Methods xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Statistical selection of biological models for genome-wide association analyses

Wenjian Bi, Guolian Kang, Stanley B. Pounds*

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

ARTICLE INFO

Keywords:
Biological models
Genome-wide association study
Multiple adjusted evidence weights
Two-stage discovery validation study

ABSTRACT

Genome-wide association studies have discovered many biologically important associations of genes with phenotypes. Typically, genome-wide association analyses formally test the association of each genetic feature (SNP, CNV, etc) with the phenotype of interest and summarize the results with multiplicity-adjusted p-values. However, very small p-values only provide evidence against the null hypothesis of no association without indicating which biological model best explains the observed data. Correctly identifying a specific biological model may improve the scientific interpretation and can be used to more effectively select and design a follow-up validation study. Thus, statistical methodology to identify the correct biological model for a particular genotype-phenotype association can be very useful to investigators. Here, we propose a general statistical method to summarize how accurately each of five biological models (null, additive, dominant, recessive, co-dominant) represents the data observed for each variant in a GWAS study. We show that the new method stringently controls the false discovery rate and asymptotically selects the correct biological model. Simulations of two-stage discovery-validation studies show that the new method has these properties and that its validation power is similar to or exceeds that of simple methods that use the same statistical model for all SNPs. Example analyses of three data sets also highlight these advantages of the new method. An R package is freely available at www.stjuderesearch.org/site/depts/biostats/maew.

1. Introduction

Genetic association studies have successfully identified single nucleotide polymorphisms (SNPs) that are associated with complex human diseases [1,2]. Traditionally, genetic association studies have used p-values to identify potentially meaningful variants. Many statistical methods have been developed to calculate p-values of the genetic association studies and each method is preferred for a specific niche of possible settings. For example, the Cochran-Armitage trend test (CATT) is most powerful if the underlying genetic model is additive [3,4]; Pearson's Chi-square test is robust to different genetic models but has less statistical power than the CATT if the underlying model is additive [5]; some entropy-based methods can be more powerful than the Pearson's Chi-square test [6]; and the set-valued based method has power similar to or greater than that of CATT in some extreme situations such as small sample size or rare variants [7].

However, the p-value only measures evidence against the null hypothesis but does not assist investigators in identifying a specific biological model [8]. For example, a very small p-value suggests that the null model is false but does not indicate whether any specific genetic model (additive, dominant, recessive, or co-dominant) is true. Correctly identifying a specific genetic model will enhance biological interpretation and can be used to select the most powerful statistical test for a follow-up validation study from the same population. Thus, some statistical methodology to identify the correct genetic model for a certain genotype-phenotype association can be very useful to investigators.

In this paper, we propose multiplicity-adjusted evidence weights (MAEW) as a novel method to empirically select the most appropriate genetic model from among five candidate models (null, additive, dominant, recessive, and co-dominant) for each genetic variant. The evidence weights are calculated based on the Bayesian Information Criterion (BIC) and adjusted for multiple-testing by estimating the empirical Bayesian probability (EBP) that the null hypothesis is true for each genetic variant [9–13]. The method provides a readily interpretable quantitative metric of the evidence supporting each of the five genetic models for each variant, that stringently controls the false discovery rate and asymptotically select the correct biological model. Additionally, by selecting the most accurate biological model of the data, our method, increases validation power in two-stage discovery-validation studies. These properties can be observed in simulation

* Corresponding author.
E-mail address: stanley.pounds@stjude.org (S.B. Pounds).

https://doi.org/10.1016/j.ymeth.2018.05.019
Received 10 January 2018; Received in revised form 13 April 2018; Accepted 22 May 2018

studies and in our analysis of the GAW17, BEN and CSSCD data set.

The remainder of this paper is organized as follows. In Section 2, we describe how to compute multiplicity adjusted evidence weights (MAEW) in detail and some statistical properties, including false discovery rate (FDR) control and asymptotic selection of the correct genetic model. In Section 3, we observe that MAEW exhibits these properties in simulations of two-stage discovery-validation studies and in replicates of split-set discovery validation analyses of three example data sets. Section 4 provides discussion and concluding remarks.

2. Methods

We first introduce BIC and give the definition of MAEW. Then, we give methods to compute MAEWs in genetic association studies. Next, we show some properties and an application in discovery-validation study.

2.1. BIC and evidence weights

BIC is a well-known statistical model selection criterion that measures how well a set of candidate models fit a set of data. Given a collection of candidate models for the data, BIC estimates the goodness of fit for each model relative to each of the other models.

Suppose that we have obtained phenotype data and genotype data for each of $l = 1, \ldots, L$ genetic loci. Also, assume that the association of phenotype with genotype at each locus may be characterized by one of $m = 0, 1, \ldots, M$ candidate statistical models. We let $m = 0$ index the null model and $m = 1, \ldots, M$ index the other models under consideration. For each locus l, we let K_{lm} denote number of parameters for model m, let n_l denote the sample size and compute corresponding maximum value L_{lm} of the likelihood function of model m given genotype data at locus l and phenotype data. For each locus and model, the BIC value $\hat{a}_{lm}$ is defined as

$$\hat{a}_{lm} = \ln(n_l)K_{lm} - 2\ln(L_{lm}). \tag{1}$$

Smaller values of the BIC indicate better model fit adjusted for model complexity (number of parameters in the statistical model). Note that for simplicity of notation, we omit the phenotype from the equations. The implicit dependency of the likelihood on the phenotype is clear by context. Suppose the minimal value of BICs for a particular locus l across all candidate models is $\tilde{a}_l = \min_m(\hat{a}_{lm})$. We then compute the BIC difference Δ_{lm} and BIC evidence weight ω_{lm} for locus l as

$$\Delta_{lm} = \hat{a}_{lm} - \tilde{a}_l, \quad \omega_{lm} = \frac{\exp\left(-\frac{1}{2}\Delta_{lm}\right)}{\sum_{r=0}^{M}\exp\left(-\frac{1}{2}\Delta_{lr}\right)}. \tag{2}$$

The evidence weight ω_{lm} can be considered as the weight of evidence in favor of model m being the best model of the data of locus l among the set of candidate models. By definition, evidence weights satisfy the mathematical properties of probabilities (each evidence weight is greater than or equal to zero, less than or equal to one, and the sum of evidence weights across the models equals one at each locus), although they do not strictly have the interpretation of probabilities. We note that the BIC and evidence weight can also be computed for models that involve many kinds of phenotypes (binary, quantitative, survival) and account for environmental covariates, thus making the methodology very generalizable.

This evidence weight procedure was originally developed to perform one model selection. Here, we wish to perform one model selection for each locus $l = 1, \ldots, L$. This multiplicity problem must be addressed. It is well-known that p-values must be adjusted for multiple-testing. This is also the case for using BIC for genome-wide association analyses. This is clearly seen in that the likelihood and p-value are closely related (consider the likelihood ratio test for example) and the BIC is directly a function of the likelihood as shown in Eq. (1). Thus, the BIC also needs to be adjusted for multiplicity. As described below, we will adapt a new procedure of empirical Bayesian probability (EBP) estimation for that purpose.

2.2. Adjusting evidence weights for multiplicity

To adjust BIC evidence weights for multiplicity, we first formally test the null hypothesis that each feature's genotype is not associated with phenotype. We use a hypothesis testing procedure designed to detect the specific alternative that is considered of greatest interest or to have the greatest statistical power for the greatest number of loci. For example, it is common to use an additive model to test the association of the phenotype with the genotype of each SNP. Another approach may be to perform the test using a model with a non-specific alternative, such as the Kruskal-Wallis test or one-way ANOVA, which is equivalent to testing for a co-dominant genetic model (Appendix A). In either case, we obtain a p-value p_l for each feature $l = 1, \ldots, L$.

Next, we develop and fit a likelihood-based adaptive histogram estimator (LB-AHE) to the set of p-values obtained above. The adaptive p-value histogram estimator is an extension of a method developed by Nettleton and colleagues that uses a p-value histogram to estimate the proportion of tests with a true null [11–13]. Here, we use a novel adaptive p-value histogram to compute an estimate q_l of the empirical Bayes probability (EBP) that the null hypothesis is true for locus l. In this application, we use EBP instead of other multiple-testing corrections because the probabilistic interpretation of EBP is most compatible with our objective to compute meaningful evidence weights. The similarities and differences between the EBP and the other false discovery metrics have been described previously [14,15]. Briefly, the EBP tends to be more conservative than Storey's q-value [16] procedure and the EBP has the large-sample property of correctly distinguishing between the null and alternative hypothesis for every test. The detailed description of the p-value histogram estimator and EBP calculation are in Appendix B.

For each locus l, we have an EBP estimate q_l that the null hypothesis is true and a set of unadjusted BIC evidence weights ω_{lm} for each of a set of candidate models $m = 0, \ldots, M$, where $m = 0$ indexes the null model. Thus, we can define the multiplicity adjusted evidence weights (MAEWs) for each locus l as

$$\tilde{\omega}_{lm} = \begin{cases} q_l, & m = 0, \\ \frac{(1-q_l)}{(1-\omega_{l0})}\omega_{lm}, & m = 1, \ldots, M, \end{cases} \tag{3}$$

where q_l is the EBP estimate of the probability that the null hypothesis is true when considering multiplicity.

The MAEW provides a richer interpretation of the data than does multiplicity-adjusted p-values. Also, incorporating the EBP multiplicity adjustment into the definition of MAEW in Eq. (3) ensures that comparing $\tilde{\omega}_{l0} = q_l$ to a specific threshold τ provides the same control of the false discovery rate at the level τ as does the method used to compute q_l. However, the MAEW provides more perspective. MAEW not only provides Type I error rate control but also quantifies the evidence in support of specific alternative statistical models that have a meaningful biological interpretation. Observing $\tilde{\omega}_{lm} \approx 1$ indicates that the data overwhelmingly support genetic model m over all other genetic models for locus l. Conversely, observing $\tilde{\omega}_{lm} \approx 0$ indicates that the data do not support genetic model m for locus l. This quantification is also useful to better understand the biological processes underlying the association and improve the power of validation studies that seek to confirm significant results.

With this richer biological interpretation comes a more complex characterization of errors than the classical setting with Type I or Type II errors. As shown in Table 1, there are many types of errors that can be incurred by a MAEW analysis that selects among five genetic models for the genotype-phenotype association at each locus. In particular, for each true genetic model, one may incorrectly select any of the four remaining models. Thus, in our simulation studies below, we will