Evaluation on the effect size of rare variants based on genome-wide association studies in WTCCC data

Shu-Hui Wen[1]
[1] Department of Public Health, Tzu-Chi University, Hualien, Taiwan
[1] Corresponding author: Shu-Hui Wen, e-mail: shwen@mail.tcu.edu.tw

Abstracts

Genome-wide association studies (GWASs) have identified hundreds of susceptibility genetic variants which were associated with complex diseases, however, most common variants explain only a modest proportion of heritability of these diseases. There are many causes for so-called missing heritability, and one reason is the ignorance of the impact of rare variants. Currently, many statistical tests developed for common variants in GWASs may not be directly applicable for rare variants due to low power. Hence, novel and powerful statistical method is needed for rare variants. In this presentation, we estimate the effect size of rare variants based on risk difference, odds ratio and Cohen's h. In addition, the relationship between the threshold of MAF, the magnitude of effect size, the sample size and the power will be examined. We utilize the coronary artery disease (CAD) case-control data from the Wellcome Trust Case Control Consortium (WTCCC) to provide an evaluation of type I error rate for each effect size. Using a total of 413,059 genetic markers on Chromosomes 1~22, the results indicated that larger variation was found for odds ratio than that of Cohen's h. In particular, for rare SNPs, the values of type one error rates are slightly higher than the nominal level 0.05, regardless of effect sizes.

Keyword: Cohen's *h*, Odds ratio, Effect size, Genome-wide association study, Type I error.

1. Introduction

Genome-wide association studies (GWASs) have been used widely in searching common single nucleotide polymorphisms (SNPs) contributing to complex diseases on the basis of current genotyping arrays in recent years (WTCCC 2007; Feng & Zhu 2010). The rationale for GWAS is common disease common variant which asserts common disease might be attributable to a small number of variants at minor allele frequencies (MAF) larger than 1%-5% with moderate risk (Manolio et al. 2009; Asimit & Zeggini 2010). Although GWASs have identified hundreds of susceptibility genetic variants which were associated with complex diseases, most common variants confer relatively small risk (odds ratio (OR) at 1.1-1.5) and explain only a modest proportion (at most 5%-10%) of heritability of these diseases (Schork et al. 2009; Hindorff et al., 2009). This leads to a question how the missing heritability can be explained, and one possible reason is due to possible contribution of variants with low MAF such as rare variants with MAF < 0.5% (Manolio et al. 2009). There is growing interest of rare variants and most of several identified rare variants have ORs above 2 (Bodmer &Bonilla 2008). In addition, development of next generation sequencing project such as 1000 GENOME project will enable the evaluation of rare variants in genetic association studies (The 1000 Genomes Project Consortium 2012).

Currently, many strategies have been developed for identifying disease-associated rare variants. These studies put focus on hypothesis testing and the aim is to enlarge the power for detecting rare variants. However, for rare variants, such as SNP with MAF 0.001, the variability of OR estimates might increase and lead to an over-estimate of effect size (ES) as noted in my previous study (Wei et al. 2010). In other words, testing approaches developed for common SNPs need modification to make sure the power is satisfied if focus on analysis of rare SNPs. We argue that commonly used measures of ESs might also need correction or evaluation when taking

care of rare SNPs. As rare SNPs might cause the bias in the estimation of ESs, the impact of this bias needs to be evaluated. The aim of this paper is to evaluate the performance of commonly used effect size (ES) measures, such as risk difference (RD), odds ratio (OR) and Cohen's h (Cohen 1988). The coronary artery disease (CAD) case-control data from the Wellcome Trust Case Control Consortium (WTCCC) are used for application (WTCCC 2007). The type I error rates for each of ES measure were presented for rare variants, as well as for common variants. Furthermore, the association tests based on each of ES measures with WTCCC CAD dataset show that the significant rare variants reveal several disease-related genes, most of which are consistent with biological networks underpining disease etiology.

2. Results

Theoretical distribution of each of ES measures, such as RD, OR and Cohen's h, was derived as asymptotic normal distribution, since all of these ES measures are functions of minor allele frequencies (MAFs) from case and control group. Taking advantage of two shared controls from WTCCC data, we are able to examine the type I error rate of genetic association test on the basis of each of these ES measures. For the CAD dataset, data cleaning was required prior to the analysis. We considered only SNPs and individuals passing WTCCC data quality control except for minor allele frequency $\geqq 0.001$. We termed common SNPs with MAF $\geqq 0.05$, and rare SNPs with $0.05 >$ MAF $\geqq 0.001$. According to the quality control criteria, a total of 1926 subjects with CAD and 2938 two shared controls were derived. In addition, we filtered the SNP set leaving a total of 413089 SNPs consisting of 52220 rare SNPs and 360839 common SNPs. First, we use two shard controls for evaluation of type I error rates of each ES for each chromosome. Figure 1 presented the estimated type I error rate for each ES for either common SNPs (dashed line) or rare SNPS (solid lines). The performance of each ES is very close for common SNPs at each chromosome. Range of type I error rates for each ES is approximately (0.047, 0.062). As for rare SNPs, range of type I error rates of each ES is (0.05, 0.093). The results indicate that either ES measure would probably produce slightly larger false positive results at examining the effect of rare variants in genetic association studies. The value of type I error from Cohen's h is slightly larger than those from RD or OR except at few chromosomes. However, we found that the biases, and mean square errors of estimates for log(OR) is quite larger than other two ES measures. In other words, the OR, ratio type of ES, might be likely to obtain biased estimate of the disease risk for rare variants.

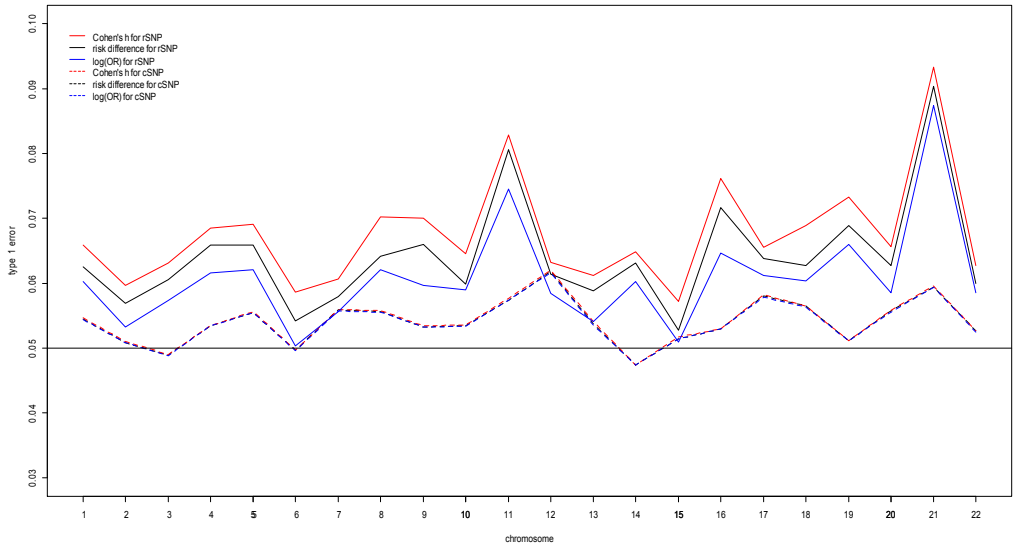

Figure 1. Type I error rates for RD, OR and Cohen's h for either rare SNPs or Common SNPs.

Next, we perform single marker association test based on all of each ES with

application to CAD GWAS data. P-value significance threshold is adjusted by the Bonferroni correction, that is 0.05/413089, either for selecting significant common SNPs or rare SNPs. To understand the biological significance of the significant rare variants reported by analysis results, we also used the HuGE Navigator database (Yu et al. 2008) and dbSNP (NCBI website: http://www.ncbi.nlm.nih.gov/projects/SNP/) database for searching disease-gene association with the CAD. The dbSNP website provides search outcomes including details about the SNP and its minor allele, physical position, neighboring SNPs, whether or not the SNP is covered by genes, the name and references of the genes, and association results from GWAS that they are from NHGRI Catalog and from association results submitted to dbGaP. Table 1 presented the testing results based on OR, RD, and Cohen's h with application to CAD dataset. We found that a total of 14 significant common SNPs were identified from each ES, and all of 14 SNPs were identical. However, the numbers of significant rare SNPs were different from either ES. The number of significant rare SNPs by Cohen's h was the most, followed by RD, and OR. As compared to OR, 4 and 12 additional SNPs were further explored by RD and Cohen's h, respectively.

Table 1. Number of significant SNPs identified by OR, RD and Cohen's h for CAD

		Effect Size (ES)		
SNP	Statistic	OR	RD	Cohen's h
Common	range of ES for sig. SNP	(0.77, 6.10)	(-0.06, 0.35)	(-0.13, 0.80)
	no. of sig. SNP	14	14	14
Rare	range of ES for sig. SNP	(1.88, 2.41)	(-0.01, 0.04)	(-0.17, 0.18)
	no. of sig. SNP	9	13	21
	no. of genes	5	7	8
	no. of validated gene[a]	3	5	5
	no. of validated[b] neighboring SNP (within 500 kb)	2	3	9

[a]: disease-gene associations that have been annotated on the HuGE Navigator database.
[b]: disease-snp associations that have been annotated on the dbSNP database.

To evaluate the biological significance of significant rare SNPs from this analysis, we found that the significant SNPs of CAD map to 3, 5, and 5 disease-gene association for OR, RD and Cohen's h, respectively, that had been annotated on the HuGE Navigator database. We search for disease-gene association according to CAD and/or relating risk factors including blood pressure (BP), body mass index (BMI), and diabetes mellitus (DM). In particular, the EIF4H gene identified by RD only has been reported to associate with CAD. Other genes have been reported to be related to other cardiovascular phenotype such as heart failure, or risk factors including BMI, DM, and BP (Table 2). In addition, we further search disease-snp association within 500 kb near significant SNPs. Nine out of 21 significant SNPs by Cohen's h was found to be associated with other cardiovascular phenotypes including coronary disease, arteries, myocardial infarction, heart failure, left ventricular, electro cardiography and echocardiography. It indicated that the Cohen's h was able to capture more related loci to CAD than RD and OR. It is worth noting that RD and Cohen's h is robust to very rare variant (e.g. MAF < 0.001), however, OR might be sensitive or not applicable as MAF is very small.

3. Conclusions

In this paper, we had evaluated statistical properties of three ES measures, OR, RD and Cohen's h for identifying rare SNPs in GWAS data. OR and RD are commonly used in GWAS, and Cohen's h is rarely used at this field. For rare variants,

the values of type I error rates from each of OR, RD and Cohen's h are slightly larger than nominal level 0.05. This suggested that the control of false positives is needed for analyzing rare variants. In addition, the estimation of ES from OR had larger mean square error than other two measures. Hence, OR might not be a reliable measure for rare variants. With application to CAD data, we found that the number of common SNPs identified by all three ES measures is the same and these significant common SNPs are identical. However, for rare SNPs, the number of significant rare SNPs is different from each ES, and Cohen's h obtains the largest set of rare SNPs. With a stringent threshold adjusted by the total number of SNPs (i.e. 413089), Cohen's h remains detecting more associated SNPs and most of the selected rare SNPs have biological meaning of CAD. In conclusion, RD and Cohen's h might be more suitable for identifying rare SNPs than OR.

References

1. Asimit, J., & Zeggini, E. (2010) "Rare variant association analysis methods for complex traits," The Annual Review of Genetics, 44, 293-308.
2. Bodmer, W., Bonilla, C. (2008) "Common and rare variants in multifactorial susceptibility to common diseases," Nature Genetics, 40(6), 695–701.
3. Cohen, J. (1988) Statistical power analysis for the behavioral sciences. 2nd Ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
4. Feng, T., & Zhu, X. (2010) "Genome-wide searching of rare genetic variants in WTCCC data," Human Genetics, 128(3), 269-280.
5. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramosa, E.M., Mehtac, J.P., et al. (2009) "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," Nature, 461, 747–753.
6. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., et al. (2009). "Finding the missing heritability of complex diseases," Proceedings of the National Academy of Sciences of the USA, 106, 9362 – 9367.
7. Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009) "Common vs rare allele hypotheses for complex diseases," Current Opinion in Genetics & Development, 19(3), 212-219.
8. The 1000 Genomes Project Consortium. (2012) "An integrated map of genetic variation from 1,092 human genomes," Nature, 491, 56-65.
9. The Wellcome Trust Case Control Consortium (2007) "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," Nature, 447, 661-678.
10. Wei, Y.C., Wen, S.H., Chen, P.C., Wang, C.H., Hsiao, C.K. (2010) "A Simple Bayesian Mixture Model with a Hybrid Procedure for Genome-wide Association Studies," European Journal of Human Genetics, 18(8), 942-947.
11. Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., Khoury, M.J. (2008) "A navigator for human genome epidemiology," Nature Genetics, 40(2),124-125. (HuGE Navigator)

Table 2: The disease association of selected genes from SNP analyses based on RD, OR and Cohen's h for CAD.

Chr.	ES	rs ID	**Gene**	location	MAF Control	MAF CAD	OR	P-value	RD	P-value	Cohen's h	P-value	Association[a]
3	**All**	**rs17042882**	**PLCL2**	**3p24.3**	**0.028**	**0.061**	**2.255**	**$4.88X10^{-15}$**	**0.033**	**$1.11X10^{-15}$**	**0.163**	**$4.00X10^{-15}$**	**Heart failure**
3	**h**	**rs16827563**	**VEPH1**	**3q24-q25**	**0.005**	**0**	**0.026**	**0.014**	**-0.005**	**$2.18X10^{-5}$**	**-0.119**	**$1.02X10^{-8}$**	**Carotid artery disease, DM**
7	**RD**	**rs17146094**	**EIF4H**	**7q11.23**	**0.017**	**0.034**	**2.036**	**$1.27X10^{-7}$**	**0.017**	**$7.15 X10^{-8}$**	**0.109**	**$1.32X10^{-7}$**	**CAD**
8	**All**	**rs16891338**	**SAMD12-AS1**	**8q24.12**	**0.023**	**0.043**	**1.908**	**$4.11 X10^{-8}$**	**0.02**	**$2.50X10^{-8}$**	**0.113**	**$4.66X10^{-8}$**	**BP**
8	All	rs16908145	FLJ45872	8q24.23	0.022	0.043	1.998	6.54 **$X10^{-9}$**	0.021	$3.46X10^{-9}$	0.12	$7.08X10^{-9}$	Neurotic disorder
15	**RD, h**	**rs7163007**	**MAP2K5**	**15q23**	**0.002**	**0.011**	**5.551**	**$2.13 X10^{-7}$**	**0.009**	**$5.33X10^{-9}$**	**0.121**	**$5.85X10^{-9}$**	**BMI, DM**
16	All	rs16955238	GAN	16q24.1	0.022	0.046	2.143	8.91 **$X10^{-11}$**	0.024	$3.41X10^{-11}$	0.135	$8.53X10^{-11}$	NA
16	h	rs7197337	ANKRD26P1	16q11.2	0.006	0	0.022	0.007	-0.006	$2.88X10^{-6}$	-0.132	$1.76X10^{-10}$	NA
19	**All**	**rs11671119**	**MEF2B MEF2NB**	**19p13.11**	**0.033**	**0.071**	**2.239**	**$<1.0X10^{-15}$**	**0.038**	**$<1.0X10^{-15}$**	**0.174**	**$<1.0X10^{-15}$**	**DM**

MAF: minor allele frequency, [a]: disease-gene associations that have been annotated on the HuGE Navigator database. Bold type means selected genes are related to CAD and/or relating risk factors including blood pressure (BP), body mass index (BMI), and diabetes mellitus (DM).