a pairwise manner but could also be involved in complicated networks of high-order interactions [24], little is known about how high-order interaction effect governs complex quantitative traits.

With the aid of high-throughout single-nucleotide polymorphisms (SNPs) genotyping technology, genome-wide association study (GWAS) provides

Corresponding author. Runqing Yang, Research Centre for Fisheries Resource and Environment, Chinese Academy of Fishery Sciences, Beijing 100141, People's Republic of China. Tel.: +86-10-68673337; Fax: +86-10-68673337; E-mail: runqing_yang@yahoo.com
*These authors contributed equally to this work.
Huijiang Gao is an Associate Professor in Institute of Animal Science, Chinese Academy of Agricultural Science. His research interest is to develop statistical methodology of QTL mapping and genetic evaluation in beef cattle.
Yang Wu is currently a Master's student in Institute of Animal Science, Chinese Academy of Agricultural Science.
Jiahan Li is an Assistant Professor of Statistics at the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame. His research interests include high-dimensional data analysis and statistical genetics.
Hongwang Li is currently a Master's student in School of Agriculture and Biology, Shanghai Jiaotong University.
Junya Li is a Professor in Institute of Animal Science, Chinese Academy of Agricultural Science. His research focuses on theory and practice of breeding and genetics in beef cattle.
Runqing Yang is a Professor in Research Centre for Fisheries Resource and Environment, Chinese Academy of Fishery Sciences. He is interested in developing Bayesian methods for mapping QTL with linkage analysis and genome-wide association study.

possibilities of identifying various interaction effects among multiple genes for complex quantitative traits. But locating a few significantly linked interactions of astronomical numbers of possible interaction combinations is an extremely challenging task. To this end, sophisticated statistical procedures and computational algorithms were developed over the past decade of analyzing GWAS data sets. In general, there are three types of statistical methods: non-parametric methods, parametric methods and Bayesian methods. The most representative non-parametric method is the combinatorial partitioning method [25] that identifies partitions of multilocus genotypes according to corresponding inter-individual variation in quantitative trait levels. Inspired by combinatorial partitioning method, a multifactor dimensionality reduction (MDR) method [6] was developed for detecting and characterizing high-order gene–gene and gene–environment interactions in case–control study. Subsequently, Lou *et al.* [26] generalized MDR to analyze both dichotomous and continuous phenotypes while adjusting for both discrete and continuous covariates. Following the partition of gene interaction by Cockerham [27] and Kempthorne [28], parametric methods were developed, where assumptions of biologically meaningful genetic models are required. Among many genetic models, the logic regression is widely used and is regarded as a fundamental parametric method for uncovering gene interaction effect for dichotomous trait [29–32]. Meanwhile, backward regression, stepwise regression and penalized regression strategies were proposed to solve for logic interactive models. Despite their successes in identifying interesting genetic variants and interactions in small data sets, all the aforementioned methods are infeasible for analyzing large genetic data sets. Bayesian method [33], on the other hand, is promising for jointly modeling and testing interactions among more genotyped SNPs for dichotomous disease traits. But because the computing burden of Bayesian analysis largely depends on the complexity of Monte Carlo Markov Chain algorithms and the sample size, this framework may not be considered for analyzing higher-order interactions among numerous genetic variants.

Compared with astronomical interaction combinations among all genotyped SNPs across the entire genome, only a few main effects and interaction effects are expected to be important for elucidating the complex genetic architecture. This is consistent with the sparsity assumption of LASSO [34–36] regressions for high-dimensional data analysis and variable selections. In particular for analyzing large-scale genomic data, LASSO with coordinate descent step [37] or Gibbs samplers [38] have been efficiently used to detect SNPs with significant main effects. Given the capability of LASSO regressions in analyzing the main effects of hundreds of thousands of SNPs simultaneously, once carefully designed, LASSO-based statistical methods can be used to identify interaction effect among SNPs. In this study, we integrate LASSO regressions into the general forward regression strategy, resulting in a stagewise interaction effect model for identifying high-order interactions for both quantitative traits and discrete traits in GWAS. Unlike many other methods where an interaction between two predictors is included only if both predictors are marginally significant, our strategy is built on weak heredity principle of interaction effects [39]. In other words, not all SNPs in an interaction have to exhibit significant marginal effects. In this way, the proposed method could efficiently and effectively select important main effects as well as higher-order interactions that may be missed by other methods. The statistical power and computational efficiency of this forward LASSO method is demonstrated by computer simulations. Moreover, it is applied to GWAS analysis of birth weight and marbling in beef cattle.

METHOD

Linear genetic model for quantitative traits

In genome-wide association studies for gene mappings, phenotypic values are observed and m SNP markers are genotyped for n individuals drawn from a randomized population. If the trait of interest is normally distributed and only additive main and interaction effects of SNPs are considered, the linear model for phenotype u_i, $i = 1, 2, \cdots, n$ can be described as

$$\begin{aligned} u_i = \varphi + \sum_{j=1}^{m} x_{ij}\alpha_j + \sum_{j=1}^{m-1}\sum_{j'=j+1}^{m} x_{ij}x_{ij'}\delta_{jj'} \\ + \sum_{j=1}^{m-2}\sum_{j'=j+1}^{m-1}\sum_{j''=j'+1}^{m} x_{ij}x_{ij'}x_{ij''}\delta_{jj'j''} + \cdots + \varepsilon_i \end{aligned} \qquad (1)$$

where φ is the population mean, α_j is additive genetic effect of the jth SNP, x_{ij} is the indicator variable corresponding to the jth SNP genotype, defined as 0 for heterozygote, -1 and 1 for the two homozygote, as usual. $\delta_{jj'}$ is additive × additive interaction effect between the jth and j'th SNPs (two-way interaction), $\delta_{jj'j''}$ is additive × additive × additive interaction effect among the jth, j'th and j''th SNPs (third-order interaction). ε_i is normally distributed residual error with mean zero and residual variance σ^2.

Usually, the number of SNPs, m, far exceeds the sample size in genome-wide association studies, so that model (1) becomes the supersaturated with $n \ll M$ being the number of genetic effects. In practice, however, only a few genetic effects are non-zero because of the limited number of QTLs governing the trait of interest. By solving the following least squares using LASSO with a coordinate descent step,

$$\min\left[\sum_{i=1}^{n}\left(u_i-\varphi-\sum_{j=1}^{m}x_{ij}\alpha_j-\sum_{j=1}^{m-1}\sum_{j'=j+1}^{m}x_{ij}x_{ij'}\delta_{jj'}\right.\right.$$
$$\left.-\sum_{j=1}^{m-2}\sum_{j'=j+1}^{m-1}\sum_{j''=j'+1}^{m}x_{ij}x_{ij'}x_{ij''}\delta_{jj'j''}-\cdots\right)^2$$
$$\left.+\lambda\left(\sum_{j=1}^{m}|\alpha_j|+\sum_{j=1}^{m-1}\sum_{j'=j+1}^{m}|\delta_{jj'}|+\sum_{j=1}^{m-2}\sum_{j'=j+1}^{m-1}\sum_{j''=j'+1}^{m}|\delta_{jj'j''}|\right)\right], \tag{2}$$

it is possible to shrink most of genetic effects to zeros. Here, λ is a tuning parameter, which will be optimized with cross-validation.

If all possible interactions are included in the model (1), 2^m genetic effects will be analyzed. Even if only two-way interactions are considered, there are $\frac{1}{2}m(m+1)$ genetic effects to be estimated and tested. As the number of SNPs, m, is usually on the order of hundreds of thousands, implementing LASSO regression directly for the genetic model (1) is neither theoretically valid nor computationally feasible given the current computational resources.

To identify important main effects as well as higher-order interactions in GWAS data analyses, we assume weak heredity on the heredity structures of interaction effects. According to Chipman (39), there are two versions of the effect heredity principle in statistics: strong heredity and weak heredity. Under strong heredity assumption, if the interaction is significant, both predictors should be marginally significant. Under weak heredity, on the other hand, at least one of these predictors is needed to be significant. Clearly, many prevailing methods for identifying interactions in GWAS data sets implicitly assume strong heredity structure, meaning that interactions are tested among a subset of marginally significant SNPs. However, throughout this article, we will only assume weak heredity. This is a weaker assumption, but could greatly facilitate computations as well as final results interpretations. In what follows we will show the weak heredity principle could reduce the model dimensionality dramatically and complete our existing knowledge on the genetic regulatory network by incorporating genetic factors that are marginally insignificant but jointly important.

Specifically, we partition the model selection problems of the full model (1) into many stages according to the order of interactions considered, and then select non-zero genetic effects at each stage by LASSO regressions. We call this strategy forward LASSO, as it is similar to forward regression for model selections. With consideration of the interactions inclusive at least one SNP of non-zero main effect, the forward LASSO shrinkage estimation for high-order interactions is simplified to carry out in the following steps:

(i) Shrink main additive effects to zero by applying LASSO regression to the main additive effect model: $u_i=\varphi+\sum_{j=1}^{m}x_{ij}\alpha_j+\varepsilon_i^{'}$

(ii) Form two-way interaction terms among the whole-genome SNPs and SNPs with non-zero main effects identified in step (i). Then shrink the two-way interaction effects as well as non-zero main effects to zero by a LASSO regression on the following model: $u_i=\varphi+\sum_{j=1}^{k}x_{ij}\alpha_j+\sum_{j=1,j\neq l}^{m}x_{ij}x_{il}\delta_{jl}+\varepsilon_i''$, with k being the number of non-zero main effects and l being the numbering of non-zero main effect.

(iii) Form the third-order interaction terms among the whole-genome SNPs and SNPs with non-zero two-way interaction effects. Then similar to step (ii), shrink third-order interactions, non-zero two-way interactions and non-zero main effects to zero in a LASSO regression containing all the aforementioned terms.

(iv) The rest can be done in the same manner.

(v) Re-estimation and significance test for all non-zero genetic effects identified by the procedure described earlier in the text using ordinary least-square method.

In implementation of the forward LASSO method, it needs to be specially noticed that genetic effects re-estimated by ordinary least-square method are biased upward because of high variable selection of forward LASSO. The permutation (or bootstrap) should be a good choice for the bias correction in step (v). Also, if too many SNPs are analyzed, as used in real data analysis, each step can be further divided into multiple stages by each non-zero main effect or interaction effect shrunk at last step.

Generalized linear genetic model for discrete traits

In genetic analyses of plants and animals, binary traits and categorical traits are commonly seen, which follow binomial and multinomial distributions, respectively. As generalizations of binary traits and categorical traits, binomial traits and multinomial traits are also frequently observed, which are defined as the proportions of the number of events happened in a number of trials. Moreover, Poisson trait is another example of discrete traits with measurements being counts in a given temporal or spatial interval. The distributions of the aforementioned phenotypic traits belong to an exponential family; thus, the generalized linear model is used to mapping QTLs responsible for the discrete traits.

A generalized linear model [40, 41] consists of three components: phenotype with a probability distribution from the exponential family, the linear predictor and the link function [40]. Similar to the model for continuous traits (1), linear predictors in this scenario can be integrated by

$$\begin{aligned}\eta_i =& \varphi + \sum_{j=1}^{m} x_{ij}\alpha_j + \sum_{j=1}^{m-1}\sum_{j'=j+1}^{m} x_{ij}x_{ij'}\delta_{jj'} \\ &+ \sum_{j=1}^{m-2}\sum_{j'=j+1}^{m-1}\sum_{j''=j'+1}^{m} x_{ij}x_{ij'}x_{ij''}\delta_{jj'j''}\end{aligned} \tag{3}$$

The link function provides the relationship between all linear predictors and the mean of the exponential distribution family, which is denoted by

$$\eta_i = g(\mu_i) \text{ or } \mu_i = g^{-1}(\eta_i) \text{ for } i = 1, 2, \cdots, n \tag{4}$$

where g is the link function and g^{-1} is its inverse (mean function). It should be noted that the link function is differentiated in distribution type of discrete traits. Moreover, the variance of discrete phenotype $V(y_i)$ can be derived for each distribution of discrete trait, which is useful for estimating model parameters as described later in the text.

The re-weighted least-square method by Wedderburn [41] is used to estimating the parameters in generalized linear model. By defining

$$D_i = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \tag{5}$$

$$\xi_i = \eta_i + D_i^{-1}(y_i - \mu_i) \tag{6}$$

$$w_i = \frac{D_i^2}{V^{-1}(y_i)} \tag{7}$$

the quadratic approximation of the log-likelihood function is obtained by Taylor expansions as

$$\sum_{i=1}^{n} w_i(\xi_i - \eta_i)^2 \tag{8}$$

In the case when the number of genetic effects is larger than sample size, the LASSO with a coordinate descent step [35, 36] can efficiently estimate few non-zero genetic effects, by minimizing

$$\begin{aligned}&\sum_{i=1}^{n} w_i(\xi_i - \eta_i)^2 \\ &+ \lambda\left(\sum_{j=1}^{m}|\alpha_j| + \sum_{j=1}^{m-1}\sum_{j'=j+1}^{m}|\delta_{jj'}| + \sum_{j=1}^{m-2}\sum_{j'=j+1}^{m-1}\sum_{j''=j'+1}^{m}|\delta_{jj'j''}|\right)\end{aligned} \tag{9}$$

with λ being as in Equation (2).

As w_i is the function of the estimated parameters, the iteration is required for shrinking most of genetic effects to be zeros, as described previously [35]. Based on the so-called iteratively re-weighted LASSO for generalized linear model, the forward LASSO described earlier in the text can be used to analyze the huge number of high-order interactions in GWAS for discrete traits.

Simulation study

The simulation is conducted to evaluate the forward LASSO method proposed here (forward for short) to the LASSO for full model (1) (full for short). For simplification of simulation, 70 SNPs with equal allele frequencies are simulated. The indicator variable x_{ij} is derived from z_{ij} following a standard multivariate normal distribution with constant correlations of 0.1 according to

$$x_{ij} = \begin{cases} 1 & z_{ij} > 0.675 \\ 0 & -0.675 \le z_{ij} \le 0.675 \\ -1 & z_{ij} < -0.675 \end{cases}$$

Two main effects, three two-order interactions, two third-order interactions and one fourth-order interaction are simulated on the SNPs 2, 11, 23, 38, 50, 60 and 69, where only the 23th and 60th SNPs have the main genetic effects. We assign genetic effects of simulated QTLs by relative heritabilities (Table 2), so that given residual variance of 1.0, in total these QTLs explain 29.4% of phenotypic variation, and the heritability of single QTLs for analyzed traits range from 2.0 to 5.4%. Then we simulate phenotypic value for the quantitative trait from a normal distribution with the expectation η_i without population mean and residual variance. For discrete phenotypic trait, binary phenotypic value is defined as 1 if the normally distributed phenotypic value is positive and as 0 otherwise.

A total of 500 replicated simulations are used to estimate QTL genetic effects and access the statistical power of QTL detections. Each simulated dataset is analyzed by two LASSO-based approaches ('forward' LASSO approach or 'full' model LASSO approach), MDR method [6] and INTERSNP method [42]. Note that, as a non-parametric method, MDR method can only report the statistical power, but not estimate the estimated QTL genetic effects. INTERSNP method has ability to analyze only up to the three-way interactions. Statistical power of QTL detection is evaluated at each locus and is defined as the proportion of all simulations where the test statistic exceeds its critical value. We set the significance level at 5%. In addition, false-positive rate is evaluated with the 500 replicated simulations under the null model without genetic effects. Note that we report means and standard deviations of QTL genetic effects based on simulations whose corresponding estimated genetic effects are significant.

The statistical powers of the four competing methods and estimated genetic effects of two LASSO-based approaches and INTERSNP method are presented in Tables 1 and 2, respectively. In general, two mapping methods exhibit the following similar statistical behaviors in analyzing both normally distributed and binary phenotypic traits. (i) statistical power of QTL detection and the precision of parameter estimation increase as the QTL heritability increases; (ii) statistical power of QTL detection is higher, and false-positive rate is lower in the same simulation scenario; (iii) large sample size is beneficial to identify QTLs; and (iv) the higher the order of interactions, the more difficult the QTL detections. Clearly, the four competing methods have comparable statistical power. The two LASSO-based approaches, however, are able to well estimate QTL genetic effects because of the incapability of the other method to handle an over-saturated ultrahigh-dimensional model. In comparison, forward LASSO method does more accurately than the LASSO method for full model. In terms of the false-positive rate, it is <10% for the four competing methods in all simulation scenarios, although the false-positive rate of forward LASSO and INTERSNP methods is slightly higher than those of other methods.

Our proposed method is computationally efficient as well. In simulated data sets, each with 70 simulated SNPs, up to four-way interactions and 2000 subjects, we compare computational time for three methods in estimating 974 121 parameters. On an Intel core 4 PC with a 3.4 GHz processor and 16.00 GB random access memory, forward LASSO method and LASSO method for full model take 10.3 s and 15.1 min on average, respectively, whereas both MDR and INTERSNP methods run ~10 min.

Real data analysis

Experimental animals are originated from Ulgai, Xilingol league, Inner Mongolia of China, which consist of 1058 young Simmental bulls being born in 2008–11. After weaning, the cattle were moved to Beijing Jinweifuren cattle farm and were fattened under the same feeding and management. Each individual was timely observed for growth and development traits until slaughtered from 16 to 18 months old. During the period of slaughter, carcass traits and meat quality traits were measured according to Institutional Meat Purchase Specifications for fresh beef guidelines. The blood samples were collected along with the regular quarantine inspection of the farms without the need of ethical approval. DNAs were extracted from these blood samples using the routine procedures. The Illumina BovineHD BeadChip was adopted for quantifying and genotyping DNAs.

Before statistical analysis, we pre-process the SNP data and remove those SNPs whose call rates are <90%, minor allele frequencies are <3%, genotype appearances are less than five individuals or departure from Hardy Weinberg Equilibrium is severe (with $<10^{-6}$ probability). Moreover, individuals with >10% missing genotypes or >2% Mendelian error rate for SNP genotypes are excluded. Finally, 986

Table 1: Statistical powers of QTL detection (%) and false-positive rates (FPR, %) obtained with the four mapping methods for the simulated data sets with 1000 and 2000 sample sizes

Trait	Sample size	Method	Q_1	Q_2	$Q_1 \times Q_2$	$Q_2 \times Q_3$	$Q_1 \times Q_4$	$Q_2 \times Q_3 \times Q_5$	$Q_1 \times Q_4 \times Q_6$	$Q_2 \times Q_3 \times Q_5 \times Q_7$	FPR
Normal	1000	Forward	43.4	50.0	47.4	39.2	23.4	25.4	25.8	14.2	8.6
		Full	48.4	53.6	52.2	45.4	21.6	21.8	26.8	16.4	8.0
		MDR	46.6	53.2	46.2	35.2	17.8	20.4	25.4	14.8	8.1
		INTERSNP	47.2	55.0	50.4	33.6	18.2	19.8	22.4	–	8.8
	2000	Forward	93.4	93.6	90.8	92.6	83.8	79.2	81.6	64.0	5.8
		Full	99.6	99.4	100.0	99.0	80.2	78.6	86.2	68.4	5.4
		MDR	100	100	96.4	98.2	81.4	80.2	85.2	69.4	5.6
		INTERSNP	99.4	100	98.2	93.6	84.6	81.2	84.4	–	5.9
Binary	1000	Forward	90.2	75.0	44.0	46.8	85.2	30.8	83.2	31.2	7.6
		Full	90.6	74.2	49.8	49.6	83.6	31.6	83.6	31.4	7.2
		MDR	86.0	73.4	44.8	46.8	81.0	30.0	83.4	28.4	7.4
		INTERSNP	85.4	78.6	42.6	47.4	80.6	32.2	82.8	–	7.9
	2000	Forward	100	98.8	88.6	92.4	100	88.8	89.8	73.4	4.0
		Full	99.4	99.2	89.0	93.0	99.0	89.6	92.4	78.2	3.6
		MDR	96.8	100	80.8	93.4	100	86.8	90.0	74.2	3.8
		INTERSNP	98.4	98.8	90.2	91.6	96.8	84.6	86.2	–	4.2

Table 2: Mean estimates and standard deviations (in parentheses) of QTL effects obtained with the three mapping methods for the simulated data sets with 1000 and 2000 sample sizes

Trait	Sample size	Method	Q_1	Q_2	$Q_1 \times Q_2$	$Q_2 \times Q_3$	$Q_1 \times Q_4$	$Q_2 \times Q_3 \times Q_5$	$Q_1 \times Q_4 \times Q_6$	$Q_2 \times Q_3 \times Q_5 \times Q_7$
		True effect	0.24	−0.26	−0.35	0.32	0.25	0.35	−0.38	0.39
Normal	1000	Forward	0.25(0.05)	−0.27(0.05)	−0.35(0.10)	0.33(0.07)	0.28(0.06)	0.38(0.07)	−0.38(0.11)	0.41(0.07)
		Full	0.26(0.04)	−0.28(0.05)	−0.37(0.08)	0.35(0.07)	0.30(0.06)	0.41(0.11)	−0.43(0.13)	0.46(0.09)
		INTERSNP	0.32(0.11)	−0.23(0.08)	−0.45(0.10)	0.39(0.08)	0.22(0.07)	0.42(0.08)	−0.46(0.15)	–
	2000	Forward	0.24(0.04)	−0.26(0.04)	−0.35(0.07)	0.32(0.06)	0.25(0.05)	0.35(0.06)	−0.37(0.08)	0.40(0.06)
		Full	0.24(0.04)	−0.26(0.04)	−0.35(0.06)	0.32(0.06)	0.26(0.05)	0.37(0.06)	−0.38(0.09)	0.40(0.11)
		INTERSNP	0.27(0.12)	−0.29(0.08)	−0.43(0.11)	0.35(0.08)	0.24(0.09)	0.37(0.05)	−0.39(0.09)	–
Binary	1000	Forward	0.27(0.06)	−0.28(0.09)	−0.33(0.17)	0.35(0.10)	0.30(0.09)	0.41(0.17)	−0.43(0.19)	0.42(0.04)
		Full	0.28(0.08)	−0.28(0.10)	−0.32(0.23)	0.35(0.16)	0.31(0.17)	0.26(0.42)	−0.31(0.38)	0.45(0.13)
		INTERSNP	0.27(0.08)	−0.30(0.10)	−0.40(0.12)	0.43(0.10)	0.43(0.15)	0.47(0.12)	0.56(0.21)	–
	2000	Forward	0.25(0.05)	−0.26(0.08)	−0.33(0.14)	0.33(0.07)	0.27(0.07)	0.37(0.09)	−0.37(0.16)	0.42(0.09)
		Full	0.25(0.06)	−0.26(0.08)	−0.31(0.20)	0.32(0.11)	0.28(0.10)	0.27(0.24)	−0.24(0.28)	0.39(0.19)
		INTERSNP	0.28(0.09)	−0.31(0.09)	−0.29(0.11)	0.35(0.08)	0.41(0.11)	0.39(0.14)	−0.45(0.04)	–

individuals and 631 396 SNPs are remained for GWAS.

Of 40 carcass traits and meat quality traits, birth weight and marbling traits are taken as examples of quantitative traits and discrete traits, respectively, to illustrate the merit of forward LASSO method. To reduce the influence of multiple uneven categories on the power of QTL detection, the marbling trait is simplified as binary traits including only two categories. Being a quantitative trait, birth weight is analyzed based on the LASSO for linear model, whereas marbling is analyzed based on the LASSO for GLM with probit link function. Systematic environment factors, including measuring year and slaughtering month old are included in the genetic model, and population stratification is taken account as well. In the GWAS, SNPs involved in up to fourth-order interactions with at least one SNP of significant main effect are searched.

Both the LASSO method for full model and MDR method fail to work because of the astronomic number of SNP combinations. Only up to third-order interactions for the two traits analyzed are identified by using forward LASSO method.

Table 3: Significant main effect SNPs and interactions among SNPs for birth weight in beef cattle

Type	QTL	SNP	Chr.	Position	Nearest gene		−log(p)	Effect	Heritability (%)
					Name	Distance (bp)			
Main effect	Q_1	BovineHD2200015010	22	52799535	BT.18085	Within	5.41	−0.94	1.23
	Q_2	BovineHD1500001499	15	6030018	BT.18504	Within	5.10	−2.52	8.87
	Q_3	BovineHD1400001847	14	6869573	KHDRBS3	571 473	3.61	−0.73	0.74
	Q_4	BovineHD2400001759	24	6565808	SOCS6	737 239	3.56	0.68	0.65
	Q_5	BovineHD1000024843	10	87201079	FLVCR2	756	3.53	−1.64	3.76
Two-way interaction	Q_6	BovineHD0300028625 × Q_5	3 × Q_5	99613961 × Q_5	TALI	Within	3.12	0.81	0.92
	Q_7	BovineHD0300011290 × Q_3	3 × Q_3	36196363 × Q_3	MGC139448	Within	2.72	−0.24	0.08
	Q_8	BovineHD1700020147 × Q_1	17 × Q_1	69028276 × Q_1	MNI	403 401	3.98	−0.56	0.44
	Q_9	BovineHD1600008117 × Q_4	16 × Q_4	28905372 × Q_4	DNAHI4	124 253	2.03	−0.38	0.20
Three-way interaction	Q_{10}	BovineHD2000000012 × Q_8	20 × Q_8	85671 × Q_8	ENSBTAG00000000617	16 726	4.93	0.18	0.05
	Q_{11}	BovineHD2600009158 × Q_9	26 × Q_9	33994296 × Q_9	TCFL2	51 628	2.74	0.03	0.00

In implementing each step of forward LASSO method, the interactions are formed separately with each SNP with non-zero main effect or interaction of previous order and all genome-wide SNPs. Further, we have randomly partitioned the real data set into two subsets and applied the proposed method to analyze these two subsets separately. The results from five repeated partitions show that there is smaller difference in the selected predictors between the two subsets, which demonstrates the stability of the mapping results.

Significant main effect SNPs and epistases of no more than three orders are summarized in Table 3 for birth weight and in Table 4 for marbling, respectively. Coincidentally, 11 significant main and interaction effects are detected for both traits and two three-way interactions are detected for both traits as well. However, the marble trait is associated with more interactions. All detected SNPs explain 17% of phenotypic variation for birth weight and 47% for marbling. Especially, the contribution of interactions among SNPs to phenotypic variation is large for marbling. Two- and three-way interactions contribute 23.75 and 14.03% of total heritability, respectively. On the other hand, all detectable SNPs for birth weight have very low heritabilities, except for the SNP with 8.87% heritability on the 15th chromosome. Biologically, these detected SNPs locate within or near genes associated with growth and development in beef cattle (Tables 3 and 4). They may jointly regulate formation and development of birth weight and marbling with these genes.

For a comparison with INTERSNP method, we select top 1000 hits from single marker analysis, and then check two-way interactions. Generally, INTERSNP method can detect more significant main-effect and interactive SNPs than forward LASSO method for the two analyzed traits. Only few same main-effect SNPs are located by using the two mapping methods, which are BovineHD2200015010 on chromosome 22 and BovineHD1400001847 on chromosome 14 for birth weight as well as BovineHD1300023732 on chromosome 13 for marbling. Because of heavy computational burden, INTERSNP method is also hard to handle more than two-way interactions with too many the selected SNPs from single-marker analysis.

DISCUSSION

Different from forward regression where only one independent variable is included in each step of regressions, forward LASSO is capable of handling hundreds of thousands of genetic effects at each stage. The number of the genetic effects estimated at each stage depends on the performance of the LASSO algorithm incorporated as well as the computational power of processors. Motivated by weak heredity principle of interaction effects, our estimation of n effects focus on interactions between each non-zero n effects of a lower order and all genome-wide SNPs, so that at least one non-zero main effect SNP is included in the interaction effect detection. Note that the significance test for non-zero genetic effects is carried out at the final stage of forward LASSO method, which can avoid the impact of non-zero genetic effects included in a later stage.

Table 4: Significant main effect SNPs and interactions among SNPs for marbling in beef cattle

Type	QTL	SNP	Chr.	Position	Nearest gene		−log(p)	Effect	Heritability (%)
					Name	Distance (bp)			
Main effect	Q_1	BovineHD0400003329	4	11041270	TFPI2	1551	3.44	−0.18	0.86
	Q_2	BovineHD0800009637	8	31873954	C6KEI7	147 589	3.09	−0.35	3.24
	Q_3	BovineHD1300023732	13	81868064	ZNF217	5060	3.06	−0.31	2.54
	Q_4	BovineHD1400004974	14	17496444	FERIL6	Within	3.69	−0.32	2.71
Two-way interaction	Q_5	BovineHD0900018633 × Q_1	9 × Q_1	67365446 × Q_1	PTPRK	Within	3.02	−0.33	2.88
	Q_6	BovineHD3000040536 × Q_2	30 × Q_2	140359650 × Q_2	BT.20005	43 739	6.14	0.57	8.59
	Q_7	BovineHD3000044113 × Q_2	30 × Q_2	148534728 × Q_2	ENSBTAG00000025951	Within	13.88	0.19	0.95
	Q_8	BovineHD0600030851 × Q_3	6 × Q_2	109607469 × Q_3	LOC510550	7140	3.55	0.54	7.71
	Q_9	BovineHD0300030768 × Q_4	3 × Q_4	107020441 × Q_4	EIBF39	11 733	3.10	0.37	3.62
Three-way interaction	Q_{10}	BovineHD2700011429 × Q_5	27 × Q_5	39417878 × Q_5	LRC3B	278 014	4.70	−0.51	6.88
	Q_{11}	BovineHD0100012484 × Q_7	1 × Q_7	43781308 × Q_7	MCATL	21 228	2.83	0.52	7.15

In forward LASSO, the quantitative traits and discrete traits are analyzed based on the LASSO for linear model and that for GLM, respectively. Although the LASSO for GLM is also appropriate for quantitative traits, it has lower computing efficiency than that for linear model. Forward LASSO for detecting high-order interactions in GWAS proposed in this study has been coded in R language, which can handle normal, binary, binomial, ordinal, multinomial and Poisson traits. The program is freely available on request from authors.

Statistically, detecting high-order interactions among a large number of SNPs in GWAS is not easy. On the one hand, pairwise combinatorial search and test for significant genetic effects are not feasible. On another hand, when estimating higher-order n effects, statistical models expect a larger than usual sample size to achieve desirable levels of statistical power and false-positive rate. Once a certain genotype combination among some SNPs is missing in experimental population or does not exist in nature, corresponding type of interaction effect can not be identified. It is found in simulations that too low frequency of genotype combination among SNPs results in false and abnormal estimate of the related interaction effect. Moreover, no interaction among more than three SNPs is found in real data analysis.

In beef cattle, for instance, Korean Hanwoo cattle [43], Korean beef cattle [44] and Australian taurine and indicine cattle [45] GWAS have been carried out for detecting genetic variations associated with beef carcass traits and meat quantity traits. Many significant main effect SNPs were identified using the simple linear regression and stepwise regression procedures. Barendse [46] have identified significant interaction effect between SNPs at CAPN1 and CAST for beef tenderness in both taurine- and zebu-derived breeds and a larger additive × dominance component of interaction effect than additive × additive and dominance × dominance components were observed. By selection designed to increase the frequencies of the minor alleles for two SNP markers in CSN151 and TG, adjusted fat thickness showed a dominance association with the TG SNP and an additive CSN1S1 × additive TG association [47]. Most currently, interaction effect analysis by GWAS has found that the 11 SNP pairs were significantly associated with carcass traits [48], although higher-order interaction effect has not been reported in beef cattle. In our real data analysis, forward LASSO method has identified third-order interaction among SNPs, which provides more heritabilities for the analyzed traits unexplained by current GWAS for main effect and two-way interaction Moreover, <50 000 SNPs were used in literature of GWAS analyses of beef cattle, but our study considers >600 000 SNPs. This will provide more insight into exploring high-order interactions for carcass traits in beef cattle.

Key points

- By evenly partitioning the interaction effect model into many stages by the number of main effects, forward LASSO method is proposed to analyze high-order interactions in GWAS.
- Our proposed method can efficiently identify high-order interactions with at least one significant main effect for both quantitative traits and discrete traits.

- Extensive simulations demonstrate that forward LASSO method could be a promising alternative of the LASSO method for full model.
- GWAS for high-order interactions of birth weight and marbling in beef cattle exhibits the use of our proposed method.

FUNDING

The study was supported by the 12th 'Five-Year' National Science and Technology Support Project (2011BAD28B04) and basic research fund program 2010jc-2 of state-level public welfare scientific research institutions, Institute of Animal Sciences, Chinese Academy of Agricultural Science. The National Natural Science Foundations of China (30972077 and 31172190) supported the research as well.

References

1. Weng G, Bhalla US, Iyengar R. Complexity in biological signaling systems. *Science* 1999;**284**:92–6.
2. Hlavacek WS, Faeder JR. The complexity of cell signaling and the need for a new mechanics. *Sci Signal* 2009;**2**:46.
3. McMullen MD, Byrne PF, Snook ME, *et al*. Quantitative trait loci and metabolic pathways. *Proc Natl Acad Sci USA* 1998;**95**:1996–2000.
4. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 2008;**9**:855–67.
5. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–404.
6. Ritchie MD, Hahn LW, Roodi N, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;**69**:138–47.
7. Martin MP, Gao X, Lee JH, *et al*. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet* 2002;**31**:429–34.
8. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;**6**:95–108.
9. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;**37**:413–17.
10. Wang WYS, Barratt BJ, Clayton DG, *et al*. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;**6**:109–18.
11. Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
12. Wan X, Yang C, Yang Q, *et al*. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* 2010;**26**:30–7.
13. Gabutero E, Moore C, Mallal S, *et al*. Interaction between allelic variation in IL12B and CCR5 affects the development of AIDS: IL12B/CCR5 interaction and HIV/AIDS. *AIDS* 2007;**21**:65–9.
14. Kao CH, Zeng ZB. Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* 2002;**160**:1243–61.
15. Yang RC. Epistasis of quantitative trait loci under different gene action models. *Genetics* 2004;**167**:1493–505.
16. Zeng Z, Wang T, Zou W. Modeling quantitative trait loci and interpretation of models. *Genetics* 2005;**169**:1711–25.
17. Mao Y, London NR, Ma L, *et al*. Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model. *Physiol Genomics* 2007;**28**:46–52.
18. Álvarez-Castro JM, Carlborg Ö. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 2007;**176**:1151–67.
19. Álvarez-Castro JM, Le Rouzic A, Carlborg Ö. How to perform meaningful estimates of genetic effects. *PLoS Genet* 2008;**4**:e1000062.
20. Zee RY, Hoh J, Cheng S, *et al*. Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. *Pharmacogenomics J* 2002;**2**:197–201.
21. Williams SM, Ritchie MD, Phillips JAIII, *et al*. Multilocus analysis of hypertension: a hierarchical approach. *Hum Hered* 2004;**57**:28–38.
22. Tsai CT, Lai LP, Lin JL, *et al*. Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation* 2004;**109**:1640–6.
23. Cho YM, Ritchie MD, Moore JH, *et al*. Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia* 2004;**47**:549–54.
24. Nunkesser R, Bernholt T, Schwender H, *et al*. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics* 2007;**23**:3280–8.
25. Nelson MR, Kardia SLR, Ferrell RE, *et al*. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001;**11**:458–70.
26. Lou XY, Chen GB, Yan L, *et al*. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am J Hum Genet* 2007;**80**:1125–37.
27. Cockerham C. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 1954;**39**:859–82.
28. Kempthorne O. The correlation between relatives in a random mating population. *Proc Royal Soc B* 1954;**143**: 103–13.
29. Ruczinski I, Kooperberg C, LeBlanc ML. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. *J Multivar Anal* 2004;**90**: 178–95.
30. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics* 2008;**9**:30–50.
31. Ives AR, Garland T. Phylogenetic logistic regression for binary dependent variables. *Syst Biol* 2010;**59**:9–26.
32. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 2005; **28**:157–70.
33. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* 2007;**39**:1167–73.
34. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B Stat Methodol* 2011;**73**:273–82.

35. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol* 2006;**68**:49–67.

36. Friedman J, Hastie T, Tibshirani. R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;**33**:1–22.

37. Wu TT, Chen YF, Hastie T, *et al*. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 2009;**25**:714–21.

38. Li J, Das K, Fu G, *et al*. The Bayesian lasso for genome-wide association studies. *Bioinformatics* 2011;**27**:516–23.

39. Chipman H. Bayesian variable selection with related predictors. *Can J Stat* 1996;**24**:17–36.

40. McCullagh P, Nelder JA. *Generalized Linear Models*. New York: Chapman and Hall/CRC, 1989.

41. Wedderburn RWM. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* 1974;**61**:439–47.

42. Herold C, Steffens M, Brockschmidt FF, *et al*. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics* 2009;**25**:3275–81.

43. Alam M, Lee YM, Park BL, *et al*. A whole genome association study to detect single nucleotide polymorphisms for carcass traits in hanwoo populations. *Asian Australas J Anim Sci* 2010;**23**:417–24.

44. Kim Y, Ryu J, Woo J, *et al*. Genome-wide association study reveals five nucleotide sequence variants for carcass traits in beef cattle. *Anim Genet* 2011;**42**:361–5.

45. Bolormaa S, Neto LRP, Zhang YD, *et al*. A genome-wide association study of meat and carcass traits in Australian cattle. *J Anim Sci* 2011;**89**:2297–309.

46. Barendse W, Harrison BE, Hawken RJ, *et al*. Epistasis between calpain 1 and its inhibitor calpastatin within breeds of cattle. *Genetics* 2007;**176**:2601–10.

47. Bennett GL, Shackelford SD, Wheeler TL, *et al*. Selection for genetic markers in beef cattle reveals complex associations of thyroglobulin and casein1-s1 with carcass and meat traits. *J Anim Sci* 2013;**91**:565–71.

48. Ghaffar MAA, Samee ASMA, Shetaewi MA, *et al*. Genome-wide analysis of single-locus and epistatic SNP effects on carcass traits in angus beef cattle. In: *Plant & Animal Genomes XIX Conference*. San Diego, USA, 2011;543.