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
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).

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.

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
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).

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

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

<table>
<thead>
<tr>
<th>No. of Groups</th>
<th>No. of Ways of Partitioning</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>255</td>
</tr>
<tr>
<td>3</td>
<td>3,025</td>
</tr>
<tr>
<td>4</td>
<td>7,770</td>
</tr>
<tr>
<td>5</td>
<td>6,951</td>
</tr>
<tr>
<td>6</td>
<td>2,646</td>
</tr>
<tr>
<td>7</td>
<td>462</td>
</tr>
<tr>
<td>8</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total (2–9)**

21,146
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 β2-adrenergic receptor gene (ADRB2) and β3-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:

$$\mathcal{N}\left(\frac{\sum a_i S_i^2 + \sum a_i^2\hat{\tau}_i}{\sum S_i^2 + \sum a_i^2}, \frac{1}{\sum S_i^2 + \sum a_i^2}\right)$$
The $g_{ij}$ is the $j^{th}$ genetic interaction effect, $\tau$ is a non-genetic fixed effect, $y_{ijk}$ is a phenotypic value, $N$ indicates Normal distribution, $\sigma^2_{genetic}$ is genetic interaction variance, and $\sigma^2$ is residual variance. The full conditional distribution of the corresponding genetic interaction variance component is as an Inverse Gamma distribution:

$$\sigma^2_{j1,...,j_n} \sim IG \left( \frac{n_j}{2} + \alpha_g, \frac{1}{2\sum g^2_j + \frac{1}{\gamma_g}} \right)$$

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


