



## Estimation of Interaction Effects among Nucleotide Sequence Variants in Animal Genomes

Chaeyoung Lee\* and Younyoung Kim

Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea

**ABSTRACT** : Estimating genetic interaction effects in animal genomics would be one of the most challenging studies because the phenotypic variation for economically important traits might be largely explained by interaction effects among multiple nucleotide sequence variants under various environmental exposures. Genetic improvement of economic animals would be expected by understanding multi-locus genetic interaction effects associated with economic traits. Most analyses in animal breeding and genetics, however, have excluded the possibility of genetic interaction effects in their analytical models. This review discusses a historical estimation of the genetic interaction and difficulties in analyzing the interaction effects. Furthermore, two recently developed methods for assessing genetic interactions are introduced to animal genomics. One is the restricted partition method, as a nonparametric grouping-based approach, that iteratively utilizes grouping of genotypes with the smallest difference into a new group, and the other is the Bayesian method that draws inferences about the genetic interaction effects based on their marginal posterior distributions and attains the marginalization of the joint posterior distribution through Gibbs sampling as a Markov chain Monte Carlo. Further developing appropriate and efficient methods for assessing genetic interactions would be urgent to achieve accurate understanding of genetic architecture for complex traits of economic animals. (**Key Words** : Animal Genomics, Bayesian Inference, Epistasis, Gibbs Sampling, Single Nucleotide Polymorphism)

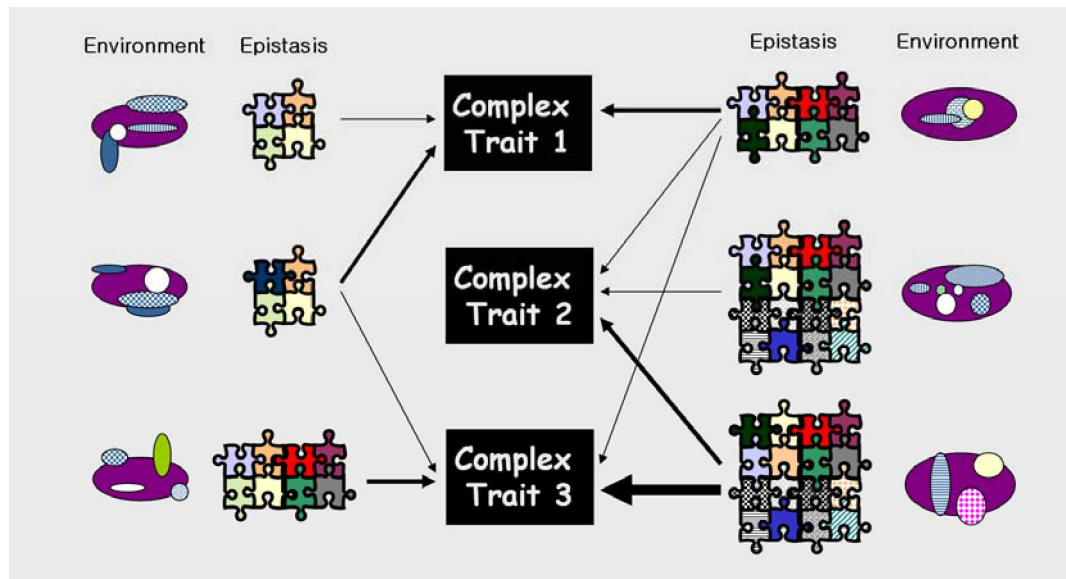
### INTRODUCTION

Providing comprehensive maps of nucleotide sequence variants in various species has been a great concern for many geneticists. As a result, genomic properties started to be partially revealed at least on the aspect of their compositions. Two independent human genomes are known to be roughly 99.9% identical, and only a small portion has variability (Kruglyak and Nickerson, 2001). Nevertheless, at least millions of variants are available, and these variants account for all the heritable phenotypic variability among individuals. Many efforts have been made to find genetic factors susceptible to complex diseases in humans, and substantial advances have been achieved in understanding the genetic dissection of complex traits of biomedical importance (McCarthy et al., 2008). Geneticists expect that such findings in human genomes may apply to other animals although their genome projects are still in the working.

One of the main goals in animal breeding and genetics is identifying the relationship of the nucleotide sequence variants with economically important phenotypes in order to select genetically superior animals whose genetic resources would be inherited to the next generation. The genetic architecture of the economic traits is quite limitedly known because of the difficulty in estimating the influence of multiple genes on such complex traits. The phenotypic variability for the complex traits might be largely explained by interactions among multiple genes under various environmental exposures (Figure 1). A genetic dissection of complex trait needs more extensive views of biology and more systematic approaches in genomic analysis. The potential interaction effects have not been analyzed in many genetic studies of complex traits because of the increasing number of genetic interaction parameters (Frankel and Schork, 1996). Therefore, the assumptions on the independence of the individual locus effects in analytical models might lead to wrong inferences on the relationship between genetic effects and phenotypic observations. It is timely to consider the next step for investigating genomewide association with complex traits. This review

\* Corresponding Author: Chaeyoung Lee. Tel: +82-2-820-0455, Fax: +82-2-824-4383, E-mail: clee@ssu.ac.kr

Received June 1, 2008; Accepted October 5, 2008



**Figure 1.** Influences of genetic interaction effects on complex traits under various environmental exposures. Puzzle piece indicates gene, and circles with different shapes, sizes, and colors indicate various environments. Thickness of arrow indicates degree of impact on the complex traits.

discusses the historical estimation of genetic interaction and difficulties in analyzing the interaction effects and introduces recently developed methods for assessing genetic interaction to animal genomics.









#### A HISTORICAL LOOK AT ESTIMATING GENE INTERACTION

Various concepts of gene interaction or epistasis have been used in quantitative genetics, and its definition was recently extended even to the interaction between different genes each from a different individual (Wolf, 2000). One example of this genome-by-genome interaction might be the regulations genetically coordinated by maternal, embryonic, and endospermic tissues in a developing seed (Walbot and Evans, 2003). In this article, we are, however, focusing on a classical meaning of genetic interaction that a genotypic effect at a gene is influenced by another genotype at another gene on the same genome (Falconer and Mackay, 1996). Another important concept of the genetic interaction here is not individual functional epistasis, but the population stochastic epistasis (Moore and Williams, 2005). The genetic interaction effects were statistically introduced by Fisher (1918) by decomposing genetic variance into additive, dominance, and epistatic variances. Then, many statistical geneticists have treated the epistatic effects as interaction terms in a regression on allelic effects and expanded to specific situations in their analytical models (Cockerham, 1954; Hansen and Wagner, 2001). These conventional genetic interaction models worked reasonably with at least a limited number of genetic variants (2-3).

#### PROBLEMS IN ESTIMATING GENE INTERACTION

Statistical modeling of genetic interaction becomes quite difficult as the number of genetic loci is increased. First of all, if we make assumption on the specific way of genetic interaction, this assumption can be inappropriate for many genetic interaction analyses because genes interact in a variety of ways. One of the difficulties lies on a large number of parameters to accommodate the various forms of genetic interactions. The increased number of parameters leads to the increased number of statistical tests and thus results in the increased number of spurious statistical significances. Various multiple comparison testing methods have been developed to reduce such false positives (Benjamini and Hochberg, 1995; Efron and Tibshirani, 2002). Although genetic interaction is statistically significant, a question arises if the genetic interaction is biologically meaningful. This is the inevitable question without any biological evidence.

Lastly and most importantly, a difficulty in estimating genetic interactions lies on sample size and statistical power. The amount of genotyping required might be reduced using a multistage discovery of nucleotide variants associated with complex traits, which maintained the statistical power of test (Hirschhorn and Daly, 2005). This strategy can be efficient for the discovery of individual locus effects, but a huge sample size is still required even in the initial stage. Another problem in practice was that most analyses have aimed to obtain the most parsimonious statistical model for genetic dissection of complex traits. This actually led to

No. of Groups		No. of Ways of Partitioning
2		255
3		3,025
4		7,770
5		6,951
6		2,646
7		462
8		36
9		1
Total (2~9)		21,146

**Figure 2.** Number of ways of partitioning 9 genotypes into 2 to 9 groups with an example of 2 nucleotide sequence variants in combinatorial partition method. Two alleles are assumed for each nucleotide sequence variant, hence there are 3 genotypes for each variant and 9 combined genotypes for 2 variants. Each cell indicates the combined genotype.

ignoring the potential genetic interaction effects in the genetic analysis, especially without single-locus additive and dominance effects (Carlborg and Haley, 2004). Another major problem has arisen with a dramatically increasing number of nucleotide sequence variants from genome projects. The classical epistatic model that included all the possible genetic interaction effects among multiple variants has shown a drawback of reduced degrees of freedom due to increased parameters for genetic interaction. This might lead to a potentially low power or a non-estimable statistic in analysis of genetic interaction. Solving or attenuating the problems addressed in this section has been major challenges for statistical geneticists, and as a result, recent advances in estimating genetic interaction effects were made possible.

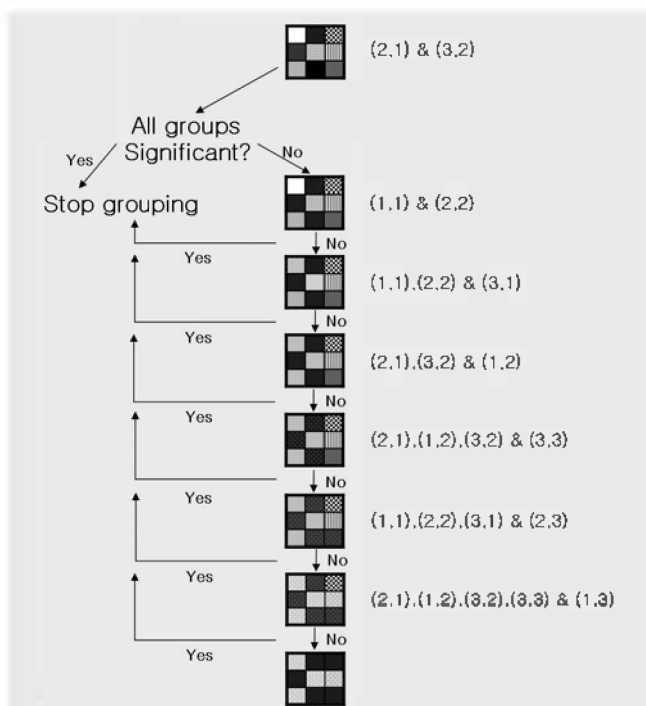
### PARTITIONING MULTI-LOCUS GENOTYPES

The methods for estimating genetic interaction effects were recently proposed by a nonparametric approach of grouping multi-locus genotypes to overcome the problems in the analysis with the conventional genetic interaction model. One of the methods by grouping multi-locus genotypes was called the combinatorial partition method (CPM). With CPM, subgroups of multi-locus genotypes that could explain phenotypic variability were identified by

evaluating all possible partitions (Nelson et al., 2001). The best genotypic partition was determined by iteratively evaluating the variability with partitioned subsets and then by cross validating genotypic partitions that explained a significant phenotypic variability. Although the CPM provides a good strategy for evaluating high-dimensional genetic interaction effects, this method has the disadvantage of computational burdens dramatically increased with a large number of nucleotide sequence variants the number of ways to partition  $\delta$  genotypes into  $\kappa$  groups can be calculated by the following formula for the Sterling's number of the second kind (Comtet, 1974).

$$S(\delta, \kappa) = \frac{1}{\kappa!} \sum_{i=0}^{\kappa-1} (-1)^i \binom{\kappa}{i} (\kappa - i)^\delta$$

This formula shows that tedious computations are required to obtain the possible partitions using CPM. For example, even with two loci as in Figure 2, the number of ways to partition 9 genotypes into 2 groups is 255, the number of ways to partition 9 genotypes into 3 groups is 3,025, the number of ways to partition 9 genotypes into 4 groups is 7,770, and so on. As a result, there are a total of 21,146 ways to partition the genotypes of only two loci into 2 to 9 groups. If we have 3 loci, then we need to evaluate



**Figure 3.** Algorithm of restricted partition method with an example of 2 nucleotide sequence variants. Two alleles are assumed for each nucleotide sequence variant, hence there are 3 genotypes for each variant and 9 combined genotypes for 2 variants. Each cell indicates the combined genotype. Every round the two groups with the smallest difference are combined into a new group until all the pairwise groups are significant. The two figures in parenthesis indicate the row and column numbers in a cell, and the cells presented in the right side are included in the pair of groups with the smallest difference.

63,438 with only 2-locus model and more than  $10^{21}$  with additional 3-locus model. This clearly demonstrates that even with 3 loci, evaluating genetic interaction effects with CPM requires too exhaustive computing.

A modified method called the restricted partition method (RPM) was developed to reduce the exhaustive computing time for searching the best among all the possible genotypic partitions in the CPM. The RPM was designed also to find partitions of multilocus genotypes that explained a significant proportion of the phenotypic variation, but it restricted its search to avoid evaluation of partitions that would not explain much of the variation (Culverhouse et al., 2004). The best partition in this method was determined by iteratively comparing genotype groups by a multiple test and combining the pair with the smallest difference into a new group (Figure 3). All pair-wise significant differences of the groups were brought to a halt of the iteration in RPM. However, the RPM has some undesirable features produced by grouping genotypes although this algorithm dramatically reduces the computational burden from CPM. Iterations of grouping can produce a merged group in which genotypes with a

significant difference in the initial stage are placed. The 31% of simulated data showed at least one merged group that included significantly different genotypes (Lee and Park, 2007). Another undesirable feature of RPM is the other way around. Two genotypes initially without statistical significance can be split into two different groups. The study of Lee and Park (2007) revealed that 32% of simulated data was classified as the undesirable cases. Furthermore, they also showed this undesirable pattern in the real clinical data of obesity. The two genotypes (CCArgArg and CCTrpArg) of  $\beta_2$ -adrenergic receptor gene (*ADRB2*) and  $\beta_3$ -adrenergic receptor gene (*ADRB3*) were separated into the risk and protective genotype groups ( $p < 0.05$ ) in spite of their corresponding initial phenotypic means ( $p > 0.05$ ). Such false positives or false negatives are more likely to be increased without a plausible biological explanation of grouping when applying the partitioning-based estimation of genetic interaction effects.

### A BAYESIAN METHOD USING GIBBS SAMPLING

Unclear biological explanation on grouping multi-locus genotypes in CPM or RPM led to skepticism about the plausibility of the grouping-based algorithm, which guided back to a parametric method for explaining genetic interaction effects. More recently, a Bayesian approach was introduced to estimating genetic interaction parameters (Lee and Park, 2007), which was originated from the animal breeding context of Bayesian inference (Lee, 2000). In this method, inferences about unknown genetic interaction effects are based on their marginal posterior distribution in a Bayesian framework. The marginalization of the joint posterior distribution is attained through Gibbs sampling that is a numerical integration method based on a Markov chain Monte Carlo (Tanner, 1993).

They first derived a general formula for the joint posterior distribution of all parameters using the Bayes theorem. Inverse Gamma distributions were assumed for the priors of variance components for both genetic interaction effects and residuals because the use of flat priors for variance components might lead to inferences based on theoretically nonexistent posterior distributions (Hobert and Casella, 1996). Full conditional posterior distribution was subsequently derived by obtaining the posterior distribution of each parameter given the data and all other parameters. The full conditional distribution for a scalar genetic interaction effect is expressed as the following Normal distribution:

$$g_j | \tau_j, \sigma_g^2, \sigma_e^2, \sigma_{g_1}^2, \dots, \sigma_{g_m}^2, \sigma_{g_{m+1}}^2, \dots, \sigma_{g_{m+n}}^2 \sim N \left( \frac{\sum_i (y_{ij} - \tau_j) \cdot \frac{\sigma_e^2}{\sigma_g^2}}{\sum n_{ij} + \frac{\sigma_e^2}{\sigma_g^2}}, \frac{\sigma_e^2}{\sum n_{ij} + \frac{\sigma_e^2}{\sigma_g^2}} \right)$$

The  $g_j$  is the  $j^{\text{th}}$  genetic interaction effect,  $\tau$ , is a non-genetic fixed effect,  $y_{ijk}$  is a phenotypic value,  $N$  indicates Normal distribution,  $\sigma_g^2$  is genetic interaction variance, and  $\sigma_r^2$  is residual variance. The full conditional distribution of the corresponding genetic interaction variance component is as an Inverse Gamma distribution:

$$\sigma_g^2 | g_1, \dots, g_{n_g} \sim IG \left[ \frac{n_g}{2} + \alpha_g, \frac{1}{\frac{1}{2} \sum_j g_j^2 + \frac{1}{\gamma_g}} \right]$$

The *IG* indicates Inverse Gamma distribution,  $\alpha_g$  is the shape parameter for genetic interaction variance component, and  $\gamma_g$  is the scale parameter for genetic interaction variance component.

For Gibbs sampling, an intensive iteration is required to generate samples using the consecutively updated full conditional distribution of the parameters. The initially generated samples are discarded until their convergence is determined, and then samples are selected at a regular interval to reduce a correlation between the consecutive samples. The posterior mean of the genetic interaction effects is recommended to be estimated based on the optimum Bayes decision rule under quadratic loss.

This Bayesian method using Gibbs sampling can be structured more in details with some specific analytical models. An example was presented in the previous study of Lee and Park (2007) where two-locus interaction model was applied to simulated data with a variety of designs. They first named the method mixed model with Gibbs sampling (MMGS), but this approach may not require to be derived in a mixed model framework. Later, it was called Bayesian approach using Gibbs sampling (BAGS). The BAGS showed a smaller prediction error for their simulated data than the grouping-based method, RPM. The larger prediction error produced by RPM might be mainly explained by losing information in grouping genotypes. This simulation study suggested that BAGS might be superior in estimating genetic interaction effects to such nonparametric partitioning approach. Furthermore, they discussed lack of biological explanation for the grouping in terms of information loss produced by merging two different genotypes into one group. Thus the grouping-based methods should be used with caution in that the information loss due to grouping has negligible effect, and justifiable biological explanation for the grouping is available. Otherwise, inferences on genetic interactions using RPM would not help determine whether their results would have viable implication to biological genetic interaction.

One of the major concerns for dealing with genetic

interaction effects is statistical power and corresponding sample size. A simulation study showed that BAGS considerably increased powers when interaction effects were tested with 2 loci comparing to the RPM (Lee and Park, 2007). Such inferior characteristic of RPM was caused mainly by grouping genotypes in the algorithm. Addition of loci would even make a larger difference in the statistical power because increased number of genotypes facilitates grouping.

For users of BAGS, Lee and Kim (2008) provided practical guidelines for determining an optimal sample size with a given statistical power and for calculating statistical power with a given sample size. They suggested a simple practical usage of the estimates using four scenarios. The two scenarios would be utilized with a known sample size, and the others with an unknown sample size. When the sample size is known, statistical power estimates can be obtained across heritability for 2-, 3-, and 4-locus balanced and unbalanced designs. If we further know the heritability, then specific values for the power can be provided. When the sample size is unknown or flexible, we can get an optimal sample size across heritability with a given statistical power. If heritability is further known, then a specific value of sample size can be provided.

We assume to apply the method to a genomewide association study with one million sequence variants and to find an optimal sample size with the power of at least 0.8 in an unbalanced data in order to find interaction effects among 4 loci. Note that we use the term interaction instead of epistasis because interaction among the sequence variants in one gene can be also easily explained with a strong linkage. Optimal sample sizes suggested by Lee and Kim (2008) were 810, 1,620, and 4,050, respectively, with heritabilities of 0.5, 0.33, and 0.28.

## CONCLUDING REMARKS

Now, we are confronting dramatically increasing markers resulted from animal genome projects, and such numerous data on genetic markers should be utilized to understand genetic architecture of their economic traits. Currently, genetic association studies for major livestock have been restricted to candidate gene analysis, and the association resulted from candidate gene studies were at most vague or contradictory in cattle (Kim et al., 2005; Cheong et al., 2008; Dario et al., 2008), swine (Li et al., 2007; Chen et al., 2008; Omelka et al., 2008), and chickens (Wang et al., 2006; Wang et al., 2007; Zhang et al., 2008). Recently, a genomewide association analysis was reported for cattle (Charlier et al., 2008). In the near future, the candidate gene association study would have a dramatical shift to the genomewide association study. Consequently, development of appropriate and efficient methods to assess

genetic interactions would be an urgent task to achieve essential understanding of their genetic architecture. Ultimately, investigations at a molecular level would offer an answer to mounting questions on true biological genetic interaction.

### ACKNOWLEDGMENTS

This study was supported by a grant (20080401034021) from BioGreen 21 Program, Rural Development Administration, Republic of Korea, and the Soongsil University Research Fund.

### REFERENCES

- Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289-300.
- Carlborg, O. and C. S. Haley. 2004. Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* 5:618-625.
- Charlier, C., W. Coppiegers, F. Rollin, D. Desmecht, J. S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J. C. Frennet, R. Hanset, X. Hubin, C. Jorgensen, L. Karim, M. Kent, K. Harvey, B. R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R. J. Harvey and M. Georges. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat. Genet.* 40:449-454.
- Chen, J. F., L. H. Dai, J. Peng, J. L. Li, R. Zheng, B. Zuo, F. E. Li, M. Liu, K. Yue, M. G. Lei, Y. Z. Xiong, C. Y. Deng and S. W. Jiang. 2008. New evidence of alleles (V199I and G52S) at the PRKAG3 (RN) locus affecting pork meat quality. *Asian-Aust. J. Anim. Sci.* 21:471-477.
- Cheong, H. S., D. H. Yoon, L. H. Kim, B. L. Park, H. W. Lee, S. Namgoong, E. M. Kim, E. R. Chung, I. C. Cheong and H. D. Shin. 2008. Association analysis between insulin-like growth factor binding protein 3 (IGFBP3) polymorphisms and carcass traits in cattle. *Asian-Aust. J. Anim. Sci.* 21:309-313.
- Cockerham, C. C. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39:859-882.
- Comtet, L. 1974. *Advanced combinatorics: the art of infinite expansions*. Boston: Reidel.
- Culverhouse, R., T. Klein and W. Shannon. 2004. Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.* 27:141-152.
- Dario, C., D. Carnicella, F. Ciotola, V. Peretti and G. Bufano. 2008. Polymorphism of growth hormone GH1-Ah1 in Jersey cows and its effect on milk yield and composition. *Asian-Aust. J. Anim. Sci.* 21:1-5.
- Efron, B. and R. Tibshirani. 2002. Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* 23:70-86.
- Falconer, D. S. and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics* (4<sup>th</sup> ed). Longmans Green, Harlow, Essex, UK.
- Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* 3:399-433.
- Frankel, W. N. and N. J. Schork. 1996. Who's afraid of epistasis? *Nat. Genet.* 14:371-373.
- Hansen, T. F. and G. P. Wagner. 2001. Modeling genetic architecture: A multilinear theory of gene interaction. *Theor. Popul. Biol.* 59:61-86.
- Hirschhorn, J. N. and M. J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6:95-108.
- Hobert, J. P. and G. Casella. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Am. Stat. Assoc.* 91:1461-1473.
- Kim, J. B., Z. X. Zeng, Y. J. Nam, Y. Kim, S. L. Yang, X. Wu and C. Lee. 2005. Association of mahogany/atractin gene (ATRN) with porcine growth and fat. *Asian-Aust. J. Anim. Sci.* 18:1383-1386.
- Kruglyak, L. and D. A. Nickerson. 2001. Variation is the spice of life. *Nat. Genet.* 27:234-236.
- Lee, C. 2000. Methods and techniques for variance component estimation in animal breeding. *Asian-Aust. J. Anim. Sci.* 13:413-422.
- Lee, C. and J. Park. 2007. Estimation of epistasis among finite polygenic loci for complex traits with a mixed model using Gibbs sampling. *J. Biomed. Inform.* 40:500-506.
- Lee, C. and Y. Kim. 2008. Optimal designs for estimating and testing interaction among multiple loci in complex traits by a Gibbs sampler. *Genomics* 92:446-451.
- Li, X. L., W. L. He, C. Y. Deng and Y. Z. Xiong. 2007. Associations of polymorphisms in the Mx1 gene with immunity traits in Large White X Meishan F<sub>2</sub> offspring. *Asian-Aust. J. Anim. Sci.* 20:1651-1654.
- McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis and J. N. Hirschhorn. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9:356-369.
- Moore, J. H. and S. M. Williams. 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 27:637-646.
- Nelson, M. R., S. L. Kardina, R. E. Ferrell and C. F. Sing. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* 11:458-470.
- Omelka, R., M. Martiniakova, D. Peskovicova and M. Bauerova. 2008. Associations between Alu I polymorphisms in the prolactin receptor gene and reproductive traits of Slovak Large White, White Meaty and Landrace pigs. *Asian-Aust. J. Anim. Sci.* 21:484-488.
- Tanner, M. A. 1993. *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*. Springer Series in Statistics, New York, NY, USA.
- Walbot, W. and N. M. S. Evans. 2003. Unique features of the plant life cycle and their consequences. *Nat. Rev. Genet.* 4:369-379.
- Wang, Y., H. Li, Y. D. Zhang, Z. L. Gu, Z. H. Li and Q. G. Wang. 2006. Analysis on association of a SNP in the chicken OBR gene with growth and body composition traits. *Asian-Aust. J. Anim. Sci.* 19:1706-1710.

- Wang, Y., D. Shu, L. Li, H. Qu, C. Yang and Q. Zhu. 2007. Identification of single nucleotide polymorphism of H-FABP gene and its association with fatness traits in chickens. *Asian-Aust. J. Anim. Sci.* 20:1812-1819.
- Wolf, J. B. 2000. Gene interactions from maternal effects. *Evolution* 54:1882-1898.
- Zhang, N. B., H. Tang, L. Kang, Y. H. Ma, D. G. Cao, Y. Lu, M. Hou and Y. L. Jiang. 2008. Associations of single nucleotide polymorphisms in BMPR-IB gene with egg production in a synthetic broiler line. *Asian-Aust. J. Anim. Sci.* 21:628-632.