A note on p-value multiple-testing adjustment for two-step genome-wide gene-environment interactions scans

Juan Pablo Lewinger[1], Eric S. Kawaguchi[1], and W. James Gauderman[1]
[1]Department of Population and Public Health Sciences, University of Southern California

Abstract

Two-step testing is the state-of-the art approach for performing genome-wide interaction scans (GWIS). It is computationally efficient and yields higher power than standard single-step-based GWIS for virtually all biologically plausible scenarios. However, while two-step tests control the genome-wide type I error rate at the desired level, the lack of associated valid p-values can make it difficult for users to compare with single step-results. We show how multiple-testing adjusted p-values can be defined for two-step test based on standard multiple-testing theory, and how they can be in turn scaled to make valid comparisons with single-step tests possible.

1 Introduction

Recent gene-environment interactions findings were made possible by the pooling of multiple studies with available genotype and harmonized exposure data, and by using powerful two-step genome-wide interaction scan (GWIS) tests.

Two-step tests for GWIS have been introduced to improve power over standard single-step tests, while controlling the family-wise error rate (FWER). Both single- and two-step tests proceed by testing one-SNP-at-a-time for $G \times E$ interaction, by modeling the genotype, an environmental exposure, and the corresponding interaction term using classical linear, logistic, or Cox regression depending on the type of outcome. However, the traditional one-step GWIS approach, which tests each $G \times E$ interaction at a significance level of $\alpha = 5 \times 10^{-8}$, is limited in power, especially for weaker or less frequent interactions.

In these procedures, independent information on $G \times E$ that is not captured by the standard $G \times E$ test is used to perform an initial screening (Step 1) to prioritize SNPs that are more likely to be involved in an interaction. These SNPs are then formally tested for an interaction (Step 2) under a modified α^*, thus reducing the multiple testing burden.

There are two widely-used procedures for prioritizing SNPs in Step 2 $G \times E$ testing after the Step 1 screening: subset and weighted hypothesis testing. Subset testing only tests a subset of the SNPs that pass a significance threshold based on the screening statistic in Step 1, while weighted hypothesis testing allocates SNPs into bins based on the magnitude of the screening statistic and tests each bin at a different significance level. While weighted hypothesis testing is often more powerful than subset testing, there is no universally most powerful approach.

In this paper, we focus on constructing adjusted p-values for a two-step GWIS based on standard theory for multiple-testing procedures. These adjusted p-values can be di-

rectly compared across different procedures and provide a useful tool for evaluating the performance of different two-step testing strategies.

Two-step tests for GWIS have been proposed to improve the power of a $G \times E$ analysis while controlling the FWER for disease (Kooperberg and LeBlanc, 2008; Murcray et al., 2009, 2011; Hsu et al., 2012; Gauderman et al., 2013; Wang et al., 2021), quantitative traits (Paré et al., 2010; Zhang et al., 2016), and time-to-event traits (Kawaguchi et al., 2022). In all of these two-step procedures, independent information on $G \times E$ not captured by the standard $G \times E$ test, is used to perform an initial screening (Step 1) to prioritize SNPs that are more likely to be involved in an interaction. These SNPs are formally tested for an interaction (Step 2) under a modified α^*, thus reducing the multiple testing burden (Kooperberg and LeBlanc, 2008; Murcray et al., 2009).

The marginal outcome-gene association statistic derived from modeling the outcome on each gene individually is a commonly-used screening statistic for quantitative (Zhang et al., 2016), binary/disease (Kooperberg and LeBlanc, 2008), and time-to-event (Kawaguchi et al., 2022) traits. For case-control studies the exposure-gene association statistic, modeling the relationship between each gene on the exposure, can also be informative (Murcray et al., 2009). Methods that utilize both outcome-gene and exposure-gene associations in a case-control study have also been developed (Murcray et al., 2011; Hsu et al., 2012; Gauderman et al., 2013). A key requirement for validity of any two-step procedure is that the statistics used in Step 1 and Step 2 are asymptotically independent (Dai et al., 2012; Kawaguchi et al., 2022).

There are two widely-used procedures for prioritizing SNPs in Step-2 $G \times E$ testing after the Step-1 screening: subset (Kooperberg and LeBlanc, 2008; Murcray et al., 2009) and weighted hypothesis testing (Ionita-Laza et al., 2007). In subset testing of the M total SNPs that are being scanned, only the $m << M$ SNPs that pass a significance threshold based on the screening statistic are tested in Step 2 using a standard $G \times E$ test. The significance threshold in step 2 for $G \times E$ discovery is calculated using a Bonferroni correction that is based on the number of SNPs that pass the screening ($\alpha^* = \alpha/m$), which is much less stringent than the threshold used in a single step approach. A trade off for a relaxed threshold is that SNPs that do not pass the step 1 screening will not be tested. An alternative approach that does not rely on a pass/no pass hard rule is weighted hypothesis testing. Here, SNPs are allocated into bins based on the magnitude of the screening statistic. Each bin has a corresponding bin-wise error rate (BWER) such that the sum across all bins does not exceed α. Lower bins are allocated a larger fraction of α (see Section **??** for more detail), so that SNPs in those bins are tested at a more liberal significance threshold. Conversely, SNPs that are placed in higher bins are tested at a much more stringent BWER. Unlike subset testing, every SNP is tested in Step 2 of the weighed approach; yet SNPs that are more likely to have an interaction based on the screening statistic will have a higher chance of being discovered. Although weighted hypothesis testing is often more powerful than subset testing (Ionita-Laza et al., 2007; Gauderman et al., 2013), there is no universally most powerful approach.

2 Methods

For concreteness, we describe our proposed adjustment of two-step-tests GxE-p-values for multiple testing in the most common scenario of a GWIS with case-control data. How-

ever, the approach presented here is completely general, applying to two-step tests with continuous and time-to-event outcomes, and with any of the multiple testing correction approaches for two-step tests proposed in the literature, including subset, weighted Bonferroni, and the recently proposed approach by Kawaguchi et. al. Following the notation and setup of Kawaguchi et. al., consider a gene-environment interaction study with a (continuous, binary, or time-to-event) trait/outcome Y, an environmental exposure of interest E, and M SNPs (G_j, $j = 1, \ldots, M$) measured or imputed for each of N subjects. A two-step GWIS consists of a screening first step based on test statistic computed for each SNP. The goal of the screening step is to identify promising SNPs, i.e. SNPs that likely to interact with the exposure to affect the outcome. The screening step (step 1) is followed by a testing step where SNPs are formally tested for interaction. Appropriate screening statistics are discussed below. In the testing step (step 2), the SNPs are tested at a significance level that depends on how promising they are, i.e. based on the value of their corresponding screening statistic. For example, one possibility in step 2 is to only test the top most promising SNPs. We refer to this as the 'subset approach'. It is the enrichment in promising SNPs granted by the screening step, combined with a reduction of the multiple testing burden in step 2 that in most scenarios yields higher power to detect interactions for a two-step compared to a single-step approach. An option that does not require committing to a subset of SNPs in step 2 – and which typically results in an even more powerful two-step method– is to sort the SNPs according to the significance of the screening statistic (from most to least significant) and to test each SNP in step 2 so that more promising SNPs are tested at a less stringent significance level. This needs to be accomplished in a way that the genomewide type-I error is preserved, which can be achieved using a weighted hypothesis testing strategy (Ionita-Laza et al., 2007). Specifically, SNPs are assigned to K 'bins' $B_1, \ldots B_k$ (K is typically a dozen or so) of sizes $b_1, \ldots b_K$ respectively, according to their step-1 screening statistic or p-value. The set of most significant b_1 SNPs are assigned to B_1, the set of most significant set of b_2 SNPs after removing the b_1 SNPs in B_1 are assigned to B_2, and so on. To control the family-wise error rate (FWER), the overall target level α, is partitioned into bin-wise target error rates (BWER) $0 < \alpha_1, \ldots, \alpha_K < 1$ so that $\alpha_1 + \alpha_2 + \ldots + \alpha_K \approx \alpha$ are chosen. A common choice is $\alpha_k = \alpha/2^{k+1}$, for $k = 1, \ldots K$, which ensures α is correctly partitioned:

$$\alpha_1 + \alpha_2 + \ldots + \alpha_K = \sum_{k=1}^{K} \frac{\alpha}{2^{k+1}} \approx \alpha.$$

Apportioning α in this way guarantees that the FWER is controlled at level α. A Bonferroni correction can be used to preserve the BWER by dividing α_k by the size m_k of bin B_k. For the bin sizes, Ionita-Laza et al. (2007) suggest using $m_k = 5 \times 2^{k-1}$. For example, the SNPs with the largest values of the screening statistic are placed in bin 1 and tested against a significance threshold of $\alpha_1/5$, the next $10 = 2 \times 5$ SNPs are tested in bin 2 at level $\alpha_2/10$ and so forth. Ionita-Laza et al. (2007) suggested setting $B_0 = 5$ so that $m_1 = 5, m_2 = 10, m_3 = 20, m_4 = 40 \ldots$. This assignment of SNPs into bins ensures that the most promising SNPs are tested at the least stringent significance level. Since the binning of SNPs can be done in several ways, we refer to the Ionita-Laza et al. (2007) approach as *rank-based (RB)* binning to contrast with significance-based binning that we propose below. Two-step RB-weighted tests are in general, more powerful than subset testing (Gauderman et al., 2013).

For a binary case-control outcome the screening statistic can be a test statistic based on the marginal test for the effect of the genotype of SNP j G_j on the outcome Y, or a test statistic for testing the association between the exposure E and G_j regardless of case-control status (i.e. in the pooled sample of cases and controls), or a combination of the two. Both of these test statistics can be shown to be independent from the step-2 statistic, which is typically the test of interaction based on a logistic regression model of the form:

$$\log\left(\frac{P(Y=1\,|\,G_j,E,X)}{1-P(Y=1\,|\,G_j,E,X)}\right)=\beta_{0j}+\beta_{G_j}G_j+\beta_E E+\beta_{G_j\times E}(G_j\times E)+\beta_X X \qquad (1)$$

for each of the M SNPs and where X represents a vector of adjustment covariates like sex, age, and variables capturing population structure like principal components. Each T_j corresponds to the test statistic for testing the null hypothesis $H_0 : \beta_{G_j\times E} = 0$. A raw/unadjusted p-value $\{p_j\}_{j=1}^M$ is computed for each SNP using a likelihood ratio test statistic (preferred for logistic regression) or an asymptotically equivalent Wald or Score statistic. Key to controlling the family-wise type I error of the two-step procedure is the statistical independence between the step-1 screening an the step-2 testing test statistics.

2.1 Multiple-testing-adjusted p-values

The most common p-value definition as the probability of observing a value of the test statistic T larger than or equal to the observed value t_{obs}, $P(T \geq t_{obs}|H_0)$, under the null hypothesis is equivalent to the alternative characterization as the smallest significance level at which one would reject the null hypothesis:

$$p=\inf\{\alpha:\ \text{Null is rejected at level } \alpha\}=\inf\{\alpha : T\geq c_\alpha\} \qquad (2)$$

where c_α is a critical value that guarantees control (exact or approximate) of the type-I error at level α for the tested hypothesis. It is this equivalent p-value characterization that leads to a proper definition of multiple-testing-adjusted p-values. Suppose we now have a collection of null hypotheses $H_0(1),\ldots,H_M(0)$ that we wish to simultaneously test. For example, the hypotheses that SNP $j, 1\leq j\leq M$ does not interact with the exposure E to affect outcome Y. A *multiple testing procedure* (MTP) for $H_0(1),\ldots,H_M(0)$ consists of a collection of test statistics T_i and multiple-testing critical values $c_i^{MTP}(\alpha)$ for each $1\leq i\leq M$, such that we reject $H_0(i)$ when $T_i > c_i^{adj}(\alpha)$ and the family-wise type I error for the family of hypotheses $H_0(1),\ldots,H_M(0), i=1,\ldots M$ is controlled at level α, where $0<\alpha<1$ is a pre-specified type-I error target. The family-wise type-I error is the probability of wrongly rejecting at least one of the true null hypotheses among $H_0(1),\ldots,H_M(0), i=1,\ldots M$. If V denotes the total number wrong rejections (false positives), the family-wise type-I error can be expressed as $P(V>0)$, where the probability is taken under the true configuration of null and alternative hypotheses among the $H_0(1),\ldots,H_M(0)$. For example, let T_i be a statistic for testing the hypothesis $H_0(i), 1\leq i\leq M$ and $c_i(\alpha)$ a corresponding critical value so that if $H_0(i)$ is true, the type-I error or probability of wrong rejection, $P(T_i > c_i(\alpha))$, is less than or equal to α for the *individual hypothesis* $H_0(i)$. Then, a multiple-testing procedure for the family of hypotheses $H_0(1),\ldots,H_M(0)$ can be obtained via a Bonferroni correction, i.e. by choosing $c_i^{MTP}(\alpha) = c_i(\alpha/M)$ as the MTP critical

values. In this example, the MTP $c_i^{MTP}(\alpha)$ critical values only depend on the significance level α and the total number of hypotheses M. In the case of two-step GWIS, the MTP critical values are dependent: $c_i^{adj}(\alpha)$ depends on the value of the screening statistic $H_0(i)$.

When a multiple testing procedure is available for a family of hypotheses $H_0(1), \ldots, H_M(0)$, multiple-testing adjusted p-values can be defined by analogy to 2. Specifically, an adjusted p-value for $H_0(i)$ is defined as:

$$p_i^{adj} = \inf\{0 < \alpha \leq 1 : \text{ Null is rejected at the multiple-testing level } \alpha\} = \inf\{0 < \alpha \leq 1 : T \geq c_\alpha^{MTP}\} \tag{3}$$

In other words, the adjusted p-value is the smallest $0 < \alpha \leq$ at which the *multiple testing procedure* would reject $H_0(i)$. In the case of the Bonferroni multiple-testing procedure above, the adjusted p-value is given by $p_i^{adj} = \inf\{0 < \alpha \leq 1 : T \geq c_{\alpha/m}\} = p_i/M$, where p_i is the unadjusted p-value $p_i = \inf\{\alpha : T \geq c_\alpha\}$. This is the well known Bonferroni multiple-testing adjustment.

Let's now consider the subset and weighted-Bonferroni two-step procedures described above. In the subset two-step procedure, only the top m SNPs out of M, as determined by the screening step are tested in step 2. A SNP is then declared significant if its step-2 p-value is smaller than α/m. The subset two-step procedure is in effect a multiple-testing procedure for testing the hypotheses $H_0(k) : \beta_{G_j \times E} = 0, k = 1 \ldots M$. The MTP critical value is given by $c_\alpha^{MTP} = c_{\alpha/m}$ for a SNP that passes the screening and $c_\alpha^{MTP} = 0$ for a SNP that that does NOT pass the screening. Thus, the multiple testing adjusted p-value is given by:

$$p_i^{adj} = \inf\{0 < \alpha \leq 1 : p_i \geq c_\alpha^{MTP}\} = \begin{cases} p_i/m & \text{SNP } i \text{ passes the screening} \\ 1 & \text{SNP } i \text{ does not pass the screening} \end{cases} \tag{4}$$

In the subset two-step procedure a multiple-testing adjusted p-value of 1 is assigned to any SNP that not pass the screening in step 1.

For the weighted-Bonferroni two-step GWIS based on the binning described above, the critical value for rejection of a SNP i depends on the bin where the SNP lands. Let's assume SNP i lands in bin B_k of size $m_k = 5 \times 2^{k-1}$. The SNP is then declared significant for interaction if its step-2 p-value is smaller than $\alpha/2^k/m_k$. Thus, the multiple testing adjusted p-value is given by:

$$p_i^{MTP} = \inf\{0 < \alpha \leq 1 : p_i < \alpha/2^k/m_k\} = \min\{m_k 2^k p_i, 1\} \tag{5}$$

3 Example

To illustrate the multiple testing p-value adjustment consider the gene by alcohol intake colorectal-cancer GWIS from the FIGI consortium (Figure 1). For this GWIS, a recent extension to the weighted-Bonferroni two-step GWIS by Kawaguchi et. al. was used. Rather than a pre-established number of SNPs per bin, (e.g. top 5 SNPs to the top bin, next top 10 to bin 2, etc.), SNPs are assigned to bins according to the range of

their step 1 p-value. The top bin consists of all SNPs with a step 1 p-value $0 < p \leq 5/1,000,000$, where 1 million SNPs is the number of tests corresponding to the conventional $5 \times 10^{-}8$ significance threshold. The second bin consists of all SNPs with a step 1 p-value $5/1,000,000 < p < 15/1,000,000$. In general bin B_k consists of all SNPs with a step 1 p-value $\sum_{j=1}^{k} 5 \times 2^{(}j-1)/1,000,000 < p < \sum_{j=1}^{k} +15 \times 2^{(}j-1)$. Thus, under the complete null hypothesis that no SNP interacts with alcohol to affect colorectal-cancer risk, in expectation 5 SNPs fall in bin 1, 10 SNPs fall in bin 2 and so on. In addition to binning based on partitioning the step 1 p-value range, the BWER is controlled using a Bonferroni correction by the effective number of SNPs.

We can see (Figure 1) that one SNP on bin 3 and two on bin 4 are significant (red dots). For example, in bin 3 the one SNP significant for SNP by alcohol intercation has a step 1 p-value of 3.43×10^{-05} and a step 2 p-value of 2.14×10^{-06}. Bin 3 has 1421 SNPs, which due to linkage disequilibrium are equivalent to 244 'effective' independent SNPs. The adjusted p-value is then given by:

$$p^{MTP} = \inf\{0 < \alpha \leq 1 : 2.14 \times 10^{-06} < \alpha/2^3/244\} = \min\{344 \times 2^3 2.14 \times 10^{-06}, 1\} = 0.0058 \tag{6}$$

This adjusted p-value is to be compared with the conventional 0.05 threshold, or equivalently $0.0058/1,000,000 = 5.9 \times 10^{-9}$ to the GWIS threshold of 5×10^{-8}.

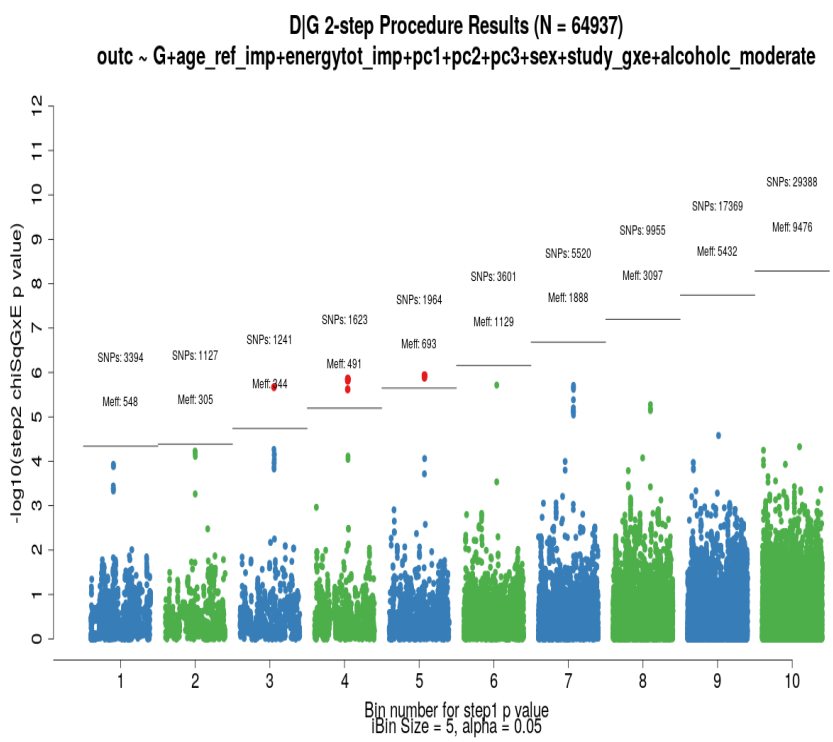

Figure 1: $G\times$ Alcohol GWIS

4 Discussion

In this paper, we presented a method to define multiple-testing adjusted p-values for two-step tests based on standard multiple-testing theory. This enables users to make valid comparisons with single-step tests, providing more informative results for interpreting GWIS results. We expect that our approach will be useful for researchers performing GWIS and will contribute to better understanding of gene-environment interactions.

Acknowledgements

The work was partially supported by NIH grants T32ES013678, P01CA196569, R01CA201407, and P30ES007048. These awards had no influence over the experimental design, data analysis or interpretation, or writing the manuscript.

References

Dai, J. Y., C. Kooperberg, M. Leblanc, and R. L. Prentice (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika 99*(4), 929–944.

Gauderman, W. J., P. Zhang, J. L. Morrison, and J. P. Lewinger (2013). Finding novel genes by testing g× e interactions in a genome-wide association study. *Genetic epidemiology 37*(6), 603–613.

Hsu, L., S. Jiao, J. Y. Dai, C. Hutter, U. Peters, and C. Kooperberg (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic Epidemiology 36*(3), 183–194.

Ionita-Laza, I., M. B. McQueen, N. M. Laird, and C. Lange (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *The American Journal of Human Genetics 81*(3), 607–614.

Kawaguchi, E. S., G. Li, J. P. Lewinger, and W. J. Gauderman (2022). Two-step hypothesis testing to detect gene-environment interactions in a genome-wide scan with a survival endpoint. *Statistics in Medicine*.

Kooperberg, C. and M. LeBlanc (2008). Increasing the power of identifying gene× gene interactions in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 32*(3), 255–263.

Murcray, C. E., J. P. Lewinger, D. V. Conti, D. C. Thomas, and W. J. Gauderman (2011). Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genetic epidemiology 35*(3), 201–210.

Murcray, C. E., J. P. Lewinger, and W. J. Gauderman (2009). Gene-environment interaction in genome-wide association studies. *American journal of epidemiology 169*(2), 219–226.

Paré, G., N. R. Cook, P. M. Ridker, and D. I. Chasman (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. *PLoS Genet 6*(6), e1000981.

Wang, J., A. Patel, J. M. Wason, and P. J. Newcombe (2021). Two-stage penalized regression screening to detect biomarker-treatment interactions in randomized clinical trials. *Biometrics*.

Zhang, P., J. P. Lewinger, D. Conti, J. L. Morrison, and W. J. Gauderman (2016). Detecting gene–environment interactions for a quantitative trait in a genome-wide association study. *Genetic epidemiology 40*(5), 394–403.