Genet. Res., Camb. (2015), vol. 97, e9. © Cambridge University Press 2015
doi:10.1017/S0016672315000038

Genome-wide interaction analysis of quantitative traits in outbred mice

WEIJUN MA[1], CHAOFENG YUAN[1], HAIDONG LIU[1], WEI ZHENG[2] AND YING ZHOU[1]*
[1]*School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China*
[2]*School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China*

(*Received 16 September 2014; revised 11 February 2015; accepted 12 February 2015*)

Summary

With a large number of quantitative trait loci being identified in genome-wide association studies, researchers have become more interested in detecting interactions among genes or single nucleotide polymorphisms (SNPs). In this research, we carried out a two-stage model selection procedure to detect interacting gene pairs or SNP pairs associated with four important traits of outbred mice, including glucose, high-density lipoprotein cholesterol, diastolic blood pressure and triglyceride. In the first stage, a variance heterogeneity test was used to screen for candidate SNPs. In the second stage, the Lasso method and single pair analysis were used to select two-way interactions. Moreover, the shared Gene Ontology information about the selected interacting gene pairs was considered to study the interactions auxiliarily. Based on this method, we not only replicated the identification of important SNPs associated with each trait of outbred mice, but also found some SNP pairs and gene pairs with significant interaction effects on each trait. Simulation studies were also conducted to evaluate the performance of the two-stage method in different situations.

1. Introduction

There are various meanings for the term interaction, here it is defined as a joint SNP–SNP or gene–gene effect that can not be readily explained by their separate marginal effect (Kahn, 1983). As is known, interaction between genes or gene and environment is one of the main factors that contribute to a trait (i.e. disease). Many strategies and methods have been applied to genome-wide interaction studies (GWIS), such as regression, machine learning, Bayesian method and SNP filtering etc. (Mckinney *et al.*, 2006; Zhang & Liu, 2007; Bai *et al.*, 2012). Balancing computational load and statistical power, SNP filtering or screening that select candidate SNPs based on *a priori* information is a promising method for genome-wide data (Herold *et al.*, 2009). Sources of the *a priori* information include statistical evidence (single marker association at a moderate level), genetic relevance (genomic location) and biologic relevance (SNP function class and pathway information).

Kooperberg & Leblanc (2008) proposed a two-stage analysis, in which they only test for interactions between SNPs that show some marginal effects. Some SNPs or genes may not show strong marginal associations when they affect disease risk through interactions with other SNPs or genes. As a result, these genes may not be identified by the single marker association screening method in GWIS. Considering both computational load and statistical power, Wu *et al.* (2009) conducted an exhaustive two-dimensional search in the first stage, which detected joint effects that may fail to emerge from single maker analysis, but may have an inflated type I error. Paré *et al.* (2010) demonstrated that, under plausible scenarios of genetic interaction, the variances of a quantitative trait are expected to differ among the three possible genotypes of a biallele SNP. Thus a variance heterogeneity test can be used to screen for potentially interacting SNPs. Then an exhaustive two-dimensional scan was conducted among the candidates identified in the first stage. Paré's screening method has the advantage that the interacting covariants need not be known or measured for a SNP to be prioritized and independence between the two steps under the null hypothesis of no interaction can guarantee correct type I error.

* Corresponding author: E-mail: yzhou@aliyun.com

In this study, we carried out genome-wide interaction association studies using 288 mice from a commercially available outbred stock with four traits and 44 428 SNPs (Zhang *et al.*, 2012). The four traits including glucose (GLU), high-density lipoprotein cholesterol (HDL), diastolic blood pressure (DBP) and triglyceride (TG) are strongly associated with some complex diseases, such as diabetes, cardiovascular disease and adiposity. Detecting interacting SNP or gene pairs is of great significance for learning about the mechanisms of some complex diseases.

A two-stage model selection procedure was used to detect interacting SNP pairs or gene pairs for each trait. In the first stage, based on Paré's idea, variance heterogeneity tests were used to screen for potentially interacting SNPs. As multiple genetic variants are expected to jointly affect a complex trait, the Lasso method (Tibshirani, 1996) and single pair analysis were both used to select significant interacting pairs among the candidates in the second stage. Several simulation scenarios were designed to evaluate the performance of the two-stage procedure, where Hardy-Weinberg equilibrium (HWE) and linkage equilibrium may be not satisfied.

2. Materials

The raw data set includes 288 NMRI mice with eight traits and 581 672 SNPs. Zhang *et al.* (2012) carried out some basic work on the data by excluding those SNPs whose HWE $\chi^2 \geq 20$, minor allele frequencies <2% and missing values >40%, and collapsing identical SNPs within 2Mb intervals, resulting in a total of 44 428 unique SNPs in the final data.

These traits include systolic blood pressure (SBP), DBP, mean arterial pressure (MAP), GLU, TG, cholesterol (CHL), HDL and urinary albumin-to-creatinine ratio (ACR). All traits except ACR were approximately normally distributed. Since the lipid traits (HDL and CHL) and blood pressure traits (SBP, DBP and MAP) were highly correlated among themselves ($r > 0{\cdot}97$), here we only report our analysis results of DBP, GLU, TG and HDL as representative of this group of traits.

3. Method

A two-stage method based on Paré's idea is introduced in this section. In the first stage, equality of the conditional variances under three possible genotypes for each SNP is tested, which is equivalent to test interaction between this SNP and another covariant. Then the Lasso method and single pair analysis are both used to select SNP pairs among the candidates chosen before. The details of the two-stage approach are given in the following sections.

(i) *Stage one: variance heterogeneity test*

The objective of the first stage is to select SNPs that are likely to have interaction effects. In detail, if we want to test whether a SNP (G represents its genotype) has an interaction effect on a quantitative trait y with a covariant C (another SNP or environment factor), the follow linear model is considered:

$$y = \beta_0 + \beta_1 G + \beta_2 C + \beta_3 GC + \varepsilon,$$

then $Var(y|G = g) = (\beta_2 + \beta_3 g)^2 Var(C|G = g) + \sigma^2$. Under the assumption of independence between G and C, we further have:

$$Var(y|G = g) = (\beta_2 + \beta_3 g)^2 Var(C) + \sigma^2.$$

It is clear that $Var(y|G = 0) = Var(y|G = 0{\cdot}5) = Var(y|G = 1)$ if and only if $\beta_3 = 0$. Thus, testing the interaction effect between G and C is equivalent to testing the equality of the conditional variances (i.e. $\sigma_0^2 = \sigma_{0{\cdot}5}^2 = \sigma_1^2$, where $\sigma_i^2 = Var(y|G = i)$). The biggest character of this approach is that we do not need to know the covariant's information when judging whether one factor has an interaction effect with another covariant or not.

The Levene's test (Olkin *et al.*, 1960) was applied to verify the equality of variances under different genotypes for each of 44 428 SNPs in the data set we considered. Suppose the sample size N can be divided into K subgroups and let N_i denote the sample size of the ith group. The Levene's test statistic is defined as:

$$W = \frac{(N-K)\sum_i N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{(K-1)\sum_i \sum_j (Z_{ij} - Z_{i\cdot})^2},$$

where $Z_{ij} = |Y_{ij} - Y_{i\cdot}|$, and Y_{ij} is the trait value for the jth individual of the ith group; $Y_{i\cdot}$ is the mean of the ith subgroup; $Z_{i\cdot}$ and $Z_{\cdot\cdot}$ are the subgroup mean and overall mean of Z_{ij}, respectively. Under the null hypothesis of variance homogeneity, the Levene's statistic follows a F distribution with K–1 and N–K degrees of freedom. In this paper we have $K = 3$ and $N = 288$.

(ii) *Stage two: interaction test*

The purpose of the second stage is to further identify interacting pairs among the candidate SNPs selected from the first stage. When building statistical models, the marginal association loci reported by Zhang *et al.* (2012) are included. Single pair scan and multiple marker analysis with the Lasso model are both used here. The details of the two linear models are as follows:

$$y = \beta_0 + \sum_{k=1}^{5} \beta_k SNP_k + \beta_{ij} X_i X_j + \varepsilon, \quad (1)$$

https://doi.org/10.1017/S0016672315000038 Published online by Cambridge University Press

$$y = \beta_0 + \sum_{k=1}^{5} \beta_k SNP_k + \sum_{k=1}^{p} \beta_k^* X_k + \sum_{i<j} \beta_{ij} X_i X_j + \varepsilon,$$
$$s.t. \sum_{k=1}^{5} |\beta_k| + \sum_{k=1}^{p} |\beta_k^*| + \sum_{i<j} |\beta_{ij}| \le \gamma, \quad (2)$$

where SNP_k, $1 \le k \le 5$ represent the five main effect loci identified by Zhang *et al.* (2012), X_k, $1 \le k \le p$ are genotypes of the SNPs selected in the first stage, $X_i X_j$ is product of the candidate pair's genotype values and $\gamma \ge 0$ is a tuning parameter to be determined separately.

Although the single pair scan and multiple marker analysis may detect some different interacting SNP pairs in the second stage, we expect that the results from the analysis of the two strategies can supplement each other, since we do not want to miss any valuable interaction effect.

4. Results

(i) *Results from stage one*

In order to avoid filtering out interesting SNPs by the variance heterogeneity tests, a relaxed type I error or family wise error rate is allowed in stage one. In the real data analysis, significance level $\alpha = 0{\cdot}05$ was chosen for each SNP. Taking trait GLU as an example, there are 2487 SNPs showing significant variance heterogeneity (see the second line of Table 1).

As some SNPs are very close in their locations on the chromosome, their genotypes are very similar among the 288 mice in the data set. For example, only three mouse genotypes are different on the three loci respectively located at 74104950bp, 74187929bp and 74201260bp on chromosome 1. In general, if one of them was significant in the variance heterogeneity test, then all would pass the first stage test. Therefore, the loci we picked up in the first stage would not be independent, and that would also increase the burden of multiple tests. So we combined those SNPs that were within 2Mb and gave the same genotype among more than 97·5% of the 288 mice. There are 779 SNPs left after reorganizing the 2487 SNPs for trait GLU (see the third line of Table 1). The allelic association between two SNP loci was used as a measure of linkage disequilibrium (LD). We observed a sharp decay in LD with increasing physical distance between the 779 SNPs on the same chromosome (see Fig. 1). The numbers of SNPs after reorganizing for each trait are listed in Table 1.

(ii) *Results from stage two*

(a) *Single pair analysis based on model (1)*

Taking trait GLU as an example, 779 SNPs were chosen from stage one and C_{799}^2 SNP pairs were tested with model (1). After calculation, 199 SNP pairs among them have P-values smaller than 1E-4. The mean square error (MSE) reduction by including each interaction term in model (1) was also calculated, respectively. Tables S1–S4 in File 1 of the supplementary material give the annotations of SNP pairs with P-values <1E-4 and significant shared Gene Ontology (GO) information for the four traits. We listed some representative SNP pairs with validated interaction effects in model (1) for each trait in Table 2.

Table 1. *The number of selected SNPs and SNP pairs.*

Trait	GLU	HDL	DBP	TG
Number of selected SNPs[a]	2487	2326	3220	2596
Number of reorganized SNPs[b]	779	900	1063	1001
Number of SNP pairs with p-value <1*E*-4 in Model (1)	199	184	205	10
Number of significant SNP pairs selected in Model (2)	10	3	5	11

[a] The number of significant SNPs in variance heterogeneity tests with significance level $\alpha = 0{\cdot}05$.
[b] The number of SNPs after collapsing those SNPs selected in the first stage with similar genotypes and close positions.

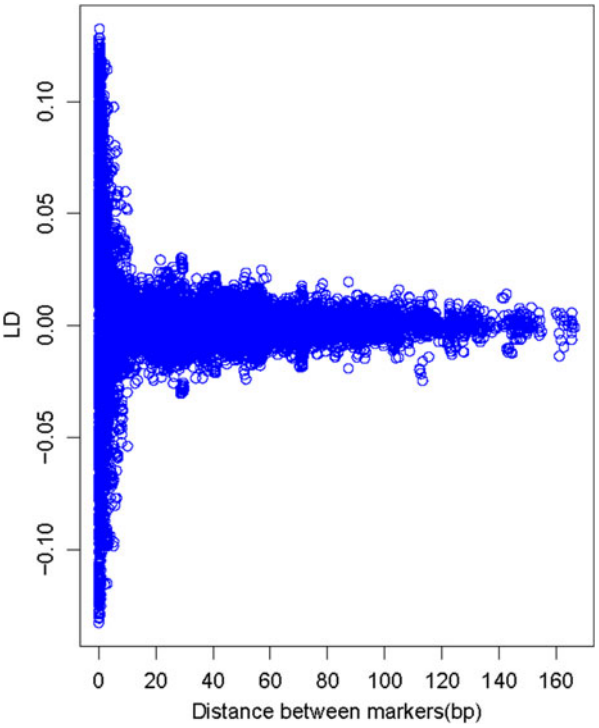

Fig. 1. LD plot for trait GLU. The allelic association between two SNP loci was used as a measure of LD. Each dot represents the LD value of a SNP pair on the same chromosome.

https://doi.org/10.1017/S0016672315000038 Published online by Cambridge University Press

Table 2. *The interacting information detected by the two-stage method via model (1).*

	Locus 1			Locus 2					MSE reduction
Trait	Chr	Position	Gene or nearest gene	Chr	Position	Gene or nearest gene	P-value[a]	N[b]	(%)[c]
GLU	17	87128520	-Epas1[d]	18	78042984	Pstpip2	1·47E-06	7	7·45
	6	147670244	Gm15762 B230104C08Rik	9	118103317	Gm17399-	3·92E-06	21	7·02
	6	147670244	Gm15762-	2	93039320	–	3·53E-06	27	6·89
	6	147670244	Gm15762-	3	75843721	Golim4 Fstl5	3·30E-06	25	6·88
	6	147639739	1700049E15Rik	5	121142226	Dtx1	4·51E-06	4	6·73
	6	147670244	Gm15762-	15	11931111	Sub1 Zfr	4·6E-06	3	5·05
HDL	4	150470889	Camta1	17	32007869	Sik1 Hsf2bp	2·06E-07	7	8·77
	1	164206594	Dnm3	17	72311236	Alk	3·01E-06	10	7·07
	4	150470889	Camta1	1	106123272	–	4·71E-06	6	6·79
	4	150470889	Camta1	1	187715529	Slc30a10 Lyplal1	4·95E-06	2	6·76
	4	150470889	Camta1	2	123671035	Gm13988 Sema6d	5·87E-06	7	6·65
	4	150470889	Camta1	14	74796750	Gm6984 Htr2a	6·08E-06	25	6·63
	5	66555742	Rbm47	13	63998360	0610007P08Rik	6·32E-06	2	6·61
DBP	13	83963970	C130071C03Rik-	3	34764407	Sox2ot-	3·72E-08	6	9·26
	13	84273654	C130071C03Rik-	9	115009565	Osbpl10	1·72E-07	24	8·85
	13	84273654	C130071C03Rik-	13	60043341	Zcchc6 Gas1	4·29E-07	3	8·27
	13	84273654	C130071C03Rik-	7	77717179	B130024G19Rik-	3·37E-06	4	6·96
	9	115029912	Osbpl10	18	65601942	Malt1	5·17E-06	6	6·69
	3	109582676	Vav3 Ntng1	16	41310194	-Lsamp	9·57E-06	4	6·31
TG	1	135851986	Prelp Fmod	14	17179014	Ngly1 Top2b	6·37E-07	4	7·95
	9	116453569	Tgfbr2 Rbms3	13	104033960	Mast4	6·72E-07	2	7·92
	9	116453569	Tgfbr2 Rbms3	12	61248230	Ociad2 Cwh43	2·49E-06	7	7·09
	5	73732287	Ociad2 Cwh43	17	32035365	Sik1 Hsf2bp	4·43E-06	24	6·72
	5	73732287	Ociad2 Cwh43	11	99397875	Krt39 Krt40	6·68E-06	6	6·46

[a] P-value for model (1) containing five main effects loci and each single SNP pair.
[b] N: Number of SNP pairs with P-value <5E-4 near the SNP pairs listed in each row.
[c] The MSE reduction by including each interaction term in model (1).
[d] -Epas1: The detected SNP is located within the intergenic region of a snRNA or miRNA and MGI gene Epas1.

(b) *Multiple marker analysis based on model (2)*

The goal of model (2) is to jointly select from the candidates. In total, 1784 variables were chosen as candidates for trait GLU, including five marginal association loci reported by Zhang *et al.* (2012), 779 unique SNPs selected in stage 1 and the top 1000 epistatic SNP pairs in model (1). Bayesian information criterion (Schwarz, 1978) was used as a criterion to choose the tuning parameter. The Lasso method generated 11 non-zero coefficients, including one marginal association locus on chromosome 5 reported before and ten interacting pairs (see Table 3). The total MSE has decreased by 26·4% compared with the previous main effect linear model. The most significant pair obtained by the single pair analysis was also selected by the Lasso method.

For trait HDL, five terms were selected by the Lasso method, including two marginal association loci reported before and three interacting pairs (see Table 3). The total MSE has decreased by 15·9% compared with the main effect linear model. The first two SNP pairs in Table 3 for trait HDL are near the first two pairs based on model (1) listed in Table 2. The analysis results for traits DBP and TG can also be found in Table 3.

Comparing Table 2 with Table 3, we found that the single pair analysis may pick up many SNP pairs, but they are mostly nested within some narrow genomic regions. While in the joint analysis via the Lasso method, SNP pairs that have high correlation with those already included in the model are less likely to be added into the model again, which may help us to find more valuable regions. Furthermore, a final model that can well explain the variability of each trait can be found by the Lasso method.

The main reason that we considered two analysis strategies in the second stage is that the results obtained from the two strategies can supplement each other, i.e. the single pair analysis and the Lasso method also detected their own interacting SNP pairs as well as the common ones. We should not neglect any detected interacting SNP pairs, since the true model between the trait and the main and interaction effects of the SNPs is completely unknown.

https://doi.org/10.1017/S0016672315000038 Published online by Cambridge University Press

Table 3. *SNP pairs selected by the two-stage method via Model (2).*

Trait	Locus 1			Locus 2			Contribution rate (%)[a]
	Chr	Position	Gene or nearest gene	Chr	Position	Gene or nearest gene	
GLU	1	82062500	Gm5530 Irs1	4	29575401	Gm11923 Gm11925	6·16
	12	30998502	Sntg2	18	73383043	-n-R5-8s1	4·45
	11	14125120	4930554G24Rik Gm12006	15	77932972	Cacng2	3·82
	17	87128520	-Epas1	18	78042984	Pstpip2	3·34
	3	143110185	Gm2574-	15	77327564		2·45
	12	113645375		16	66284359	-Cadm2	2·36
	5	121142226	Dtx1	6	147639739	1700049E15Rik	2·27
	4	96540503	Gm12695 Gm10192	16	39177513	Igsf11-	1·33
	6	147670244	Gm15762 B230104C08Rik	7	135179430		1·30
	4	65160973	Astn2	19	18752026	2410127L17Rik	1·06
HDL	4	150470889	Camta1	17	32007869	Sik1 Hsf2bp	9·49
	1	163546745	-Gm15429	17	72311236	Alk	6·15
	10	6331899	Mthfd1l	9	115878311	Gadl1	5·90
DBP	3	111579000		16	41310194	-Lsamp	7·28
	9	115009565	Osbpl10	13	84273654	C130071C03Rik-	6·65
	2	129096371	Ckap2l	15	19802913	Cdh10-	5·37
	11	42826374		7	77717179	B130024G19Rik-	4·26
	11	42826374	-Gm9972	13	84273654	C130071C03Rik-	0·04
TG	1	181436563	Smyd3	5	123349051	Kdm2b	6·09
	1	195979059	4631405K08Rik Plxna2	15	81491304	Ep300 L3mbtl2	4·82
	7	143625565	Mki67-	12	17139883	2410004P03Rik Kcnf1	4·70
	13	103304354	Cd180	15	70171048	-Fam135b	4·70
	3	129089662	Enpep Elovl6	12	39963955	Etv1-	4·06
	11	99310169	Krt20 Krt23	16	49935728	Cd47 Bbx	3·93
	9	89545898	AF529169 Tmed3	11	120587173	Dcxr	3·55
	4	10784867	Gm12919 2610301B20Rik	9	120440907		2·03
	6	65836389	Prdm5	14	17179014	Ngly1 Top2b	1·14
	9	116453569	Tgfbr2 Rbms3	13	104033960	Mast4	1·14
	3	129089662	Enpep Elovl6	9	116453569	Tgfbr2 Rbms3	0·12

[a] The increased rate of MSE after deleting each pair from the final full model selected by the Lasso method.

5. Simulation studies

In this section, simulation studies are designed to investigate the effects of minor allele frequency (MAF), main effects, HWE and LD on the two-stage method we used in the real data analysis.

(i) *Simulation design*

We assumed that a quantitative trait (denoted Y) is affected by four causal SNPs, with two of them having main effects only (denoted X_1 and X_2) and two of them having interaction (denoted X_3 and X_4). Phenotype data were generated from the following linear model:

$$Y = \beta_0 + \sum_{k=1}^{4} \beta_k X_k + \beta_{34} X_3 X_4 + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, $\sigma^2 = 10$, $\beta_0 = 3$. The main effect of the ith SNP (denoted β_i, $i = 1,\ldots,4$) was chosen according to its heritability (denoted H_i), which means the proportion of the trait's variance is explained by the ith locus. As loci 1 and 2 do not have an interaction effect, we set $H_1 = H_2 = 10\%$.

We considered four different situations of loci 3 and 4. For each scenario, 300 individuals were generated, which is similar to the previous real data set and the process was repeated 1000 times. If loci 3 and 4 both came through the variance heterogeneity test with a significant interaction effect in the linear model, we would conclude that an interacting pair has been discovered, and then the discovery rate among 1000 replications was calculated. Four different situations were considered, as discussed in the following sections.

(a) *Situation 1: loci 3 and 4 are independent and have an interaction effect only* ($H_3 = H_4 = 0$)

In this situation, β_3 and β_4 are set to zero and X_i can be encoded as -2q, 1-2q and 2-2q with probability p^2, 2pq and q^2, respectively, such that the mean genotypic value equals to zero. We can easily get:

$$H_i = \frac{2pq\beta_i^2}{Var(Y)} \quad i = 1, ..., 4, \quad \text{and} \quad H = \frac{4p^2q^2\beta_5^2}{Var(Y)},$$

where $Var(Y) = 2pq(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2) + 4p^2q^2\beta_5^2 + 10$ and H denotes the heritability of the interaction effect of loci 3 and 4.

https://doi.org/10.1017/S0016672315000038 Published online by Cambridge University Press

Table 4. *The discovery rates of 1000 simulations in situation 1.*

q^a	H^b (%)						
	10	15	20	25	30	35	50
0·1	0·173	0·012	0·006	0·006	0·006	0·014	0·050
0·2	0·073	0·192	0·334	0·496	0·621	0·712	0·896
0·3	0·051	0·150	0·320	0·508	0·670	0·831	0·985
0·4	0·030	0·120	0·234	0·522	0·759	0·900	1·000

[a] MAF.
[b] The proportion of variance (heritability) explained by interaction effect of loci 3 and 4.

Table 5. *The discovery rates of 1000 simulations with $q = 0{\cdot}3$ in situation 2.*

H (%)	$H_3 + H_4^a$ (%)				
	10	15	20	30	40
5	0·012	0·032	0·066	0·020	0·392
7	0·340	0·094	0·178	0·382	0·594
10	0·118	0·240	0·364	0·606	0·860
15	0·342	0·494	0·632	0·904	0·990
20	0·566	0·704	0·852	0·984	1

[a] The total proportion of variance explained by the marginal effects of loci 3 and 4.

For each MAF q, we respectively took H (%) = 10, 15, 20, 25, 30, 35, 50. For example, when $q = 0{\cdot}3$, $H = 10\%$, we can obtain $\beta = 1{\cdot}22$, 1·22, 0, 0, 2·005, which satisfies the assumptions of heritabilities. As estimates of powers, the discovery rates among 1000 replications were presented in Table 4.

(b) *Situation 2: loci 3 and 4 are independent and have main effects ($H_3 = H_4 \neq 0$)*

In this situation, we investigate the influence of marginal effects of interacting factors on the power of the variance heterogeneity test. We fixed q at 0·3, because the effect of MAF has been studied in situation 1. For each H (%) (5, 7, 10, 15, 20), we respectively took $H_3 + H_4$ (%) = 10, 15, 20, 30, 40. The simulation results are listed in Table 5.

(c) *Situation 3: loci 3 and 4 are independent and HWE is not satisfied for locus 3*

Firstly, we define a measure of skew for HWE. Let A and a denote the two alleles of locus 3 with $P(A) = p$ and $P(a) = q$. By fixing one of the three genotype probabilities, we can define skew coefficient of HWE according to the other two genotype probabilities. For example, Supposing $P(Aa) = 2pq$, skew coefficient r of HWE is defined by the following expressions:

$$p(AA) = p^2 - r,\ P(aa) = q^2 + r.$$

If $r \geq 0$, then $r < p^2$, otherwise $r > -q^2$. $r = 0$ indicates that HWE holds. The simulation results of two different cases by respectively fixing $P(Aa) = 2pq$ and $p(aa) = q^2$ are both given in Table 6.

Table 6. *The discovery rates of 1000 simulations with different r in situation 3.*

r^a	−0·03	−0·02	−0·01	0	0·01	0·02	0·03	0·04
	0·521	0·591	0·637	0·670	0·701	0·756	0·783	0·806
r^b	−0·3	−0·2	−0·1	0	0·1	0·2	0·3	0·4
	0·515	0·599	0·664	0·670	0·638	0·566	0·424	0·254

[a] The skew coefficient of HWE when fixing $P(Aa) = 2pq$.
[b] The skew coefficient of HWE when fixing $P(aa) = q^2$.

Table 7. *The discovery rates of 1000 simulations with $q = 0{\cdot}3$ in situation 4.*

H (%)	θ^a								
	0·9	0·85	0·8	0·75	0·7	0·65	0·6	0·55	0·5
20	0·003	0·009	0·014	0·011	0·013	0·016	0·017	0·016	0·018
25	0·008	0·017	0·028	0·035	0·052	0·065	0·068	0·063	0·068
30	0·012	0·036	0·056	0·084	0·014	0·161	0·188	0·196	0·210
35	0·017	0·077	0·152	0·227	0·304	0·351	0·407	0·436	0·448
40	0·039	0·152	0·313	0·447	0·548	0·601	0·674	0·718	0·713

[a] The LD coefficient between loci 3 and 4.

(d) *Situation 4: loci 3 and 4 are linked and have interaction effect only ($H_3 = H_4 = 0$)*

The performance of the two-stage method on two linked loci with various LD coefficients θ is considered. The detailed information to chose genotypes of loci 3 and 4 and regression coefficients are listed in the supplementary material (File 2). For each H, we respectively took, $\theta = 0{\cdot}9$, 0·85, 0·8, 0·75, 0·7, 0·65, 0·6, 0·55, 0·5. The simulation results are listed in Table 7.

(ii) *Simulation results*

The discovery rate among 1000 replicates is considered as an estimate of power. From Table 4, we can see that the powers increase with increasing interaction effect and MAF except when MAF is small. This result is intuitive. When MAF is 0·1, the power always stays at a low level, even if the interaction effect is highly significant ($H = 50\%$). So the method

https://doi.org/10.1017/S0016672315000038 Published online by Cambridge University Press

Table 8. *The discovery rates of 1000 simulations for different sample sizes.*

	Sample size			
H (%)	300	500	750	1000
10	0·051	0·106	0·168	0·262
20	0·320	0·579	0·822	0·929
30	0·670	0·929	0·994	1

we used may be ineffective in detecting interaction effects among rare variants.

Table 5 shows that the marginal effects of interacting factors can affect the power and the dependence is monotonic in most cases. While some violation may exist for small H. The reasons that contribute to this contra-intuitive phenomenon are explained as follows (Struchalin *et al.*, 2010): from the linear model given before, we can get $\sigma^2_{AA} = Var(y|G=1) = (\beta_2+\beta_3)^2 + \sigma^2$, $\sigma^2_{Aa} = Var(y|G=0{\cdot}5) = (\beta_2 + 0{\cdot}5\beta_3)^2 + \sigma^2$, $\sigma^2_{aa} = Var(y|G=0) = \beta_2^2 + \sigma^2$. Let

$$f(\beta_2) = \frac{\sigma^2_{AA} - \sigma^2_{Aa}}{\sigma^2_{AA}}.$$

When MAF is small, $f(\beta_2)$ is one of the main factors that affects the power of the variance heterogeneity test, and interacting loci tend to be discovered with large $f(\beta_2)$. So an optimum β_2 may exist.

From Table 6 we can see that there is no simple change of powers when HWE is not satisfied. As the skew coefficient r of HWE influences the distribution of the three genotypes, the power of detecting interaction would increase if the skew coefficient r makes the distribution more uniform, otherwise the power would decrease. For example, in the first case, where $P(Aa) = 2pq$, $P(AA) = p^2 - r$ and $P(aa) = q^2 + r$. When $r > 0$, $P(aa)$ would increase r from $q^2 = 0{\cdot}09$ and $P(AA)$ would decrease from $p^2 = 0{\cdot}49$, which makes the three probabilities more close. So the power increases with the increasing of r.

It can be seen from Table 7 that the two-stage method performs poorly when independence assumption of loci is not satisfied. Even for $H = 25\%$, the discovery rate is <10% and the discovery rate also decreases with the increasing of the LD coefficient θ.

In addition, the performance of the two-stage method for cases of different sample sizes was also considered in our simulation. Table 8 lists the change of powers when we increase the number of individuals from 300 to 500, 750 and 1000, where $q = 0{\cdot}3$.

6. Discussion

In this paper, a two-stage method based on Paré's idea was used to conduct a genome-wide interaction analysis for four important traits of NMRI outbred mice. For each trait, we found some SNP pairs that are potentially valuable in further exploring the relationships between genes and these traits. From a bioinformatics point of view, the GO information was also used to auxiliarily explain the selected interacting pairs. Special simulation scenarios were also conducted in this research to evaluate the effects of practical factors such as HWE and LD on the two-stage screening method.

The NMRI outbred mice data set we used in this paper originated from a single population that descended from two males and seven females imported from Lausanne, Switzerland (Lynch, 1969). Therefore, it has minimal population structure and a small proportion of private alleles. Furthermore, all the mice are bred in the same situation, which can avoid wild factors. From this aspect, the population size needed is much smaller than would be needed in human association studies. Association mapping with a population of outbred mice is similar to human genome-wide association studies (GWAS) in many respects; therefore, we expect that the two-stage method and the analysis results in this paper can provide valuable reference for human GWAS.

Simulation results show that the power is affected by the MAF of a SNP and it grows sharply with increasing MAF. In the real data set, we just analysed SNPs with a MAF >2%, thus we may have neglected some rare variants with interaction effects. Fang *et al.* (2011) modified Paré's two-stage approach such that it can be applied to rare variants. They just simply collapsed the rare variants in a gene to a part that is considered as a single variant, so the marginal effect of a rare variant can not be measured. Currently, the issues surrounding rare variant analysis are arousing many researchers' attention.

In our statistical analysis of the mouse data, we assumed that all SNPs are independent. SNPs in different chromosomes may satisfy this assumption, while some SNPs in the same chromosome may have LD, especially for those whose positions are very close. In NMRI mice, there is probably no significant LD between markers approximately more than 10Mb apart (Zhang *et al.*, 2012), and the distance is approximately 0·5Mb in humans (Dawson *et al.*, 2002). So our statistical method may neglect some interacting SNPs that are located on the same chromosome and whose positions are very close. Our further research will be to test interaction effects among dependent variants.

The authors would like to thank the joint Editor and referees for comments that greatly improved the presentation of the paper. This research was supported by the National Natural Science Foundation of China (nos. 11201129 and 11371083), the Natural Science Foundation of

Heilongjiang Province of China (A201207), the Scientific Research Foundation of Department of Education of Heilongjiang Province of China (no. 1253G044) and the Science and Technology Innovation Team in Higher Education Institutions of Heilongjiang Province (no. 2014TD005).

Declaration of interest

None.

Supplementary material

The online supplementary material can be found available at http://journals.cambridge.org/GRH

References

Bai, J. Y., Kooperberg, C., Leblanc, M. & Ross, L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene–environment interaction. *Biometrika* **99**, 929–944.

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lõhmussaar, E., Zernant, J., Tõnisson, N., Remm, M., Mägi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R. & Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548.

Fang, S. R., Ma, W. J. & Zhang, S. L. (2011). A two-stage approach to detect gene–gene and gene–enviroment interaction: application to GAW17 data set. *Journal of Natural Science of Heilongjiang University* **28**, 767–770.

Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P. & Becker, T. (2009). INTERSNP: genome-wide interaction analysis guides by *a priori* information. *Bioinformatics* **25**, 3275–3281.

Kahn, H. A. (1983). *An Introduction to Epidemiologic Methods*. NY: Oxford University Press.

Kooperberg, C. & Leblanc, M. (2008). Increasing the power of identifying gene–gene interactions in genome-wide association studies. *Genetic Epidemiology* **32**, 255–263.

Lynch, C. J. (1969). The so called Swiss mouse. *Laboratory Animal Care* **19**, 214–220.

Mckinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. (2006). Machine learning for detecting gene–gene interaction: a review. *Applied Bioinformatics* **124**, 214–220.

Olkin, I., Sudhish, G., Ghurye, W., Hoeffding, W. & Madow, H. (1960). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. CA: Stanford University Press.

Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. *PLoS Genetics* **6**, e1000981.

Schwarz, Z. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Struchalin, M. V., Dehghan, A., Witteman, J. C., Duijin, C. & Aulchenko, Y. S. (2010). Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genetics* **11**, 92.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B* **58**, 267–288.

Wu, Z., Aporntewan, C., Ballard, D. H., Lee, J. Y., Lee, J. S. & Zhao, H. (2009). Two-stage joint selection method to identify candidate markers from genome-wide association studies. *BMC Proceedings* **3**, S29.

Zhang, W., Korstanje, R., Thaisz, J., Staedtler, F., Harttman, N., Xu, L., Feng, M., Yanas, L., Yang, H., Valdar, W., Churchill, G. A. & Dipetrillo, K. (2012). Genome-wide association mapping of quantitative traits in outbred mice. *G3 (Bethesda)* **2**, 167–174.

Zhang, Y. & Liu, J. S. (2007). Bayesian inference of epistatic interactions in case control study. *Nat Genetics* **39**, 1167–1173.