Genomics & Informatics Vol. 7(3) 148-151, September 2009

The Effect of Increasing Control-to-case Ratio on Statistical Power in a Simulated Case-control SNP Association Study

Moonsu Kang, Sunhee Choi and InSong Koh*

Department of Physiology, College of Medicine, Hanyang University, Seoul 133-791, Korea

Abstract

Generally, larger sample size leads to a greater statistical power to detect a significant difference. We may increase the sample size for both case and control in order to obtain greater power. However, it is often the case that increasing sample size for case is not feasible for a variety of reasons. In order to look at change in power as the ratio of control to case varies (1:1 to 4:1), we conduct association tests with simulated data generated by PLINK. The simulated data consist of 50 disease SNPs and 300 non-disease SNPs and we compute powers for disease SNPs. Genetic Power Calculator was used for computing powers with varying the ratio of control to case (1:1, 2:1, 3:1, 4:1). In this study, we show that gains in statistical power resulting from increasing the ratio of control to case are substantial for the simulated data. Similar results might be expected for real data.

Keywords: association study, ratio of control to case, simulated data, SNP, statistical power

Introduction

The power of a study is the probability that the test will reject a null hypothesis that is in fact false. As power increases, the probability of a Type II error (false negative rate $=\beta$) decreases (Fig. 1). Therefore power is 1-β. Decreasing β error is equivalent to increasing statistical power (Fig. 2).

Power depends on several factors such as prevalence, magnitude of effect, sample size, and required level of statistical significance α. When computing statistical power in matched case-control studies (Dupont, 1988), we need to know a pre-specified type I error rate, the ratio of control to case, estimated number of cases, the prevalence of exposure in the control group, minimum odds ratio declared to be significant and correlation coefficient for exposure between cases and their matched controls. Hennessy S described the effect of increasing the ratio of control to case for different values of correlation coefficients and prevalence among controls in matched case-control studies (Hennessy S *et al.*, 1999). For a detailed review of power and sample size computation in either genetic studies or genetic epidemiology, please refer to Shork *et al.* (2002), Ambrosius *et al.* (2004), De La Vega *et al.* (2005), and Burton *et al.* (2009). In our study, we may focus on how sample size affects statistical power, given a set of population parameters.

Generally, increase in sample size for both case and control leads to increase in statistical power. There are some situations, however, where increasing sample size for case is not available. For example, in rare diseases, the cost of including additional controls is low whereas that of including cases is high. In such instances, we increase sample size for control only and then see if the effect on statistical power is the same as that obtained when the sample size for both case and control increases. Specifically, we examine if increase in the ratio of control to case has an effect on increasing power. We simulate SNP data as below and assess the effect of the ratio of control to case on statistical power.

Fig. 1. Type I error and Type II error.

		Decision (H_0)	
		Reject	Not reject
H_0	True	α (Type Ⅰ error, false positive)	1-α
	False (=H_1)	**1-β** **(Power)**	β (Type Ⅱ error, false negative)

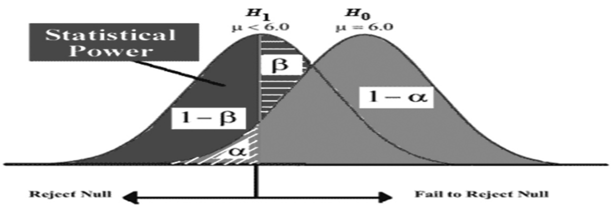

Fig. 2. Statistical power.

*Corresponding author: E-mail insong@hanyang.ac.kr
Tel +82-2-2220-0615, Fax +82-2-2281-3603

Accepted 1 September 2009

Table 1. Simulated data

	ID	null_1	null_2	⋯	null_300	disease_1	disease_2	⋯	disease_50
Case	1	d D	D D	⋯	d D	d D	d d	⋯	d d
	2	D D	d D	⋯	d D	d d	d d	⋯	D d
	3	d D	d D	⋯	D D	d D	d d	⋯	d d
	4	D D	D D	⋯	d d	d d	d d	⋯	d d
	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	50	D D	D D	⋯	D D	d D	d d	⋯	d d
Control	51	D D	D D	⋯	D D	d D	d d	⋯	d D
	52	D D	D D	⋯	D D	d d	D d	⋯	d d
	53	D D	D D	⋯	d D	d d	d d	⋯	D d
	54	d D	D D	⋯	D D	d D	d d	⋯	d d
	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	250	D D	d D	⋯	d d	d D	D d	⋯	D D

Null, non-disease SNP; Disease, disease SNP; d, minor allele; D, major allele.

Table 2. The number of significant SNP for each ratio of control-case ($p < 0.05$) [Allele model]

control : case		1:1 (50:50)	2:1 (100:50)	3:1 (150:50)	4:1 (200:50)
Significant SNP	Total	50	53	54	52
	Disease SNP	31	35	36	39
	Non-disease SNP	19	18	18	13

Methods

PLINK (Purcell *et al.*, 2007) was used for generating simulated data with 50 disease SNPs and 300 non-disease SNPs (Table 1). In this data, we fixed the sample size for case as 50 but the sample size for control size varies from 50, 100, 150 to 200 in order to investigate the effect of the ratio of control to case on power. We set prevalence as 0.5000, 0.3333, 0.2500, and 0.2000 for four models, respectively. First, assuming allele model, we computed the number of significant SNPs for disease SNPs, non-disease SNPs, and overall SNPs as the ratio of control to case changes. We also computed the estimated average power which is equal to *E(S)/m* (eq1), where S is the number of SNPs declared to be significant among disease SNPs and *m* is the number of disease SNPs. Second, we examined the power for each disease SNP using Genetic Power Calculator (Purcell S *et al.*, 2003) in order to see how change in the ratio of control to case affects the power for the genetic models (allele, genotype, dominant, and recessive).

Results and Discussion

Table 2 shows that the number of significant SNPs increases as the ratio of control to case for allele model rises from 1:1 to 3:1. The increase in the ratio of control to case for disease SNPs leads to increase in number of significant SNPs, while the increase in the ratio of control to case for non-disease SNPs leads to decrease in the number. Therefore, the number of significant SNPs decreases when the ratio of control to case increases from 3:1 to 4:1. We might expect that the gain in average power shown in eq1 increases as the ratio of control to case increases, since the effect of the ratio of control to case for disease SNPS on the number of significant SNPs is substantial. Assuming allele model, the average power shown in eq1 increases for disease SNPs as the ratio of control to case increases (Fig. 3). In other words, the power curve tends to increase gradually.

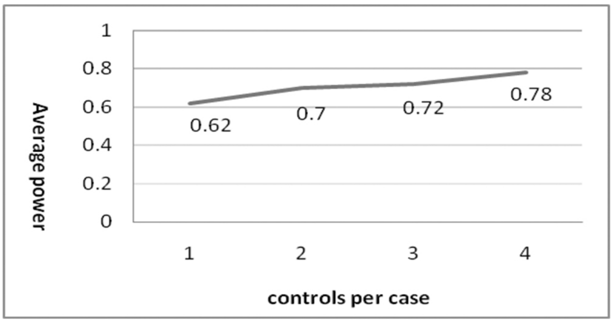

Fig. 3. Average power for disease SNP ($p < 0.05$) [Allele model].

On the other hand, in regard to computing statistical power for each disease SNP, for example, the disease SNP 3 (Fig. 4) is on the increase for all models. The disease SNP 17, the power increases for all models except dominant model (Fig. 5). The SNP 30 (Fig. 6) is on the increase for all models. The disease SNP 34 has the increasing pattern except recessive model (Fig. 7). The SNP 45 (Fig. 8) is on the increase for all models. We

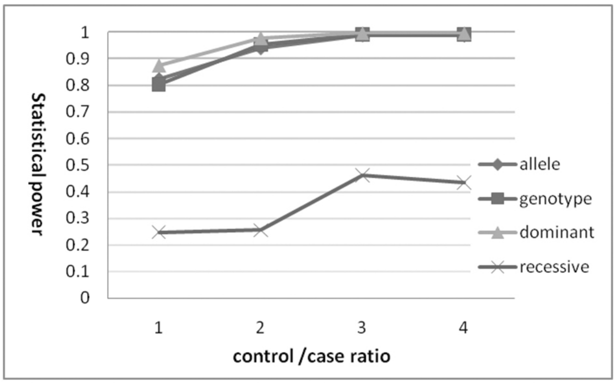

Fig. 4. Statistical power for Disease SNP 3.

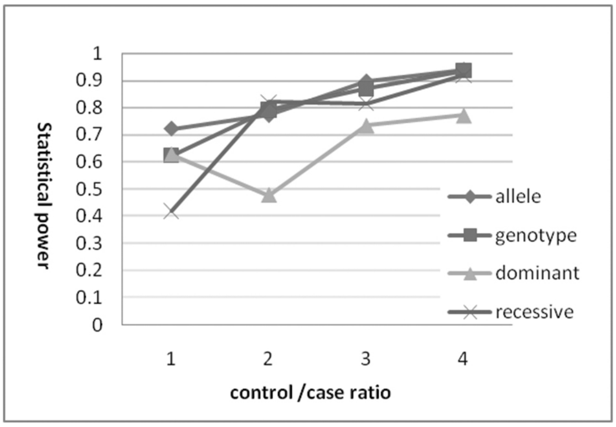

Fig. 5. Statistical power for Disease SNP 17.

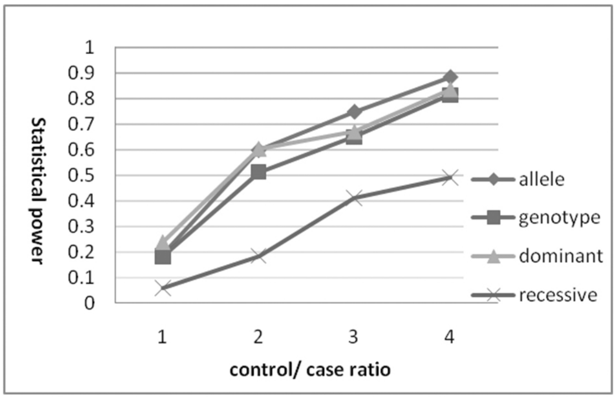

Fig. 6. Statistical power for Disease SNP 30.

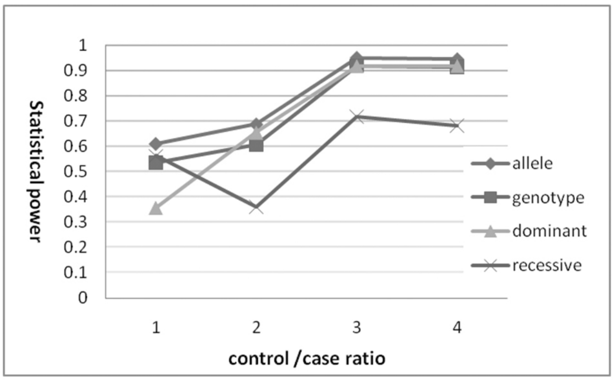

Fig. 7. Statistical power for Disease SNP 34.

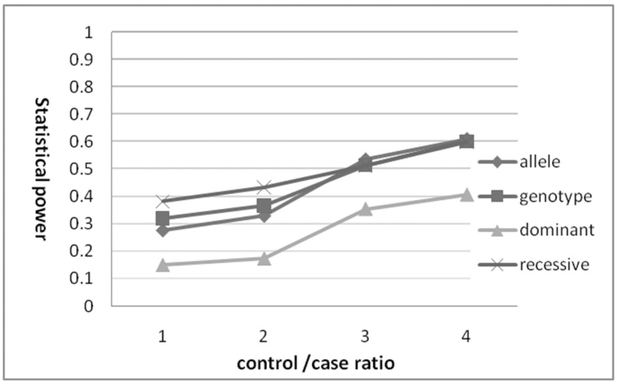

Fig. 8. Statistical power for Disease SNP 45.

use Chi-square test for testing allelic association and genotype analyses. We do not show all other disease SNPs in the paper but in general, statistical power for most disease SNPs is likely to increase by increase in the ratio of control to case. In summary, we show that given other factors such as prevalence, magnitude of effect, and required level of statistical significance α, significant increase in statistical power can be obtained by increasing the ratio of control to case. Henceforth, investigators conducting such a study where cases are limited might consider increase in the ratio of control to case. Further investigation may be needed for real data. And other factors which affect the power need to be considered.

Acknowledgements

This research was supported by grants from MOHWF, Korea. (00-PJ3-PG6-GN02-0002)

References

Ambrosius, W.T., Lange, E.M., and Langefeld, C.D. (2004). Power for genetic association studies with random allele frequencies and genotype distributions. *Am. J. Hum. Genet.* 74, 683-693.

Burton, P.R., Hansell, A.L., Fortier I., Manolio, T.A., Khoury, M.J., Little, J., and Elliott, P. (2009). Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. *Int. J. Epidemiol.* 38, 263-273.

De La Vega, F.M., Gordon, D., Su, X., Scafe, C., Isaac, H., Gilbert, D.A., and Spier, E.G. (2005). Power and sample size calculations for genetic case/control studies using gene-centric SNP maps: Application to human chromosomes 6, 21, and 22 in three populations. *Hum. Hered.* 60, 43-60.

Dupont, W.D. (1988). Power calculations for matched case-control studies. *Biometrics* 44, 1157-1168.

Dupont, W.D., and Plummer, W.D.Jr. (1990). Power and sample size calculations. A review and computer program. *Control Clin. Trials.* 11, 116-128.

Hennessy, S., Bilker, W.B., Berlin, J.A., and Storm B.L. (1999). Factors influencing the optimal control to case ratio in matched case-control studies. *Am. J. Epidemiol.* 149, 195-197.

Lewis, C.M. (2002). Genetic association studies: design, analysis and interpretation. *Brief Bioinform.* 3, 144-153.

Park, K., and Kim, H. (2007). A review of power and sample size estimation in genomewide association studies. *J. Prev. Med. Public Health.* 40, 114-121.

Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19, 149-150.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daily, M.J., and Sham, P.C. (2007). PLINK: A toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559-575.

Schork, N.J. (2002). Power calculation for genetic association studies using estimated probability distributions. *Am. J. Hum. Genet.* 70, 1480-1489.