Contributed paper

A DETERMINISTIC ALGORITHM FOR OPTIMALITY OF THRESHOLD IN A GWAS EXPERIMENT

Z. Loh, J. H. J. van der Werf and S. Clark

School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia

SUMMARY

While genome-wide association study (GWAS) is an important tool for gene discovery for economic traits in livestock, its use of large numbers of genetic markers necessitates the use of multiple testing correction methods. Several of these methods have been suggested, but their optimality is not as well studied. The aim of this study is to present a deterministic algorithm to provide a framework for estimating the power and false positive rate (FPR) in a GWAS, and using these estimates to test the optimality of these correction method based on the Receiver Operating Characteristic (ROC) curve. This study suggests that both Bonferroni correction and Benjamini-Hochberg False Discovery Rate are overly conservative even if under the assumption of independence between markers.

INTRODUCTION

Genome-wide association studies (GWAS) are commonly used to identify genes associated with quantitative traits. Due to the increasingly large number of markers used in GWAS however, it had been plagued by an unprecedented level of a multiple testing problem. To avoid the correspondingly increased number of false positives, a multiple testing method that increases the threshold for significance had been utilized in GWAS (Gondro 2015; Tam *et al*. 2019; Visscher *et al*. 2017).

The Bonferroni correction was originally proposed due to its effectiveness in controlling the false positives (Narum 2006), but has since been widely criticized for its conservativeness (de Smet *et al*., 2004; Narum 2006; Tam *et al*. 2019). Alternative correction methods with reduced stringency in their threshold such as the frequently used Benjamini-Hochberg False Discovery Rate (BH-FDR) method have been suggested. A test on threshold optimality, defined as its ability to optimally balance the power and FPR of GWAS is lacking. Such an optimal threshold may depend on sample size, QTL effect distribution and marker allele frequencies.

The aim of this study is to test the degree of optimality of thresholds provided by Bonferroni and BH-FDR methods under varying relevant parameters. Optimality will be derived from an estimate of power and FPR of a GWAS using a deterministic algorithm, and using these estimates to establish the optimality of these thresholds.

THEORY

In this study a threshold would be considered optimal if it could balance the power and FPR in a GWAS. Given a threshold THR, alongside with effect size of the marker a, phenotypic variance $Var(p)$, allele frequency p, sample size of GWAS N and number of QTL $nqtl$, the power of GWAS can be defined as follow:

$$power = \frac{Number\ of\ true\ QTLs\ that\ \ exceed\ THR\ \left(-log_{10}(pvalue)\right)}{nqtl}$$

The expected $pvalue$ of a locus could in turn be calculated using the following equation:

$$pvalue = \ 2 - 2t_{CDF}\left(a\sqrt{\frac{2p(1-p)(N-2)}{Var(P) - 2p(1-p)a^2}}\ , N-2\right)$$

Where $t_{CDF}(t, n)$ is the cumulative density function (CDF) of Student's t-distribution with test statistic t and degree of freedom n. While $nqtl$ is not estimated in this study, deterministic algorithms for this estimation are available with assumption on the distribution of QTL effect sizes, for example see Hall *et al.* (2016). With the same threshold THR, the FPR could be defined as follow:

$$FPR = \frac{Number\ of\ null\ marker\ that\ exceed\ THR}{Number\ of\ null\ marker}$$

As this model assumed independence between markers, linkage disequilibrium is not assumed, and null marker are modelled with effect size 0. Modelling of simulated null markers suggested that FPR followed a 1-CDF of gamma distribution with shape and scale parameter of 1 and 0.4344 respectively, and FPR depends only on THR. Thus the equation of FPR can be rewritten as follow:

$$FPR = (1 - gamma_{CDF}(THR;\ 1, 0.4344))$$

Where $gamma_{CDF}(x;\ k, \theta)$ is the CDF of gamma distribution at point x with shape and scale parameter k and θ. To test the optimality of THR, a receiver operating characteristic (ROC) curve was used. The conventional ROC curve have its FPR and power plotted at x and y-axis, respectively, with optimal threshold being the point where the tangent of the curve equal to 1 (as described by de Smet *et al.* (2004) and mathematically proven by Kaivanto (2008)). Another interpretation which was used in this study, is the difference between number of true and false positives, which represent the numerator of $power$ and FPR respectively. The optimal threshold can then be defined as the argument of the maxima of this differences, where the power is maximized and FPR minimized. This interpretation can also take into account the unequal chance between finding true QTLs and null markers. A sample of this reinterpreted ROC curve would be provided in Figure 1.

VALIDATION OF THE MODEL

The model was validated through simulation using Python (Version 3.7.3), where the optimality of threshold calculated by Bonferroni and BH-FDR was compared under varying parameters.

A GWAS experiment with N sample size was simulated with a genotype array with M number of independent markers with their allele frequencies following a beta-distribution. A vector of effect sizes was assigned to $nqtl$ markers, which were considered QTL with their effect sizes following a gamma distribution while other markers had effect size of 0. Only markers with effect size of > 0.1 σ were considered in the calculation of power. For all simulations the heritability of the trait was set to 0.3. Using the genotype array, effect sizes and heritability, a vector of phenotypes was calculated, and a GWAS was conducted using Single SNP Linear Regression with the genotype array and phenotype vector. Using Bonferroni correction and BH-FDR at alpha = 0.05, the number of true and false positives were recorded. The ROC score was calculated by subtracting number of false positives from number of true positives. Correction methods with higher ROC score are deemed having its threshold more optimal and provide better balance between power and FPR. This simulation was repeated 200 times. When a parameter is under study the other parameters were kept at the Default Value. The parameters tested are presented in Table 1.

Table 1. Parameter tested in this experiment

Parameter	Default Value	Alternative Value
Sample Size	2000	800
Number of Markers	20k	80k
Distribution of QTL Effect Size	Gamma(0.4, 1)	Gamma(0.8, 1)
Distribution of Allele Frequency	Beta(0.5, 0.5)	Beta(0.2, 0.2)
Number of QTLs	100	2000

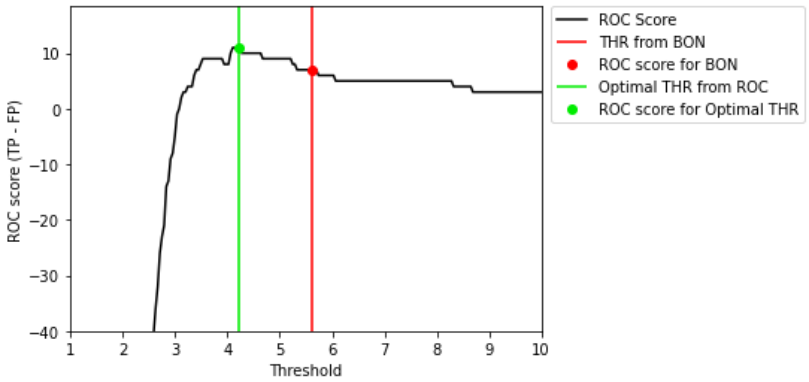

Figure 1. The reinterpreted ROC curve under default scenario with Bonferroni correction, with TP and FP representing number of true and false positives respectively

RESULTS AND DISCUSSION

The number of true and false positives from each correction method are provided in Table 2, and the ROC score and threshold of each correction methods were provided in Table 3.

Table 2. The number of true positives (TP) and false positives (FP) for each correction methods under varying parameter values[1]

Parameter Tested	Values	Multiple Testing Correction Method					
		Optimal Threshold from ROC		Bonferroni Correction		BH-FDR	
		TP	FP	TP	FP	TP	FP
Sample Size (Default)[1]	2000	11.36	0.86	7.82	0.09	9.77	0.63
(Alternative)	800	4.70	0.42	2.64	0.04	3.35	0.20
Number of Markers	80k	9.53	0.72	6.92	0.05	8.30	0.44
Distribution of QTL Effect Size	Gamma(0.8,1)	11.70	1.01	7.52	0.02	9.72	0.50
Distribution of Allele Frequency	Beta(0.2, 0.2)	9.96	0.61	7.37	0.07	8.84	0.48
Number of QTLs	2000	6.77	2.31	1.08	0.02	1.47	0.07

[1] The default values are provided in Table 1.

Table 3. The threshold (THR) and ROC score for each correction methods under varying parameter values[1]

Parameter Tested	Values	Multiple Testing Correction Method					
		Optimal Threshold from ROC		Bonferroni Correction		BH-FDR	
		THR	ROC	THR	ROC	THR	ROC
Sample Size (Default)[1]	2000	4.29	11.18	5.60	7.49	4.62	9.10
(Alternative)	800	4.73	4.05	5.60	2.61	5.09	2.86
Number of Markers	80k	5.06	8.82	6.20	6.88	5.28	7.86
Distribution of QTL Effect Size	Gamma(0.8,1)	4.32	10.84	5.60	7.79	4.58	9.41
Distribution of Allele Frequency	Beta(0.2, 0.2)	4.46	9.35	5.60	7.30	4.64	8.36
Number of QTLs	2000	3.93	4.58	5.60	0.93	5.34	1.24

[1] The default values are provided in Table 1.

Compared to both the Bonferroni and the BH-FDR methods, the threshold optimal to the ROC curve has a significantly higher number of false positives in all scenarios, which is associated with a significantly lower threshold. This suggests that the threshold optimal to ROC is less stringent compared to both correction method. Despite this, as suggested by the increased ROC score, the increment of power of GWAS due to the decreased threshold is more significant than the increment of FPR, which could suggest that both Bonferroni correction and BH-FDR are overconservative for all the scenarios in this study.

Between the two existing correction methods, BH-FDR provided a better balance between power and FOR when compared to the Bonferroni correction. While the number of false positives also increased in this correction method, as suggested by Huang *et al*. (2018), the increment in true positives is more significant than the increment of false positives. While with the Bonferroni correction, the power is significantly lower than with BH-FDR, it also had a significantly smaller proportion of false positives. Indeed, the Bonferroni correction had successfully maintained the number of false positives between 0.02 and 0.09 in all scenarios, whereas BH-FDR failed to maintained it in all the scenarios.

While this experiment has illustrated the optimality of threshold from the multiple correction methods, there were several assumptions being made. One of the main assumptions is the independence of the markers, which is unlikely to occur in actual GWAS. Huang *et al.* (2018) suggested threshold from correction methods that assumed independence between markers had increased conservativeness compared to those without such assumption. Despite this, even if this assumption is held, as in this experiment, both correction methods are still overconservative in respect with the optimal threshold. Further study on the effect of correlated markers on the optimality of thresholds from these correction methods would be required.

CONCLUSION

This study had provided a framework for estimating the power and false positive rate of GWAS using a deterministic algorithm, and using these measures to test the optimality of threshold from two common multiple testing correction methods. This study had demonstrated the excessive conservativeness in both correction methods, especially in Bonferroni correction. The BH-FDR attained a better balance between true and false positives in the setting of independent markers and thus a more optimal threshold. Despite this the optimality of these threshold from correlated markers still warranted further study.

REFERENCES

de Smet F., Moreau Y., Engelen K., Timmerman D., Vergote I. and de Moor B. (2004) *Br. J. Cancer* **91**: 1160.

Gondro C. (2015). Primer to analysis of genomic data using R. Springer International Publishing, Cham.

Hall D., Hallingbäck H.R. and Wu H.X. (2016) *Tree Genet. Genomes* **12**: 110.

Huang Q.Q., Ritchie S.C., Brozynska M. and Inouye M. (2018) *Nucleic Acids Res*. **46**: e133.

Kaivanto K. (2008) *J. Clin. Epidemiol*. **61**: 516.

Narum S.R. (2006) *Conserv. Genet*. **7**: 783.

Tam V., Patel N., Turcotte M., Bossé Y., Paré G. and Meyre, D. (2019) *Nat. Rev. Genet*. **20**: 467.

Visscher P. M., Wray N. R., Zhang Q., Sklar P., McCarthy M. I., Brown M. A. and Yang J. (2017) *Am. J. Hum. Genet*. **101**: 5