BMC Genetics

Methodology article

Regression-based approach for testing the association between multi-region haplotype configuration and complex trait

Yanling Hu[1], Sinnwell Jason[2], Qishan Wang[1], Yuchun Pan*[1], Xiangzhe Zhang[1], Hongbo Zhao[1], Changlong Li[3] and Libin Sun[4]

Address: [1]School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, PR China, [2]Mayo Clinic, Division of Biomedical Statistics and Informatics, Rochester, Minnesota, USA, [3]Zhejiang Provincial Laboratory of Experimental Animals & Non-clinical Studies, Hangzhou, PR China and [4]Shanghai Institute of Veterinary Hygiene, Shanghai, PR China

Email: Yanling Hu - ylhu0323@sjtu.edu.cn; Sinnwell Jason - sinnwell.Jason@mayo.edu; Qishan Wang - wangqishan@sjtu.edu.cn; Yuchun Pan* - panyc@sjtu.edu.cn; Xiangzhe Zhang - xiangzhezhang@sjtu.edu.cn; Hongbo Zhao - zhaohb@sjtu.edu.cn; Changlong Li - li-changlong@126.com; Libin Sun - slb__2002@163.com

* Corresponding author

Published: 17 September 2009

BMC Genetics 2009, **10**:56 doi:10.1186/1471-2156-10-56

Received: 20 January 2009
Accepted: 17 September 2009

This article is available from: http://www.biomedcentral.com/1471-2156/10/56

Abstract

Background: It is quite common that the genetic architecture of complex traits involves many genes and their interactions. Therefore, dealing with multiple unlinked genomic regions simultaneously is desirable.

Results: In this paper we develop a regression-based approach to assess the interactions of haplotypes that belong to different unlinked regions, and we use score statistics to test the null hypothesis of non-genetic association. Additionally, multiple marker combinations at each unlinked region are considered. The multiple tests are settled via the *minP* approach. The *P* value of the "best" multi-region multi-marker configuration is corrected via Monte-Carlo simulations. Through simulation studies, we assess the performance of the proposed approach and demonstrate its validity and power in testing for haplotype interaction association.

Conclusion: Our simulations showed that, for binary trait without covariates, our proposed methods prove to be equal and even more powerful than htr and hapcc which are part of the FAMHAP program. Additionally, our model can be applied to a wider variety of traits and allow adjustment for other covariates. To test the validity, our methods are applied to analyze the association between four unlinked candidate genes and pig meat quality.

Background

Haplotypes, the linear arrangement of alleles on the same chromosome inherited as a unit, provide a natural framework for testing the association between genetic markers and complex traits more efficiently than separate marker analysis[1]. There is strong evidence that several mutations in *cis* position within a single gene can interact to create a "super allele" that has a large effect on the observed phenotype. The biological explanation for these haplotype effects is that several mutations in a gene cause several amino acid changes in the ultimate protein product, and the joint effect of these amino acid changes can have a much larger influence on the function of the protein product than any single amino acid change. This emphasizes the importance of examining candidate genes by SNP haplotyping. Some studies focus on haplotypes

within a given genomic region [2-7]. Because complex traits are presumed to be the results of interaction by a set of genes which may be located in different regions, some methods aim to test gene-gene interaction, and interactions of single markers from different unlinked regions [8-11]. Specifically, Becker *et al*[12] reported a method to deal with haplotype interaction in unlinked regions for a binomial trait. They find the best haplotype combination from the unlinked regions by permutation, which is a modification of Ge *et al*[13]. However, this method could only be applied to case-control association testing, and could not include other covariates.

Generalized Linear Model (GLM) is an extension of the general linear modeling process that allows models to be fitted for several kinds of traits, such as Gaussian, Poisson, Binomial, etc., and allows various covariates. Schaid *et al*[5] introduced score statistics, which are receiving increased attention because they require only computation of the null estimates and are asymptotically equivalent to Wald and likelihood ratio statistics under both null and Pitman alternative hypotheses. Some methods that use score tests based on GLM to test haplotype-trait association have demonstrated the validity and power of this statistic[6]. However, these methods only considered one genomic region. If considering multi-region multi-marker haplotype configurations, a severe multiple-testing problem will occur. To obtain uncorrected *P*-values for a specific marker combination, we use an unnested simulation introduced by Becker and Knapp[2], which is based on the algorithm proposed by Ge *et al*[13].

We propose an alternative approach that uses score statistics based on GLM to build the statistic T over which some of the unlinked regions are considered and some markers are chosen at the selected regions. Since the distribution of T is generally unknown and is generally not comparable, we replace T with P^{min} which is inherited from the algorithm of Becker and Knapp[2]. This simulation method has already been validated by Manly[14] and Hoh *et al*[15], and has systematically been applied to some genetic data[2,12,16,17].

Simulation study

Simulation schemes

We conduct a simulation study to evaluate the power and type I error of the association test and to compare our approach with others. The haplotype data are generated in a way similar to that of Roeder *et al*.[18] and Tzeng *et al*[6]. In every simulation scheme, we consider two unlinked regions. We consider that markers are in strong linkage disequilibrium within each region, but markers from different regions are in linkage equilibrium. Therefore, we separately produce two regions by using a modified Hudson's MS program[19]. This program generates data under a coalescent model in which the recombination occurs uniformly over the region. The 4 samples sizes are 50, 100, 200 and 400, respectively. The scaled recombination rate, $\rho = N_e\delta/bp$, is set to 4×10^{-3} for the recombination cold spots, and 100 times greater in the hot spots, with the effective population size N_e is 1×10^4. The scaled mutation for the entire region, $4N_e\mu/bp$, is set to be 6×10^{-4}. Once the haplotypes have been generated, the first step is to restrict the disease or minor allele frequency. In this simulation, we set the allele frequency as 3 levels: 0.1, 0.3, and 0.5. We assume that the middle locus in every gene is the liability locus. Once a liability locus is chosen, a haplotype is defined as a segment of three adjacent SNPs in which the second SNP is the liability locus.

After randomly pairing haplotypes to form individual genotypes, we generate both continuous and binary trait values.

Continuous traits

For the Type I error test, we consider two simple models of quantitative traits simulated independently of the liability locus. Let model 1 include only an environmental effect e: $\boldsymbol{Y} = \boldsymbol{e}$. Let model 2 additionally incorporate a covariate Z: $\boldsymbol{Y} = \gamma \times \boldsymbol{Z} + \boldsymbol{e}$. In the models, e follows a standard normal distribution with mean 0 and variance 1, and Z is generated from a standard normal distribution. For assessing power and the effective selection of the best combination of markers, we also consider two models of quantitative traits simulated in association with the liability locus. Let model 3 decompose the trait value into MRHC effect and environmental effect e: $Y = g + e$, and model 4 additionally incorporate a covariate Z: $Y = g + \gamma \times Z + e$. In these models, g is the sum of all considered genes' effects. For the ith gene, g_i has a discrete distribution and equals u_2^i, u_1^i, u_0^i with probabilities q^2, $2q(1\text{-}q)$ and $(1\text{-}q)^2$, respectively. As in models 1 and 2, e follows a normal distribution with mean ε and variance σ_e^2, and Z is generated from a standard normal distribution. For simplicity, we set $u_k^i = k - 1$, $\varepsilon = 0$, and $\gamma = 1$. The trait values are generated using the normal penetrance function. $f(Y \mid 1, \ldots, m) = N(\sum_{i=1}^{m}(u_1 + \ldots + u_m), \sigma_e^2)$ for the first model and $f(Y \mid 1, \ldots, m) = N(\sum_{i=1}^{m}(u_1 + \ldots + u_m) + \gamma \times Z, \sigma_e^2)$ for the second model, where m is the number of the genes. We determine σ_e^2 through the heritability h^2 of all liability loci, which we set at 0.4.

Binary traits

We generate binomial phenotypes on the basis of the above continuous traits, and consider four models where the disease prevalence is set to 0.10. If the values of the above continuous traits are more than a given threshold, we set the traits as disease cases, otherwise we set them as control. Let model 5 be the binary trait created from a continuous trait simulated as in model 1, model 6 from model 2, model 7 from model 3, and model 8 from model 4. Binary traits are simulated until an equal number of cases and controls are reached.

The detail of the models can be seen in table 1.

Under all scenarios, we compute the global *P* value with 1500 permutation replicates for each simulated data set. Empirical significance levels and power were computed as the portion of simulated data sets for which the global *P* value was less than or equal to 0.05.

Results

Comparison of three models htr, hapcc and HAPGLM

In order to check the validity and the accuracy of our HAPGLM approach, we first carry out simulations under the null hypothesis and compare it with hapcc and htr, which were implemented in the beta version of FAMHAP. htr performs a haplotype trend regression test proposed by Zaykin *et al*[7], and hapcc performs a χ^2 test for haplotypes proposed by Becker *et al*[12,16]. Here, we use model 5 and 7 to simulate the trait. Haplotypes and trait values are compared according to the frequency (q = 0.1, 0.3, 0.5) of the disease allele, and sample size (n = 50, 100, 200, 400).

First, we discuss the results under the effect of minor allele frequency for type I error in table 2. Under 12 scenarios, the type I error of the three models is near 0.05, and there are not significant differences between the three models. The results show that our model can approximate to hapcc and htr in accuracy. For the power comparison, table 3 presents worst performance when disease allele frequency is high with small sample sizes, where the power of the global test is not stabilized, especially for the *hapcc*. The reason is that the disease individual prevalence is 0.10, and the percent for the disease is somewhat small, making it difficult to find the significant difference. However, when the sample size is more than 100, the power is near one for the three models. There are no significant differences among hapcc, htr and our method.

Three factors analysis for global test

To evaluate the test performance, we describe the results from our power and type I error study that use the above methods with various parameters. Type I error test includes 48 scenarios which include 4 models, 3 minor allele frequencies and 4 sample sizes. As shown in table 4, the type I error stabilizes in all the scenarios. Power test includes 48 scenarios which include 4 models, 3 disease allele frequencies and 4 same sizes. For the power calculations in table 5, the power is adversely affected by the small sample size and high disease allele frequency. Otherwise, if the sample size is at least to 100, the power is preserved. Therefore, we set sample size to test recombination affection and the specific MRHC testing as 100.

Recombination analysis for global test

In order to check the effect of recombination on the model, we first consider two different recombination levels at which the diversity of the haplotype is high and low. "High diversity" indicates that a minor or disease locus is located in the region of recombination hot spots and that the number of distinct haplotypes is 6-8; "low diversity" indicates that a minor or disease locus is located in a haplotype block and the number of distinct haplotype is 3-5. In this simulation, we consider three SNPs in each region and we assume equal recombination level. The first recombination level has 2-4 different haplotypes. The haplotype distribution of the second recombination level consists of 6-7 theoretically possible haplotypes. From table 6 and 7, there are some differences between the two diversities on type I error and power, but there are no significant differences between the high and low diversities.

Table 1: the 8 models in the trait producing in simulation

Model name	Trait type	Factor consider
Model 1	Continuous traits	include only an environmental effect e
Model 2	Continuous traits	include an environmental effect e and a covariate
Model 3	Continuous traits	include MRHC effect and environmental effect e
Model 4	Continuous traits	include MRHC effect and environmental effect e and a covariate
Model 5	Binary Traits	Produce from model 1 above a given threshold
Model 6	Binary Traits	Produce from model 2 above a given threshold
Model 7	Binary Traits	Produce from model 3 above a given threshold
Model 8	Binary Traits	Produce from model 4 above a given threshold

Table 2: Type I error of three models via 1500 simulations at α = 0.05

Sample Size	Minor Allele	Model Type		
		hapcc	***htr***	***HAPGLM***
50	q = 0.1	0.097(0.079-0.117)	0.048(0.036-0.063)	0.048(0.036-0.063)
	q = 0.3	0.032(0.022-0.045)	0.039(0.028-0.053)	0.044(0.032-0.059)
	q = 0.5	0.029(0.020-0.041)	0.041(0.030-0.055)	0.047(0.042-0.071)
100	q = 0.1	0.036(0.025-0.049)	0.035(0.024-0.048)	0.047(0.035-0.062)
	q = 0.3	0.031(0.021-0.044)	0.072(0.057-0.090)	0.043(0.031-0.057)
	q = 0.5	0.042(0.030-0.056)	0.047(0.035-0.062)	0.051(0.038-0.067)
200	q = 0.1	0.031(0.021-0.044)	0.026(0.017-0.038)	0.055(0.042-0.071)
	q = 0.3	0.051(0.038-0.067)	0.033(0.023-0.046)	0.047(0.035-0.062)
	q = 0.5	0.040(0.029-0.054)	0.067(0.052-0.084)	0.037(0.026-0.051)
400	q = 0.1	0.035(0.024-0.048)	0.037(0.026-0.051)	0.042(0.030-0.056)
	q = 0.3	0.047(0.035-0.062)	0.028(0.019-0.040)	0.041(0.030-0.055)
	q = 0.5	0.028(0.019-0.040)	0.041(0.030-0.055)	0.042(0.030-0.056)

That is to say, the proposed method is not significantly affected by the recombination level.

Analysis for specific MRHC

The score test can easily compute the specific MRHC. In order to study the performance of the proposed method in detecting individual MRHCs, we set 3 disease allele frequencies and 4 model fits. The power for specific MRHC is presented in table 8. It is of interest that the proposed method is robust for detecting the specific MRHC.

Select the best combination of markers

In order to check the accuracy of our methods to select the best MRHC, the markers of the 2 genes in simulations are 123|456, and the marker 2 and 5 are as the liability loci. The frequency of the disease allele is 0.3. The result shows that for these 2 liability loci combination, the statistic *T* is the largest and the *P* value is the smallest. These are presented in table 9. The combination without these two loci presents low *T* and high *P* value.

Application to a pig meat quality dataset

Meat quality is very important in the pig meat production industry. Many candidate genes have been identified that could be used to improve this trait through marker assisted selection (MAS)[20]. The Heart Fatty Acid-Binding (*H-FABP*) gene encodes a type of cytosol protein that transports fatty acids from the cell membrane to other sites where 3-acyl-glyceride and phospholipids are synthesized and fatty acids are oxidized. Gerbens [21-23] discovered *Msp* I, *Hae* II and *Hinf* I polymorphisms of the *H-FABP* gene that is related to intramuscular fat content. Melanocortin-4 Receptor (MC4R) is believed to be a link between feed intake and body weight[24].

Polymorphism of the *MC4R* gene has been reported to be associated with back fat thickness[25]. Adipocyte Determination and Differentiation factor-1 (*ADD1*) can activate or restrain some genes in fat and glucose metabolism. Research has suggested that the *ADD1* gene can be used as a candidate gene for pork quality[26]. Calpastatin (*CAST*), which is an endogenous inhibitor (Ca^{2+} dependent

Table 3: Power of three models via 1500 simulations at α = 0.05

Sample Size	Disease Allele Frequency	Model Type		
		hapcc	***htr***	***HAPGLM***
50	q = 0.1	0.947	0.947	0.947
	q = 0.3	0.866	0.907	0.968
	q = 0.5	0.666	0.391	0.596
100	q = 0.1	0.977	0.971	0.952
	q = 0.3	0.963	0.927	0.941
	q = 0.5	0.834	0.728	0.917
200	q = 0.1	0.980	0.980	0.968
	q = 0.3	0.981	0.981	0.968
	q = 0.5	0.961	0.912	0.954
400	q = 0.1	0.967	0.981	0.948
	q = 0.3	0.981	0.983	0.974
	q = 0.5	0.934	0.925	0.967

Table 4: Type I error of global test via 1500 simulations at α = 0.05

Sample Size	Minor Allele Frequency	Model Type			
		Model 1	Model 2	Model 5	Model 6
50	q = 0.1	0.034(0.024-0.047)	0.031(0.021-0.044)	0.048(0.036-0.063)	0.039(0.028-0.053)
	q = 0.3	0.050(0.037-0.065)	0.041(0.030-0.055)	0.044(0.032-0.059)	0.044(0.032-0.059)
	q = 0.5	0.052(0.039-0.068)	0.062(0.048-0.079)	0.047(0.035-0.062)	0.056(0.043-0.072)
100	q = 0.1	0.040(0.029-0.054)	0.032(0.022-0.045)	0.047(0.035-0.062)	0.037(0.026-0.051)
	q = 0.3	0.038(0.027-0.052)	0.044(0.032-0.059)	0.043(0.031-0.059)	0.037(0.026-0.051)
	q = 0.5	0.048(0.036-0.063)	0.044(0.032-0.059)	0.051(0.038-0.067)	0.052(0.039-0.068)
200	q = 0.1	0.045(0.033-0.060)	0.041(0.030-0.055)	0.055(0.042-0.071)	0.031(0.021-0.044)
	q = 0.3	0.048(0.036-0.063)	0.041(0.030-0.055)	0.047(0.035-0.062)	0.042(0.030-0.056)
	q = 0.5	0.047(0.035-0.062)	0.049(0.036-0.064)	0.037(0.026-0.051)	0.041(0.030-0.055)
400	q = 0.1	0.027(0.018-0.039)	0.040(0.029-0.054)	0.042(0.030-0.056)	0.038(0.027-0.052)
	q = 0.3	0.043(0.031-0.059)	0.037(0.026-0.051)	0.041(0.030-0.055)	0.044(0.032-0.059)
	q = 0.5	0.041(0.030-0.055)	0.052(0.039-0.068)	0.042(0.030-0.056)	0.048(0.036-0.063)

cysteine proteinase), plays a central role in the regulation of calpain activity in cellsand is considered to be one of the major modulators of the calpains[27,28]. The *CAST* gene represents an excellent candidate gene for studying variation in pork quality. We aim to find association between multi-region haplotype effects from these candidate regions and meat quality.

Our data set is a sample which includes 93 unrelated fatteners from the following breeds/populations: 18 Meishan, 21 Sutai, 14 Yorkshire × Sutai, 16 Landrace × Sutai and 24 Duroc × Landrace × Yorkshire pigs. 8 polymorphic markers of the preceding genes in the populations have been reported[29,30], which are: 2 in *ADD1*, 1 in *H-FABP*, 1 in *MC4R* and 4 in *CAST*. We code these polymorphic markers as (A1, A2, H1, M1, C1, C2, C3, and C4). The χ^2 test of these polymorphic markers show that there are significant differences in 5 polymorphic markers (except for A1, A2, C4) between the five populations. We set up single locus models including sex and breed as environmental covariates, for every polymorphic marker, and use statistical software SAS macro GLM for calculations. The results show significant effects at 0.05 level for: A1, H1, C2, C3 and C4 in back fat thickness (BK); A1, C1 and C3 in meat color (MC); A1, H1, C1 and C4 in intramuscular fat content (IMF), and A1 in protein content. To apply our methods, we also incorporate two environmental covariates (sex and breed), and use back fat thickness, tenderness, drip loss, meat color, intramuscular fat content, pH 1 hour after slaughter, pH 24 hours after slaughter, and the content of protein as the dependent variable in the regression model. Table 10 illustrates the marker combination in which raw *P* values are lower at 0.01 level, Compared to single locus model, MRHC analysis can detect more markers which are significantly associated with traits.

Additionally, our methods are used to reconstruct the distinct MRHC from the above 4 genes which are in unlinked

Table 5: Power of global test via 1500 simulations at α = 0.05

Sample Size	Disease Allele Frequency	Model Type			
		Model 3	Model 4	Model 7	Model 8
50	q = 0.1	0.977	0.824	0.947	0.478
	q = 0.3	0.964	0.920	0.968	0.891
	q = 0.5	0.947	0.625	0.596	0.741
100	q = 0.1	0.934	0.936	0.952	0.937
	q = 0.3	0.979	0.957	0.941	0.889
	q = 0.5	0.965	0.836	0.917	0.887
200	q = 0.1	0.978	0.963	0.968	0.916
	q = 0.3	0.968	0.972	0.968	0.967
	q = 0.5	0.967	0.971	0.954	0.960
400	q = 0.1	0.968	0.967	0.948	0.944
	q = 0.3	0.971	0.947	0.974	0.920
	q = 0.5	0.948	0.958	0.967	0.962

Table 6: Type I error of global test for different recombination at α = 0.05

Haplotype Diversity	Minor Allele Frequency	Model Type			
		Model 1	Model 2	Model 5	Model 6
High	0.1	0.046(0.034-0.061)	0.053(0.040-0.069)	0.045(0.033-0.060)	0.043(0.031-0.057)
	0.3	0.051(0.038-0.067)	0.043(0.031-0.057)	0.047(0.035-0.062)	0.038(0.027-0.052)
	0.5	0.058(0.044-0.074)	0.052(0.039-0.068)	0.042(0.030-0.056)	0.042(0.030-0.056)
Low	0.1	0.047(0.035-0.062)	0.041(0.030-0.055)	0.042(0.030-0.056)	0.043(0.031-0.057)
	0.3	0.039(0.028-0.053)	0.044(0.032-0.059)	0.037(0.026-0.051)	0.047(0.035-0.062)
	0.5	0.040(0.029-0.054)	0.051(0.038-0.057)	0.044(0.032-0.059)	0.050(0.037-0.065)

regions. In table 11, we illustrate the 5 top statistics for MRHCs which consider 8 markers combination.

Discussion

Presently many publications have proven that the genetic dissection of complex traits depends not only on the identification of genes involved in disease susceptibility but also on the elucidation of the synergistic role that genes play with other genes and with environmental factors[8-11,31-33]. Therefore, considering unlinked genomic regions simultaneously is desirable. There are two models hapcc and htr in FAMHAP program which can compute the haplotype interaction in unlinked region. For hapcc, Becker *et al.*[12] chose the usual χ^2 test statistic for contingency tables which can be applied only to case-control traits. For htr, the haplotype trend regression test proposed by Zaykin *et al*[7], chose *F* statistic and could be used for qualitative and quantitative traits, but won't allow other covariates in the model. Our proposed methods are based on score equations for GLMs which allow adjustment of covariates and can model qualitative and quantitative traits. For binary trait without covariate, the type I error and power comparison show that our model has the same power as hapcc and htr, and type I error is as expected. For a small number of markers, the run times for hapcc, htr and HAPGLM are approximately equal. Additionally, our model has obvious advantages. First, our model can be applied to analyze haplotype association across independent regions with adjusting of covariates for a wider variety of traits. Second, the score statistic (z_k^2) of the individual MRHCs can be easily computed.

Our model adopted the simulation method proposed by Becker *et al.*[2,12,17], which can be computationally feasible to deal with the multiple marker combination. For our program, the evaluation of a single simulated data set with 15 markers in 3 regions will take no more than 10 seconds on average on two nodes with 3.0 GHz Intel with 512 MB main memory. In general, it will be possible to simultaneously consider about 600 to 1,200 hypotheses on a standard PC. Our program is very flexible to allow selection of loci and genes for analysis. However, in regions with too many possible haplotype combinations, our program runs out of memory. We need to consider more aggressive trimming parameters or other haplotype estimation algorithms. For example, we will improve our program by using the haplotype ancestral cluster idea to cluster rare haplotypes with similar ancestral haplotypes, which was used by Tzeng *et al.*[34]. Haplotypes of the entire block can be represented by a smaller set of SNPs which are referred to as tag SNPs[35]. In order to analyze more markers, it will be helpful to select tag markers at each region and to carry out the analysis on the set of these markers. Tag SNPs selection will save run-time, and we plan our further research along this path.

In this research, we proposed markers from different regions which are proposed to be in linkage equilibrium.

Table 7: Power of global for different recombination level at α = 0.05

Haplotype Diversity	Disease Allele Frequency	Model Type			
		Model 3	Model 4	Model 7	Model 8
high	0.1	0.971	0.860	0.955	0.957
	0.3	0.967	0.962	0.961	0.961
	0.5	0.948	0.968	0.953	0.965
low	0.1	0.973	0.977	0.972	0.972
	0.3	0.963	0.944	0.951	0.958
	0.5	0.977	0.933	0.948	0.953

Table 8: Power for the specific MRHC at α = 0.05

Disease Allele Frequency	Model Type			
	Model 3	Model 4	Model 7	Model 8
0.1	0.981	0.856	0.978	0.946
0.3	0.943	0.962	0.916	0.937
0.5	0.933	0.954	0.946	0.952

Markers from different regions can be in linkage disequilibrium, but the methods can allow such markers as if they were in the same region.

Since the current model assumes that the subjects are independent of each other (i.e., unrelated), it is critical to extend the current approach to account for the correlations between subjects, given their family data. Therefore, further studies are needed to address the impact and modeling strategies with regard to the assumptions in the model. This model is restricted to outcomes that can be placed in the generalized linear model framework. An extension to failure-time data could also be placed in the framework.

Conclusion

It is quite common that the genetic architecture of complex traits involves many genes and their interactions. Therefore, dealing with multiple unlinked genomic regions simultaneously is desirable. We developed a regression-based approach which can be applied to a wider variety of traits and allow adjustment for other covariates to assess the interactions of haplotypes that belong to different unlinked regions. Multiple marker combinations at each unlinked region are also considered. In addition, HAPGLM can be downloaded for free at: ftp://public.sjtu.edu.cn/, user: ylhu0323, password: public.

Methods

Contribution to multi-region haplotype configurations

Consider R unlinked genomic regions, and m_r observed markers for each of n unrelated individuals in region r. Further, we suppose that markers within each region are in strong linkage disequilibrium, while markers from different regions are in linkage equilibrium. What we wish to investigate is whether some of the genomic regions are associated with the phenotype of interest via some of the markers from each region.

Let G^r be the multi-locus genotype of an individual at region r, and h^r be a haplotype of region r. If the haplotypes h_j^r and h_k^r are compatible with G^r, (h_j^r, h_k^r) is then called a haplotype explanation of G^r. Obviously, for a given G^r, there may be several haplotype explanations.

First, we use the expectation-maximization (EM) algorithm[5] to obtain the maximum-likelihood estimates of the haplotype frequencies at each of the unlinked regions. After we pooled the rare haplotypes (with estimated frequencies <0.001) into a single group, we adopt the following formula ([12] to compute the likelihood weights of all haplotype explanations of each region for each individual,

$$w_{j,k}^r = \frac{(2-\delta_{j,k}^r) f_j^r f_k^r}{\sum_{(h_j^{\sim r}, h_k^{\sim r}) \in C^r} (2-\delta_{j,k}^{\sim r}) f_j^{\sim r} f_k^{\sim r}} \qquad (1)$$

where G^r is the multilocus genotype of a fixed individual at region r. Let $C^r = \left\{ (h_j^r, h_k^r) : h_j^r \middle| h_k^r = G^r \right\}$ be the set of unordered haplotype explanation, which are compatible with G^r. let f_j^r be the estimated frequency of haplotype h_j^r, the sum in the denominator runs over all possible haplotype explanations, and the Kronecker symbol δ is defined as $\delta_{j,k} = 1$ if $j = k$, and $\delta_{j,k} = 0$ if $j \neq k$, where j and k is the pair of haplotypes at region r. The "$\sim r$" is all the pair haplotypes in region r for each individual.

Let $G = (G^1, G^2, \text{and}, G^R)$ be the multi-region genotype of an individual and $[(h_j^1, h_k^1), (h_p^2, h_q^2), \cdots (h_s^R, h_t^R)]$ be a possible multi-region haplotype explanations (MRHEs) that are compatible with $(G^1, G^2, \text{and}, G^R)$. Let $(h_j^1, h_k^2, ..., h_l^R)$ denote a multi-region haplotype configuration (MRHC).

Table 9: The best ten combinations in two regions with six markers

Model	1	2	3	4	5	6	7	8	9	10
Model 5	2|5	12|5	23|5	123|5	245	234|5	123|45	12|56	123|56	12|456
Model 6	5	2	2|5	3|5	2|6	23	12	56	2|4	45
Model 7	2|5	12|45	13|456	123|45	2|45	23|45	123|56	12|456	123|45	12|56
Model 8	12|5	2|5	23|5	2|45	12|45	2|56	123|5	23|56	123|45	12|56

The possible number for the combination is 2^6-1 with 6 markers in 2 regions.

Table 10: All marker combinations with raw *P* values less than 0.01

Trait Type	Number	Marker Combination
BK[a]	15	3|67, 1|3|568, 1|3|7, 3|58, 2|3|58, 3|568, 3|57, 1|3|5, 1|3|57, 3|567, 1|3|56, 2|3|68, 1|3|78, 2|3|78, 3|578
Tend[b]	34	4|68, 4|56, 578, 568, 4|568, 1|4|58, 1|4|68, 4, 3|56, 1|2|57, 4|567, 3|58, 3|4|5, 1|57, 3|4|58, 3|568, 4|67, 5|67, 4|78, 58, 5678 4|678, 4|58, 4|57, 3|4|7, 3|4|6, 3|578, 1|3|4|5, 1|4|5, 1|3|678, 3|4|578, 1|3|4|8, 1|3|56,1|3|5
DL[c]	4	1|3|678, 3|4|578, 567, 58
MC[d]	10	1|3|4|5, 3|4|678, 1|3|57, 1|568, 1|4|567, 1|4|57, 1|4|578, 1|56. 1|3|56, 1|3|67
IMF[e]	4	1|3|4|5, 1|4|56, 4, 568
pH1[f]	32	1|3|4|5, 1|3|5, 1|3|56, 1|3|57, 1|3|6, 1|3|67, 1|3|7, 1|4|567, 1|4|57, 1|4|578, 1|4|58, 1|4|6, 1|56, 1|578, 3|4|56, 3|4|6, 3|4|678,3|4|7, 3|4|8, 3|56, 3|57, 3|568, 3|57, 3|58, 3|6, 3|67, 3|678, 3|7, 3|78, 4|56, 4|58, 4|6
pH24[h]	5	1|4|8, 3|4|8, 4|6, 58, 8
Protein[i]	8	1|3|6, 5, 1|5, 12|3, 1|4|8, 3|58, 1|3|8, 12|3|4|5

[a] back fat thickness, [b] tenderness, [c] drip loss, [d] meat color, [e] intramuscular fat content, [f] PH after 1 hour's slaughter, [h]PH after 24 hours' slaughter, [i] the content of protein. The eight markers (A1, A2, H1, M1, C1, C2, C3, C4) as 1 2 3 4 5 6 7 8, and same as follows.

An MRHE is formed by two MRHCs from the two gametes, but there may be 2^R MRHCs for a given MRHE, some of which may be the same. We construct an $n \times h$ matrix X'_g with rows referring to the n individuals and columns referring to the different MRHCs. Cell x_{ef} of this matrix denotes the contribution of individual e to MRHC f, and can be calculated according to the following equation,

$$2\prod_{r=1,\ldots,R} w_{jk}^{r}\frac{(1+\delta_{j,k}^{r})}{2} \qquad (2)$$

Regression model with MRHC

Let γ denote an $n \times 1$ vector of measured phenotypes of a trait, α denote an $h \times 1$ vector of the effects for the MRHCs, β denote the regression parameters for the intercept and environmental factors, X'_g be the contribution matrix obtained above, and X'_e denote the design matrix corresponding to measured environmental factors. Then we have the following generalized linear model (GLM):

$$g(EY) = \eta = X'_g\alpha + X'_e\beta \qquad (3)$$

Let $Z = X_e|X_g$ and $\gamma = (\alpha|\beta)$. Then, the likelihood of trait γ_i for subject i, given the vector Z_i, can be expressed as a GLM for exponential family data[36] according to

$$L(\gamma_i \mid Z_i) = \exp\left[\frac{\gamma_i\eta_i - b(\eta_i)}{a(\varphi)} + c(\gamma_i, \varphi)\right] \qquad (4)$$

where a, b, and c are known functions, and φ is the dispersion parameter. To implement the score statistics for different types of traits, we need only assume a distribution for the trait and to make the appropriate substitutions for the expected value of the trait, $\tilde{\gamma}$, the dispersion parameter $a(\varphi)$, the ratio $b"(\eta)/a(\varphi)$ and the link function. (see table 1 of [5]).

Score test for incorporating contribution of MRHC

We derive score statistics to test the null hypothesis of no association between MRHC and trait, $H_0: \alpha = 0$. Let ζ denote the vector of nuisance parameters (β, μ, φ). The likelihood function for (α, ζ) on the basis of the data (Y, X'_g, X'_e) according to Tzeng *et al*[6] is

$$L(\alpha, \zeta, X_g, X_e) = \prod_{i=1}^{n}\left\{\sum_{x_{g,i}\in G_i} f(\gamma_i \mid x_{g,i}, x_{e,i}; \alpha, \zeta)\times P(x_{g,i})\right\} \qquad (5)$$

Table 11: The specific MRHC of 4 regions of haplotype interaction

Trait Type	The specfic haplotype configurations in 4 genes
BK	AG|A|A|AGGA, AG|A|G|GGGA, AG|A|G|AGGA, AG|G|G|AGAA, AG|G|G|AGAA
Tend	AG|A|G|AGGA, AG|A|G|AGGA, AG|A|G|AGGA, GG|A|A|AAAA[a], GG|G|A|AAAA[a]
DL	AG|A|G|AGAA, AG|A|G|AGGA, AG|G|G|AGAA, GG|G|A|AAAA[a], AA|A|A|AAAA[a]
MC	AG|A|A|AGGA, AG|A|G|AAGA, AG|A|G|AGAA, AG|A|G|AGGA, AG|G|G|AGAA[a]
IMF	AG|A|G|AGAA, AG|A|G|AGGA, AA|G|G|AAAG[a], AA|G|A|AAAG[a], AG|G|G|AAGG[a]
pH1	AG|A|G|AGAA, AG|A|G|AGGA, AG|G|G|AGAA, AG|G|G|AAAA[a], AG|A|A|AAGA
pH24	AG|A|G|AGAA, AG|A|G|AGGA, AG|G|G|AGAA, AG|A|A|AGGA[a], AG|A|A|AAGA[a]
Protein	AG|A|G|AGAA, AG|A|G|AGGA, AG|A|A|AGGA, AG|A|A|AGGA[a], AA|A|A|AAGA[a]

[a] denote the statistics z_k is very low.

Where $P(x_{g,\ i})$ is the contribution of individual i to MRHCs.

The score function for α is the partial derivative of the likelihood equation (5), with respect to α. The resulting score statistic, denoted by S_{α}, is the score function evaluated at the restricted maximum-likelihood estimate under the null hypothesis. S_{α} is the statistic that we use to test MRHC effect; in appendix A, we show the following result:

$$S_{\alpha}(Y, X_g, X_e, \alpha, \zeta) = \sum_{i=1}^{n} \frac{\gamma_i - E(\gamma_i)}{a(\varphi)} E(X_{g,i} \mid G_i) \Bigg|_{\substack{\alpha = \tilde{\alpha} \\ \zeta = \tilde{\zeta}}} \tag{6}$$

where $\tilde{\alpha}$ and $\tilde{\zeta}$ are the restricted maximum-likelihood estimated under the null hypothesis, and $E(X_{g,\ i}|G_i)$ is the contribution of individual i to MRHC under the observed multi-region genotypes, G.

To test the association between MRHC and trait that adjusts for other covariates, we need to compute the variance of S_{α} under the null hypothesis H_0: $\alpha = 0$. Under H_0, S_{α} is asymptotically distributed as multivariate normal[36]. We consider the generalized score test, which would ensure the asymptotic null χ^2 distribution even under model misspecification[5]. Define $\theta = (\alpha, \zeta)$ and let V_{α} denote the variance of S_{α}, the equation can be expressed according to Louis[37] and Tzeng *et al*[6] as:

$$V_{\alpha} = (D_{\alpha\alpha} - I_{\alpha\zeta} I_{\zeta\zeta}^{-1} D'_{\alpha\zeta} - D_{\alpha\zeta} I_{\zeta\zeta}^{-1} I'_{\alpha\zeta} + I_{\alpha\zeta} I_{\zeta\zeta}^{-1} D_{\zeta\zeta} I_{\zeta\zeta}^{-1} I'_{\alpha\zeta} \Big|_{\substack{\alpha = \tilde{\alpha} = 0 \\ \zeta = \tilde{\zeta}}} \tag{7}$$

where,

$$D = \begin{pmatrix} D_{\alpha\alpha} & D_{\alpha\zeta} \\ D'_{\alpha\zeta} & D_{\alpha\alpha} \end{pmatrix} = \sum_{i=1}^{N} S_i(\gamma_i, x_{g,i}, x_{e,i}, \theta) S'_i(\gamma_i, x_{g,i}, x_{e,i}, \theta)$$

$$I = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\zeta} \\ I_{\alpha\zeta} & I_{\zeta\zeta} \end{pmatrix} = -\sum_{i=1}^{n} E\left[\frac{\partial S_i(\gamma_i, x_{g,i}, x_{e,i}, \theta)}{\partial \theta'} \right]$$

In appendix B, we show the above result.

With the above results, we can compute a global score statistic according to

$$T = S_{\alpha} V_{\alpha}^{-1} S_{\alpha} \tag{8}$$

The score statistic is distributed asymptotically as χ^2 with degrees of freedom equal to the rank of V_{α}.

Schaid *et al*[5] proved that the score function for α and the score function for haplotype probabilities are independent under the null hypothesis, so that the covariance between the two score functions is zero. Since the contribution to each MRHC is estimated from the haplotype frequency that is used to calculate the score statistic S_{α}, the variance of the score statistic is not penalized by the use of estimated haplotype frequencies.

In this framework, we can readily compute score statistics for each MRHC according to[5]:

$$z_k = S_{\alpha,k} / \sqrt{V_{\alpha,k,k}} \tag{9}$$

where z_k follows $\chi^2(1)$ under the null hypothesis H_0: $\alpha = 0$.

The P value P_0^B is assessed via simulation. In each replicate of this simulation, a sample is constructed in which the sample trait and environmental covariate of each individual are randomly permuted at the same time, and the score test statistic is computed again. Let T_i^B denote the value of the test statistic obtained for the ith replicate. Then P_0^B is the fraction of permutation replicates resulting in a test statistic greater than or equal to the test statistic of the real data, i.e., $P_0^B = \left| \{ i : T_i^B \geq T_0^B \} \right| / t$, with t denoting the number of permutation replicates and $| P_i^B = \left| \left\{ s : 0 \leq s \leq t, s \neq i, T_s^B \geq T_i^B \right\} \right| / t |$ denoting the number of elements of $\{ i : T_i^B \geq T_0^B \}$.

Testing more than one hypothesis

If we select m markers in several genes, there would be 2^m-1 marker combinations. To test 2^m-1 combinations with associated raw P values, and declare the global P value the significance level for our analysis would lead to another multiple-testing problem. In order to avoid nested simulation, we use the method which Becker and Knapp[2] adapted from Ge *et al*[13]. The basic idea is that, to test B = 2^m-1 marker combinations, global P is estimated by the proportion of permutation samples with min $_B P_t$ smaller than that in the observed data, where t is the simulation time. For each marker combination $B \in B$ and for each permutation replicate i = 1,..., t, the raw P value of the ith permutation replicate is calculated as

$$T = S_{\alpha} V_{\alpha}^{-1} S_{\alpha} \tag{10}$$

For $i > 0$, $P_i^{\min} := \min_{B \in \mathrm{B}} P_i^B$ is the minimum of the uncorrected P values over all MRHC in the ith permuta-

tion replicate. So the *P* value for the global hypothesis H_0 is calculated as:

$$P = \left| \left\{ s : 1 \le s \le t, P_i^{\min} \le P^{\min} \right\} \right| / t \qquad (11)$$

This permutation method is explained in more detail in[2].

Authors' contributions

YH developed the statistical model, carried out the software implementation, and made the simulation design and drafted the manuscript. JS helped with discussion both in theoretical developments and English copyediting. YP helped with discussion in theoretical developments, as well as in drafting the manuscript. QW, XZ and HZ contributed with discussion on theoretical aspects and drafting the manuscript. CL and LS contributed with the experimental data. All authors read and approved the manuscript.

Appendix A

Let $S_\alpha(Y, G, X_e, \alpha, \zeta)$ denote the score function of the data (Y, G, X_e) for α. As set forth by[37], $S_\alpha(Y, G, X_e, \alpha, \zeta)$ is the expectation of the complete-data score function given the observed data--that is,

$$\begin{aligned} S_\alpha(Y, G, X_e, \alpha, \zeta) &= \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \alpha} \log L(\alpha, \zeta; y_i, x_{g,i}, x_{e,i} \,|\, g_i \right] \\ &= \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \alpha} (\log f(y_i \,|\, x_{g,i}, x_{e,i}; \alpha, \zeta) + \log P(x_{g,i}) \,|\, g_i \right] \\ &= \sum_{i=1}^{n} E\left[\frac{y_i - b'(\eta)}{a(\varphi)} X_{g,i} \,|\, g_i \right] \\ &= \sum_{i=1}^{n} \frac{Y_i - E(y_i)}{a(\varphi)} E(X_{g,i} \,|\, g_i) \end{aligned}$$

Appendix B

For the expected Fisher information function of the observed data (*Y, G, Xe*), *I* is

$$I = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\beta} & I_{\alpha\varphi} \\ I'_{\alpha\beta} & I_{\beta\beta} & I_{\beta\varphi} \\ I'_{\alpha\varphi} & I'_{\beta\varphi} & I_{\varphi\varphi} \end{pmatrix},$$

where

$$I_{\alpha\varphi} = 0_{L\times 1},$$
$$I_{\beta\varphi} = 0_{(1+P)\times 1}$$

The hybrid estimate of *I* is obtained by replacing the nonzero entries of *I* with the observed Fisher information (denoted by *i*):

$$I = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\beta} & 0 \\ I'_{\alpha\beta} & I_{\beta\beta} & 0 \\ 0 & 0 & I_{\varphi\varphi} \end{pmatrix}.$$

Hence, equation (7) can be simplified as

$$V_\alpha = D_{\alpha\alpha} - i_{\alpha\beta} i_{\beta\beta}^{-1} D'_{\alpha\beta} - D_{\alpha\beta} i_{\beta\beta}^{-1} i'_{\alpha\beta} + i_{\alpha\beta} i_{\beta\beta}^{-1} D_{\beta\beta} i_{\beta\beta}^{-1} i'_{\alpha\beta}.$$

Recall that

$$D = \sum_{i=1}^{n} S_i(y_i, g_i, x_{e,i}, \theta) S_i'(y_i, g_i, x_{e,i}, \theta)$$

and that[37] proposed

$$S_i(y_i, g_i, x_{e,i}, \theta) = E\left[S_i(y_i, x_{g,i}, x_{e,i}, \theta \,|\, g_i \right]$$

so that

$$\begin{aligned} i &= \sum_{i=1}^{n} \left\{ E\left[-\frac{\partial S_i(y_i, x_{g,i}, x_{e,i}, \theta}{\partial \theta} \,|\, g_i \right] \right. \\ &\quad - E\left[S_i(y_i, x_{g,i}, x_{e,i}, \theta) S_i'(y_i, x_{g,i}, x_{e,i}, \theta) \,|\, g_i \right] \\ &\quad \left. + E\left[S_i(y_i, x_{g,i}, x_{e,i}, \theta) \,|\, g_i \right] E\left[S_i'(y_i, x_{g,i}, x_{e,i}, \theta) \,|\, g_i \right] \right\} \\ D_{\alpha\alpha} &= \sum_{i=1}^{n} \left(\frac{y_i - b'(\eta_i)}{a(\varphi)} \right)^2 E\left(x_{g,i} \,|\, g_i \right) E(x'_{g,i} \,|\, g_i) \\ D_{\alpha\beta} &= \sum_{i=1}^{n} \left(\frac{y_i - b'(\eta_i)}{a(\varphi)} \right)^2 E\left(x_{g,i} \,|\, g_i \right) x'_{e,i} \\ D_{\beta\beta} &= \sum_{i=1}^{n} \left(\frac{y_i - b'(\eta_i)}{a(\varphi)} \right)^2 x_{e,i} x'_{e,i} \\ i_{\alpha\beta} &= \sum_{i=1}^{n} \frac{b''(\eta)}{a(\varphi)} E(x_{g,i} \,|\, g_i) x'_{e,i} \\ i_{\beta\beta} &= \sum_{i=1}^{n} \frac{b''(\eta)}{a(\varphi)} x_{e,i} x'_{e,i} \end{aligned}$$

Acknowledgements

In this research, we had much help. During writing and running the program, Ning Xu, Haitao Wang, Bofei Xiao, etc. gave much help. Becker, Jing Li, Hongyu Zhao, etc. gave us many helpful suggestions in our study. This work is supported by the National Natural Science Foundation of China (grant no.30671492), the National High Technology Research and Development Program of China (863 project) (grant no. 2006AA10Z1E3, 2008AA101002) and the National 973 Key Basic Research Program (grant no, 2006CB102102, 2004CB117500).

References

1. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S: **A first-generation linkage disequilibrium map of human chromosome 22.** *Nature* 2002, **418(6897):**544-548.
2. Becker T, Knapp M: **A Powerful Strategy to Account for Multiple Testing in the Context of Haplotype Analysis.** *Am J Hum Genet* 2004, **75(4):**561-570.
3. Bell JT, Wallace C, Dobson R, Wiltshire S, Mein C, Pembroke J, Brown M, Clayton D, Samani N, Dominiczak A: **Two-dimensional genome-scan identifies novel epistatic loci for essential hypertension.** *Hum Mol Genet* 2006, **15(8):**1365-1374.
4. Carlborg O, Haley CS: **Epistasis: too often neglected in complex trait studies.** *Nat Rev Genet* 2004, **5(8):**618-625.
5. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: **Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous.** *Am J Hum Genet* 2002, **70(2):**425-434.
6. Tzeng JY, Wang CH, Kao JT, Hsiao CK: **Regression-Based Association Analysis with Clustered Haplotypes through Use of Genotypes.** *Am J Hum Genet* 2006, **78(2):**231-242.
7. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: **Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals.** *Hum Hered* 2002, **53(2):**79-91.
8. Cordell HJ, Barratt BJ, Clayton DG: **Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects.** *Genet Epidemiol* 2004, **26(3):**167-185.
9. Devlin B, Roeder K, Wasserman L: **Analysis of multilocus models of association.** *Genet Epidemiol* 2003, **25(1):**36-47.
10. Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex diseases.** *Nat Genet* 2005, **37(4):**413-417.
11. Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN: **Nonparametric Tests of Association of Multiple Genes with Human Disease.** *Am J Hum Genet* 2005, **76(5):**780-793.
12. Becker T, Schumacher J, Cichon S, Baur MP, Knapp M: **Haplotype Interaction Analysis of Unlinked Regions.** *Genet Epidemiol* 2005, **29(4):**313.
13. Ge Y, Dudoit S, Speed TP: **Resampling-based multiple testing for microarray data analysis.** *Test* 2003, **12(1):**1-77.
14. Manly BFJ: **Randomization, Bootstrap And Monte Carlo Methods in Biology.** New York: Chapman & Hall/CRC; 2007.
15. Hoh J, Wille A, Ott J: **Trimming, weighting, and grouping SNPs in human case-control association studies.** *Genome Res* 2001, **11(12):**2115-2119.
16. Becker T, Cichon S, Jonson E, Knapp M: **Multiple Testing in the Context of Haplotype Analysis Revisited: Application to Case-Control Data.** *Ann Hum Genet* 2005, **69(6):**747-756.
17. Becker T, Herold C: **Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power.** *Eur J Hum Genet* 2009, **17(8):**1043-1049.
18. Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B: **Analysis of single-locus tests to detect gene/disease associations.** *Genet Epidemiol* 2005, **28(3):**207-219.
19. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation.** *Bioinformatics* 2002, **18(2):**337-338.
20. Dekkers JCM: **Commercial application of marker-and gene-assisted selection in livestock: Strategies and lessons.** *J Anim Sci* 2004, **82:**E313-328.
21. Gerbens F: **Characterization, chromosomal localization, and genetic variation of the porcine heart fatty acid-binding protein gene.** *Mamm Genome* 1997, **8(5):**328-332.
22. Gerbens F, de Koning DJ, Harders FL, Meuwissen TH, Janss LL, Groenen MA, Veerkamp JH, Van Arendonk JA, te Pas MF: **The effect of adipocyte and heart fatty acid-binding protein genes on intramuscular fat and backfat content in Meishan crossbred pigs.** *J Anim Sci* 2000, **78(3):**552-559.
23. Gerbens F, van Erp AJ, Harders FL, Verburg FJ, Meuwissen TH, Veerkamp JH, te Pas MF: **Effect of genetic variants of the heart fatty acid-binding protein gene on intramuscular fat and performance traits in pigs.** *Am Soc Animal Sci* 1999, **77:**846-852.
24. Seeley RJ, Yagaloff KA, Fisher SL, Burn P, Thiele TE, van Dijk G, Baskin DG, Schwartz MW: **Melanocortin receptors in leptin effects.** *Nature* 1997, **390(6658):**349.
25. Kim KS, Larsen N, Short T, Plastow G, Rothschild MF: **A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits.** *Mamm Genome* 2000, **11(2):**131-135.
26. Foretz M, Guichard P, Ferre P, Foufelle F: **SREBP-1c is a major mediator of insulin action on the hepatic expression of gluckinase and lipogenesis related genes.** *Proc Natl Acad Sci USA* 1999, **96:**12737-12742.
27. Forsberg NE, Ilian MA, Ali-Bar A, Cheeke PR, Wehr NB: **Effects of cimaterol on rabbit growth and myofibrillar protein degradation and on calcium-dependent proteinase and calpastatin activities in skeletal muscle.** *J Anim Sci* 1986, **67(12):**3313-3321.
28. Murachi T, Tanaka K, Hatanaka M, Murakami T: **Intracellular Ca2+-dependent protease (calpain) and its high-molecular-weight endogenous inhibitor (calpastatin).** *Adv Enzyme Regul* 1980, **19:**407-424.
29. Li CL, Pan YC, Me H: **Polymorphism of the H-FABP, MC4R and ADD1 genes in the Meishan and four other pig populations in China.** *S Afr J Anim Sci* 2006, **36(1):**1.
30. Wang QS, Pan YC, Sun LB, Meng H: **Polymorphisms of the CAST gene in the Meishan and five other pig populations in China: short communication.** *S Afr J Anim Sci* 2007, **37(1):**27-30.
31. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ: **Estimation and Tests of Haplotype-Environment Interaction when Linkage Phase Is Ambiguous.** *Hum Hered* 2003, **55(1):**56-65.
32. Lobach I, Carroll RJ, Spinka C, Gail MH, Chatterjee N: **Haplotype-based regression analysis and inference of case-control studies with unphased genotypes and measurement errors in environmental exposures.** *Biometrics* 2008, **64(3):**673-684.
33. Zhou W, Thurston SW, Liu G, Xu LL, Miller DP, Wain JC, Lynch TJ, Su L, Christiani DC: **The Interaction between Microsomal Epoxide Hydrolase Polymorphisms and Cumulative Cigarette Smoking in Different Histological Subtypes of Lung Cancer 1.** *Cancer Epidemiol Biomarkers Prev* 2001, **10(5):**461-466.
34. Tzeng JY: **Evolutionary-based grouping of haplotypes in association analysis.** *Genet Epidemiol* 2005, **28(3):**220-231.
35. Patil N, Berno AJ, Hinds DA: **Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.** *Science* 2001, **294(5547):**1719-1723.
36. McCullagh P, Nelder JA: **Generalized Linear Models.** London: Chapman & Hall/CRC; 1989.
37. Louis TA: **Finding the observed information matrix when using the EM algorithm.** *J Roy Stat Soc B Stat Meth* 1982, **44(2):**226-233.