# OPTIMAL SAMPLE SIZE FOR TESTING MULTILOCUS GENOTYPIC EFFECTS BY A BAYESIAN METHOD USING A GIBBS SAMPLER

Jihye Ryu, Younyoung Kim, and Chaeyoung Lee*

*Department of Bioinformatics and Life Science, Soongsil University*
*511 Sangdo-dong, Dongjak-gu, Seoul 156-743, Republic of Korea*
**Email: clee@ssu.ac.kr*

The potential epistasis that may explain a large portion of the phenotypic variation for complex traits has been ignored in many genetic association studies. A Baysian method using a Gibbs sampler was introduced to draw inferences about multilocus genotypic effects based on their marginal posterior distributions by a Gibbs sampler. This method would be applied to studies for interaction effects among limited number of loci although theoretically they are applicable to all the possible interaction of millions of single nucleotide variants resulted from genomewide association study. A simulation study was conducted to provide an optimal sample size for experimental designs with this method. Data were simulated with more than 42,240 data sets produced by combined designs of number of loci (2, 3, 4, and 5 loci), within genotype variance (10 ~ 40, 16 levels), and sample size (5 ~ 100, 20 levels) in unbalanced designs with various portions of null genotypic cells (0 ~ 50%, 11 levels). Mean empirical statistical power was estimated for each data set in testing the posterior mean estimate of combination genotypic effect. Additionally, mean square prediction error was obtained from estimating the posterior mean estimate. The optimal sample sizes were provided with the prediction error > 2.0 and the statistical power >0.8 under various designs. The Baysian method using a Gibbs sampler was suggested for testing and estimating epistatic effects among limited number (2~4) of loci. Practical guidelines for determining the optimal sample size with a specific power are provided when population geneticists apply the Baysian method to their genetic association studies.

## 1. INTRODUCTION

Genetic architecture for complex traits 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.[1]

The classical epistatic model included all the possible genetic interaction effects among multiple loci, which led to a drawback of drastically reduced degrees of freedom. Restricted partition method (RPM) as a nonparametric approach was recently developed for estimating epistasis, and it overcame the problem in the conventional epistatic analysis.[2]

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.[3] This approach dramatically reduced predictional errors in estimating interaction effects comparing to RPM.[3] A guideline was provided for experimental designs under various situations when conducting genetic association study with multi-locus interaction effects by the Bayesian approach with a Gibbs sampler.[4] The simulation study for experimental designs was conducted to examine the accuracy of predicting the interaction effects and to estimate the corresponding statistical power by the method. The degree of balance was, however, quite limited in the study.

In reality, the genetic data are most likely unbalanced. Furthermore, null genotypic cells 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, and more practical guidelines are presented for unbalanced data including null genotypic cells.

## 2. METHODS

A Monte Carlo simulation was conducted to generate unbalanced data with null genotypic cells. Phenotype assuming 2-locus model to 5-locus model was generated by adding a environmental fixed effect, a genotypic mean, and an error. The genotypic means assigned to the corresponding 9, 27, 81 and 243 genotypes were generated from the Normal distribution with the variance of 10. 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.

* Corresponding author.

Portions of null genotypic 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 genotypic cells (11 levels). One hundred replicates were simulated for each set. A random number generator based on Box-Muller method was used to generate random Gaussian deviates.[5]

The simulated data were analyzed by the Bayesian method by Gibbs sampling to estimate genetic parameters in multilocus epistatic models.[3] 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.

## 3. RESULTS AND DISCUSSION

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 1) and strongly (Figure 2) unbalanced data. The power estimate obtained from the strongly 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. The power estimates increased with a reduced number of loci or with a reduced within genotype variance.

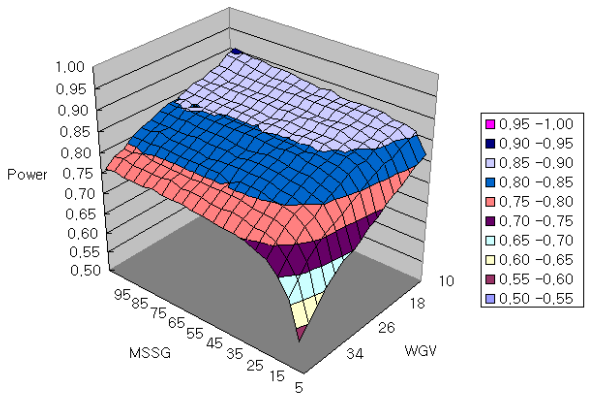


**Fig. 1.** Empirical statistical power for testing genotypic difference from the data simulated for mildly unbalanced 4-locus design with 0% null genotypic cells by the Bayesian method using Gibbs sampling. The power was estimated with the false positive probability of 0.05. WGV stands for within genotype variance, and MSSG stands for mean sample size for genotype.

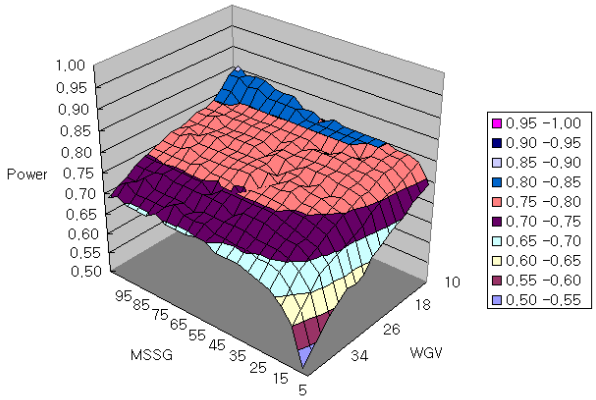


**Fig. 2.** Empirical statistical power for testing genotypic difference from the data simulated for strongly unbalanced 4-locus design with 0% null genotypic cells by the Bayesian method using Gibbs sampling. The power was estimated with the false positive probability of 0.05. WGV stands for within genotype variance, and MSSG stands for mean sample size for genotype.

The frequency of null genotypic cells influenced on the statistical power estimates. If half of the genotypic cells were null, then the power increased as shown in Figure 3.

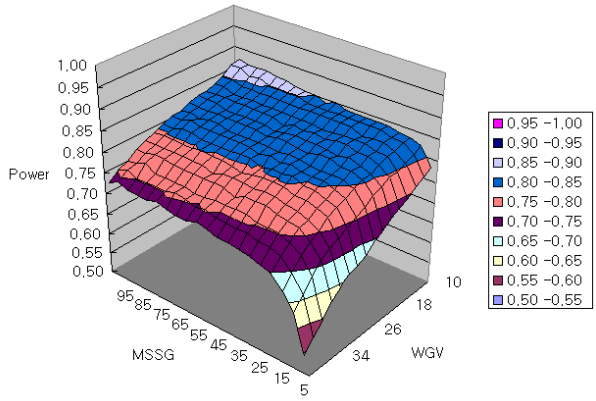


**Fig. 3.** Empirical statistical power for testing genotypic difference from the data simulated for strongly unbalanced 4-locus design with 50% null genotypic cells by the Bayesian method using Gibbs sampling. The power was estimated with the false positive probability of 0.05. WGV stands for within genotype variance, and MSSG stands for mean sample size for genotype.

The mean values estimated in this 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.

## Acknowledgments

This study was supported by a grant from BioGreen 21 Program, Rural Development Administration, Republic of Korea (Grant Code No. 20080401034021).

## References

1. Frankel WN, Schork NJ. Who's afraid of epistasis?, *Nat Genet* 1996; **14**: 371-373.
2. Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* 2004; **27**: 141-152.
3. Lee C, Park J. Estimation of epistasis among finite polygenic loci for complex traits with a mixed model using Gibbs sampling. *J Biomed Inform* 2007; **40**: 500-506.
4. Lee C, Kim Y. Optimal designs for estimating and testing interaction among multiple loci in complex traits by a Gibbs sampler. *Genomics* 2008; **92**: 446-451.
5. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, 1992.