Asian-Aust. J. Anim. Sci.
Vol. 25, No. 11 : 1511-1514 November 2012
http://dx.doi.org/10.5713/ajas.2012.12133

www.ajas.info
pISSN 1011-2367 eISSN 1976-5517

Empirical Statistical Power for Testing Multilocus Genotypic Effects under Unbalanced Designs Using a Gibbs Sampler

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

ABSTRACT: Epistasis that may explain a large portion of the phenotypic variation for complex economic traits of animals has been ignored in many genetic association studies. A Baysian method was introduced to draw inferences about multilocus genotypic effects based on their marginal posterior distributions by a Gibbs sampler. A simulation study was conducted to provide statistical powers under various unbalanced designs by using this method. Data were simulated by combined designs of number of loci, within genotype variance, and sample size in unbalanced designs with or without null combined genotype cells. Mean empirical statistical power was estimated for testing posterior mean estimate of combined genotype effect. A practical example for obtaining empirical statistical power estimates with a given sample size was provided under unbalanced designs. The empirical statistical powers would be useful for determining an optimal design when interactive associations of multiple loci with complex phenotypes were examined. (**Key Words:** Bayesian, Epistasis, Genetic Association, Gibbs Sampling, Statistical Power)

INTRODUCTION

Genetic architecture for complex economic traits of animals might be understood based on accurate estimates of interaction effects. However, the most parsimonious statistical models have been suggested in many analyses for genetic dissection of complex traits and the potential interaction effects were excluded in analytical models (Frankel and Schork, 1996).

The classical epistatic model included all the possible genetic interaction effects among multiple loci. For example, a two-locus epistatic model included genetic interaction effect between locus 1 and locus 2 (I_{12}), and a three-locus epistatic model included genetic interaction effects among locus 1, locus 2, and locus 3 (I_{12}, I_{13}, I_{23}, I_{123}). This led to a drawback of drastically reduced degrees of freedom with an increased number of loci. A restricted partition method as a nonparametric approach was recently developed for estimating epistasis, and it overcame the problem in the conventional epistatic analysis (Culverhouse et al., 2004).

More recently, a Bayesian approach using Gibbs sampling was proposed to overcome the shortage of degrees of freedom by treating the epistatic effects as random effects (Lee and Park, 2007). This parametric method dramatically reduced predictional errors in estimating interaction effects compared to the restricted partition method. A guideline was provided for experimental designs under various situations when conducting a genetic association study with multi-locus interaction effects by the Bayesian approach with a Gibbs sampler (Lee and Kim, 2008). The simulation study for experimental designs was conducted to examine the accuracy of predicting the interaction effects and to estimate the corresponding statistical power, the probability of accepting true interaction effects, by the method. The degree of unbalance was, however, not considered in the study.

In reality, the genetic data are most likely unbalanced. Furthermore, null combined genotype cells, i.e. combined genotypes with no observation in a multi-locus model, increase as the number of loci increases. In the current study, we conducted a simulation study to show empirical power and sample size for the use of the Bayesian method by Gibbs sampling under practical situations with a variety of unbalanced data including null combined genotype cells. Ultimately, the results will help to determine optimal designs for identifying genetic variants associated with complex traits.

* Corresponding Author: C. Lee. Tel: +82-2-820-0455, Fax: +82-2-824-4383, E-mail: clee@ssu.ac.kr
Submitted Mar. 8, 2012; Accepted May 5, 2012; Revised May 29, 2012

MATERIALS AND METHODS

Simulation

A Monte Carlo simulation was conducted to generate unbalanced data with null combined genotype cells. Phenotype assuming 2-locus model to 5-locus model was generated as follows: $y_{ijk} = a_i+g_j+e_k$ where y_{ijk} is k^{th} phenotypic value within fixed effect i and random effect j, a_i is i^{th} fixed environmental effect, g_j is j^{th} random combined genotype effect, and e_k is k^{th} random error. The combined genotype effects assigned to the corresponding 9, 27, 81 and 243 genotypes for 2-, 3-, 4-, and 5- locus models were generated from the Normal distribution with the variance of 10, i.e. $g_j \sim N(0,10)$ where $j = 1,\ldots 3^m$, m is the number of loci (= 2, 3, 4, or 5). The error was also generated from the Normal distribution with the variance ranged from 10 to 40. Simulation was devised under various unbalanced designs (mild, medium, and strong). Their average sample size for each genotype was 5, 10, 15, …, or 100. Portions of null combined genotype cells ranged from 0 to 50% with an increment of 5%. A total of 42,240 data sets were simulated from combinations of number of loci (4 levels), variance within genotype (16 levels), sample size (20 levels), degree of unbalance (3 levels), and portions of null combined genotype cells (11 levels). One hundred replicates were simulated for each set. A random number generator based on the Box-Muller method was used to generate random Gaussian deviates (Press et al., 1992).

Analytical method

The simulated data were analyzed by the Bayesian method by Gibbs sampling to estimate genetic parameters in multilocus epistatic models. This method was devised to draw inferences about the epistatic effects based on their marginal posterior distributions and to attain the marginalization of the joint posterior distribution through Gibbs sampling (Lee and Park, 2007). We conducted the Gibbs sampling by intensive iterations of sampling from full conditional posterior distributions as follows:

i) Set arbitrary initial values for fixed effects (**a**), random genotype effects (**g**), genotypic variance component (σ_g^2), and residual variance component (σ_e^2).

ii) Generate and update residual variance component using the following full conditional posterior distribution.

$$\sigma_e^2 \left| \mathbf{a}, \mathbf{g}, \mathbf{y} \sim IG \left[\frac{n}{2} - \alpha_e, \frac{1}{\frac{1}{2}(\mathbf{y}-\mathbf{Xa}-\mathbf{Zg})'(\mathbf{y}-\mathbf{Xa}-\mathbf{Zg}) + \frac{1}{\gamma_e}} \right] \right. \quad (1)$$

where n is the number of phenotypic observations, σ_e^2 is residual variance component, **a** is the vector of fixed effects, **g** is the vector of random genotype effects, and **y** is the vector of phenotypes. IG [.] indicates inverse Gamma distribution, and γ_e and α_e are scale and shape parameters for the prior distribution of σ_e^2. **X** and **Z** are known design matrices relating the fixed and random effects to their corresponding phenotypes.

iii) Generate and update genotypic variance component using the following full conditional posterior distribution.

$$\sigma_g^2 \left| \mathbf{g} \sim IG \left[\frac{3^m}{2} + \alpha_g, \frac{1}{\frac{1}{2}\mathbf{g}'\mathbf{g} + \frac{1}{\gamma_g}} \right] \right. \quad (2)$$

where γ_g and α_g are scale and shape parameters for the prior distribution of σ_g^2.

iv) Generate and update random genotype effects using the following full conditional posterior distribution. For example, generate and update g_1, generate and update g_2, …, and generate and update g_{3^m}.

$$g_j \left| \mathbf{a}, \mathbf{g}_{-j}, \sigma_g^2, \sigma_e^2, \mathbf{y} \sim N \left(\frac{\sum_i \sum_k (y_{ijk} - a_i)}{\sum_i n_{ij} + \frac{\sigma_e^2}{\sigma_g^2}}, \frac{\sigma_e^2}{\sum_i n_{ij} + \frac{\sigma_e^2}{\sigma_g^2}} \right) \right. \quad (3)$$

where n_{ij} is the number of records within the fixed effect i and the genotype effect j.

v) Generate and update fixed effects using the following full conditional posterior distribution.

$$\mathbf{a} \left| \mathbf{g}, \sigma_e^2, \mathbf{y} \sim N\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{Zg}), (\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2 \right) \right. \quad (4)$$

vi) Repeat steps i) through v).

The Gibbs sampling was run a total of 52,000 iteration rounds. Samples generated from a warming-up period of the first 2,000 rounds were all removed to avoid a noise before convergence. Only samples at every 50 rounds after the warming-up period were retained to reduce lag correlation among the thinned samples.

RESULTS AND DISCUSSION

Statistical power estimates

Empirical statistical powers were estimated by testing genotypic difference from the unbalanced data simulated with 2 to 5 loci by the Bayesian method by Gibbs sampling. For example, the empirical statistical powers are presented for mildly (Figure 1A) and strongly (Figure 1B) unbalanced data. The power estimate obtained from the strongly

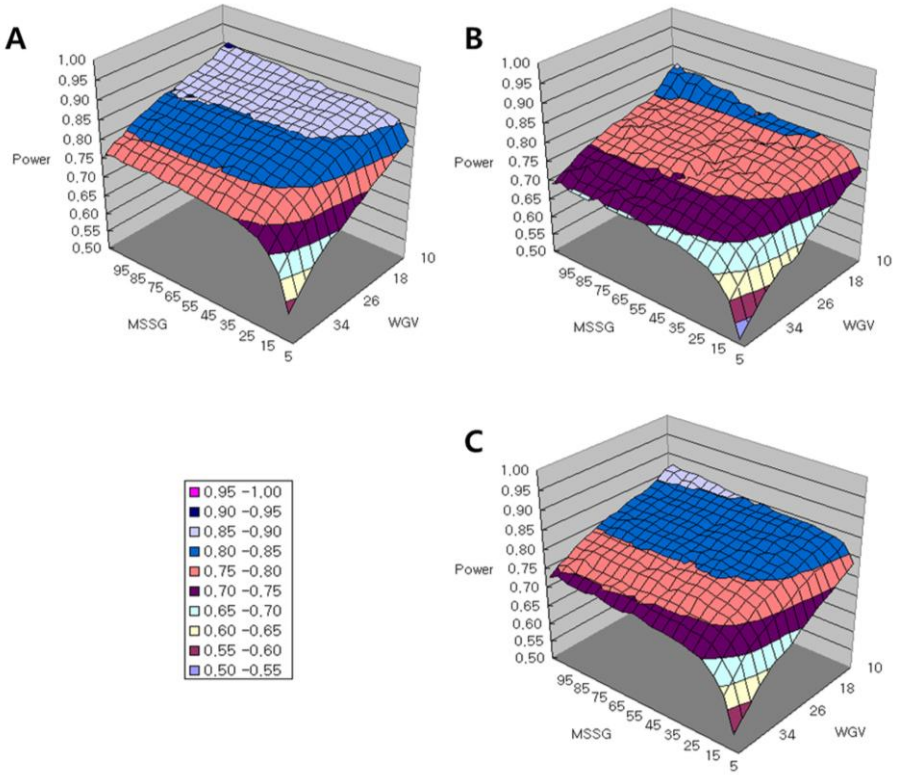

Figure 1. Empirical power for testing combined genotype difference from simulated data using (A) mildly unbalanced 4-locus design with 0% null combined genotype cells, (B) strongly unbalanced 4-locus design with 0% null combined genotype cells, and (C) strongly unbalanced 4-locus design with 50% null combined genotype cells. The power was estimated with the false positive probability of 0.05. WGV stands for within genotype variance, and MSSG stands for mean sample size for genotype.

unbalanced data was smaller than the corresponding estimate from the mildly unbalanced data regardless of the sample size, the number of loci, and within genotype variance. Of course, power estimates obtained from the balanced data by Lee and Kim (2008) were closer to those from the mildly unbalanced data than those from the strongly unbalanced data. The power estimates increased with a reduced number of loci or with a reduced within genotype variance. This concurred with results of the study in which powers were estimated under balanced designs (Lee and Kim, 2008). The frequency of null combined genotype cells influenced the statistical power estimates. If half of the combined genotype cells were null, then the power increased as shown in Figure 1C.

The mean empirical statistical powers estimated in the current study highlight the heterogeneous statistical powers in unbalanced designs under a range of plausible scenarios. This critical influence of the unbalanced designs on statistical power could be strengthened by the study of Wang et al. (2012) where the identification of genetic association with heading date of barley depended on the degree of balance. The estimates obtained in the current study might be applied to finding an optimal design for estimating and testing multi-locus interaction effects. The sample size and the number of loci would be important components affecting the statistical power in practice. This study also suggested the Baysian method using a Gibbs sampler for testing epistatic effects among limited number (up to 4) of loci.

A practical example

Consider a practical example for experimenters to apply the statistical power estimates in an epistatic association study. Let's assume a research plan to determine a statistical power with the given sample size of 800. Assuming a heritability of 0.25, the power is predicted as 0.83 for mildly unbalanced data and as 0.76 for strongly unbalanced data. They are obtained from Figure 1 and 2 because the sample size of 800 corresponds to MSSG = 89 and the heritability of 0.25 corresponds to WGV = 30. This practical guideline for determining the optimal sample size with a given power would be useful for population geneticists to apply the method in genetic association studies.

Concluding remarks

Epistasis that has been ignored in most genetic association studies might explain a large portion of genetic variation for complex traits. Furthermore, epistasis more clearly explains associations with individual variants (Chen

and Ishwaran, 2012). For example, a redundant epistasis might disclose a spurious effect produced by linkage or confounding among variants. Although estimates of epistatic effects do not entirely reflect biological interactions (He et al., 2010), an accurate estimation of the epistasis would be of value in identifying the genetic factors for complex traits. Efficient identification of genetic factors with epistasis greatly depends on determining power and the corresponding optimal sample size (Lee and Kim, 2008). Especially, this issue becomes more critical under unbalanced designs (Wang et al., 2012). Statistical powers and the practical guideline in the current study would be useful for determining the optimal sample size with a given power in genetic association studies. This would be helpful in meeting the need for increased sample size in genomewide association analysis, which is accelerating due to a rapid development of sequencing technology and an increase of variants in the analysis. However, the current study was restricted to some specific unbalanced designs. A generalized computer program would be in order for determining an optimal design by degree of unbalance and further for dealing with user-created designs.

ACKNOWLEDGEMENTS

This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (Grant No. 2009-0071063).

REFERENCES

Chen, X. and H. Ishwaran. 2012. Random forests for genomic data analysis. Genomics http://dx.doi.org/10.1016/j.ygeno.2012.04.003.

Culverhouse, R., T. Klein and W. Shannon. 2004. Detecting epistatic interactions contributing to quantitative traits. Genet. Epidemiol. 27:141-152.

Frankel, W. N. and N. J. Schork. 1996. Who's afraid of epistasis? Nat. Genet. 14:371-373.

He, X., W. Qian, Z. Wang, Y. Li and J. Zhang. 2010. Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. Nat. Genet. 42:272-276.

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.

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.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery. 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, UK.

Wang, H, K. P. Smith, E. Combs, T. Blake, R. D. Horsley and G. J. Muehlbauer. 2012. Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor. Appl. Genet. 124:111-124.