P-Values for Two-Step Genome-Wide Gene–Environment Interaction Scans

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

Abstract

Two-step testing has emerged as the leading approach for genome-wide gene-environment interaction scans, providing greater power than standard single-step methods in nearly all plausible scenarios. This gain in power has enabled several recent discoveries in gene-environment interactions. However, despite controlling the genome-wide type I error rate, two-step tests typically lack valid p-values, hindering comparisons with single-step results. We demonstrate how multiple-testing adjusted p-values can be defined for two-step tests using standard multiple-testing theory, and how these can be further scaled to allow valid comparisons with single-step tests.

Keywords: gene–environment interaction, genome-wide interaction scans, two-step test, multiple-testing, adjusted p-value

1 Introduction

Recent gene-environment ($G \times E$) findings (Carreras-Torres et al., 2023; Virolainen et al., 2023) have been facilitated by two key developments: (1) the pooling of large sample sizes across studies with harmonized genotype and exposure data, and (2) the adoption of two-step genome-wide interaction scan (GWIS) approaches. Two-step approaches have been proposed to increase power for detecting $G \times E$ interactions while controlling the family-wise error rate across the genome for the most common outcome types, including binary (Gauderman et al., 2013; Hsu et al., 2012; Kooperberg & LeBlanc, 2008; Murcray et al., 2009, 2011; Wang et al., 2022), quantitative (Paré et al., 2010; Zhang et al., 2016), and time-to-event traits (Kawaguchi et al., 2022).

In contrast to standard single-step GWIS—which tests each SNP–exposure interaction at a stringent genome-wide significance level (e.g., $\alpha = 5 \times 10^{-8}$; see (Dudbridge & Gusnanto,

*Email: lewinger@usc.edu
†Email: byuan616@usc.edu
‡Email: ekawaguc@usc.edu
§Email: jimg@usc.edu

2008))—two-step methods first use a screening step to prioritize SNPs, followed by a testing step that uses more permissive, data-adaptive significance thresholds for the most promising candidates.

This two-step strategy yields higher power under most biologically plausible $G \times E$ scenarios (Gauderman et al., 2013; Hsu et al., 2012; Kawaguchi et al., 2022; Kooperberg & LeBlanc, 2008; Murcray et al., 2009, 2011; Paré et al., 2010; Wang et al., 2022; Zhang et al., 2016). Multiple combinations of screening statistics and testing strategies have been proposed, including subset (Murcray et al., 2009) and weighted hypothesis testing (Ionita-Laza et al., 2007; Kawaguchi et al., 2023). However, none of these methods currently provide a principled framework for computing multiple-testing-adjusted p-values, thereby limiting interpretability and comparability across approaches.

In this note, we propose adjusted p-values for two-step GWIS procedures based on standard multiple-testing theory. Our approach is completely general and can be applied with any combination of the existing screening and testing strategies. These adjusted p-values enhance transparency and interpretability, allowing for valid comparisons across $G \times E$ methods.

2 Methods

Let Y denote the outcome (binary, continuous, or time-to-event), E a single environmental exposure, and $G_1, \ldots, G_M$ the genotypes for M SNPs measured or imputed in N subjects. A two-step GWIS consists of:

Step 1 (Screening): For each SNP, a screening statistic is calculated - typically based on its marginal association with the outcome Kooperberg and LeBlanc, 2008, the exposure Murcray et al., 2009 or a combination of both Gauderman et al., 2013. The screening statistic ranks SNPs according to how promising they are for $G \times E$ interaction testing; under most scenarios, SNPs truly involved in interactions will tend to exhibit larger screening statistics.

Step 2 (Testing): SNPs are formally tested for interaction using a significance level α^* that depends on their screening result in Step 1.

Two common strategies for implementing Step 2 are:

Subset Testing: Only the top $m \ll M$ SNPs exceeding a predefined screening threshold are tested for interaction. A Bonferroni correction is applied using m as the denominator:

$$\alpha^* = \frac{\alpha}{m}.$$

Weighted Hypothesis Testing: All M SNPs are tested, but with bin-specific significance thresholds based on their Step 1 rankings. Specifically, SNPs are sorted

by their screening statistic and assigned to K bins, $B_1, \ldots, B_K$, with corresponding sizes $m_1, \ldots, m_K$. The top m_1 SNPs (most promising) are placed in B_1, the next m_2 in B_2, and so on. Bin $k = 1, \ldots, K$ is tested at a bin-wise error rate (BWER), α_k, where $0 < \alpha_K < \cdots < \alpha_1$, so that higher-priority bins are tested at less stringent significance levels. The α_k values form a partition of the overall significance level α, such that $\sum_{k=1}^{K} \alpha_k = \alpha$. A common bin-wise partition of the α level and choice of bin sizes is (Ionita-Laza et al., 2007):

$$m_k = \begin{cases} B_0 \cdot 2^{k-1} & \text{for } k = 1, 2, \ldots, K-1 \\ M - B_0 \left(2^{K-1} - 1\right) & \text{for } k = K \text{ (leftover SNPs assigned to bin } K) \end{cases} \tag{1}$$

$$\alpha_k = \begin{cases} \frac{\alpha}{2} \frac{1}{2^{k-1}} & \text{for } k = 1, \ldots, K-1 \\ \frac{\alpha}{2} \frac{1}{2^{K-1}} & \text{for } k = K \end{cases} \tag{2}$$

with a typical choice of $B_0 = 5$ and where $K = \min(k \in \mathbb{N} : B_0 \cdot 2^k > M)$. Within each bin, SNPs are tested using a Bonferroni correction—that is, the bin-wise error rate α_k is further divided by the number of SNPs assigned to the bin.

For example, with these choices for α_k and m_k, the top 5 SNPs are placed in bin 1 and tested at level $\alpha_1/5$; the next 10 SNPs are placed in bin 2 and tested at $\alpha_2/10$; the next 20 SNPs go to bin 3 and are tested at $\alpha_3/20$, and so on. Overall, this scheme preserves the FWER $\leq \alpha$, while assigning more permissive significance thresholds to SNPs considered more promising based on Step 1, thereby increasing power. Recent extensions to this strategy (Kawaguchi et al., 2023) aim to reduce power loss due to SNPs with strong marginal effects but no true interaction, and to account for correlations due to linkage disequilibrium (LD) during the testing step.

2.1 Multiple Testing

The key to constructing valid p-values for two-step procedures is to recognize them as special cases of multiple testing procedures. Suppose we aim to test a family of null hypotheses $H_0(1), \ldots, H_0(M)$, for example, that SNP j, for $j = 1, \ldots, M$, does not interact with exposure E to affect outcome Y. The family-wise error rate (FWER) is the probability of making at least one false rejection:

$$\text{FWER} = P(V > 0),$$

where V denotes the number of false positives. A FWER-controlling *multiple testing procedure* (MTP) (Dudoit & van der Laan, 2008) specifies a set of test statistics T_j and corresponding critical values $c_j^{\text{MTP}}(\alpha)$ such that rejecting $H_0(j)$ when $T_j \geq c_j^{\text{MTP}}(\alpha)$ ensures FWER $\leq \alpha$.

Example: Bonferroni Procedure. Let $c_j(\alpha)$ denote the critical value for testing the individual hypothesis $H_0(j)$, such that $P(T_j \geq c_j(\alpha) \mid H_0(j)) \leq \alpha$. While c_j may vary with j, it is often constant across tests. Then, the Bonferroni multiple-testing critical value for $j = 1, \ldots, M$ is given by,

$$c_j^{\mathrm{MTP}}(\alpha) = c_j\left(\frac{\alpha}{M}\right), \tag{3}$$

which ensures control of the FWER at level α regardless of test dependence.

Example: Two-step Subset Procedure. In this procedure, only the top $m \ll M$ SNPs—selected based on a screening statistic—are tested for interaction in Step 2. A SNP that passes the screening step is declared significant if its Step 2 p-value satisfies $p_j < \alpha/m$, corresponding to a critical value:

$$c_j^{\mathrm{MTP}}(\alpha) = \begin{cases} c_j(\alpha/m), & \text{if SNP } j \text{ passes the screening step,} \\ \infty, & \text{if SNP } j \text{ does not pass the screening step.} \end{cases} \tag{4}$$

This procedure has been shown to control the FWER (Kooperberg & LeBlanc, 2008; Murcray et al., 2009), and thus defines a valid MTP in which the critical value depends on the Step 1 screening outcome.

Example: Two-step Weighted Bonferroni Procedure. In the weighted Bonferroni two-step procedure, the Step 2 critical value depends on the bin to which SNP j is assigned based on its screening statistic. Specifically, under the binning scheme defined by (2) and (1), suppose SNP j is assigned to bin B_{k_j}, which has size m_{k_j} and is allocated a bin-wise error rate α_{k_j}. The corresponding critical value is:

$$c_j^{\mathrm{MTP}}(\alpha) = c_j\left(\frac{\alpha_{k_j}}{m_{k_j}}\right). \tag{5}$$

This approach also guarantees FWER control (Gauderman et al., 2013; Ionita-Laza et al., 2007), and thus defines a valid MTP in which the critical values depend on the Step 1 screening results.

2.2 Adjusted p-values

Since two-step procedures are special cases of multiple testing procedures, their adjusted p-values can be derived from the general framework of multiple-testing-adjusted p-values. We begin by recalling the standard definition of an *unadjusted* p-value.

Classically, the p-value is defined as the probability, under the null hypothesis, of observing a test statistic T at least as extreme as the observed value t_{obs}. If the null hypothesis is $H_0 : \theta \in \Omega_0$, where Ω_0 is the subset of parameter values consistent with H_0, then

$$p = \sup_{\theta \in \Omega_0} P(T \geq t_{\mathrm{obs}}).$$

An equivalent and useful characterization (Casella & Berger, 2002) defines the p-value as the smallest significance level α at which the observed test statistic would lead to rejection of the null:

$$p = \inf \{0 < \alpha \leq 1 : T \geq c_\alpha\}, \tag{6}$$

where c_α is the critical value corresponding to level α, such that type I error is controlled at α under H_0.

This formulation extends naturally to multiple testing. Let T_j be the test statistic for hypothesis $H_0(j)$, and $c_j^{\mathrm{MTP}}(\alpha)$ the MTP-defined critical value. Then the adjusted p-value for $H_0(j)$ is:

$$p_j^{\mathrm{adj}} = \inf \left\{0 < \alpha \leq 1 : T_j \geq c_j^{\mathrm{MTP}}(\alpha)\right\}. \tag{7}$$

That is, p_j^{adj} is the smallest FWER-controlled level α at which $H_0(j)$ would be rejected.

Example: Bonferroni Adjustment. For the classical Bonferroni procedure,

$$T_j \geq c_j\left(\frac{\alpha}{M}\right) \quad \Longleftrightarrow \quad p_j \leq \frac{\alpha}{M},$$

where $p_j = \inf\{\alpha : T_j \geq c_j(\alpha)\}$ is the unadjusted p-value for $H_0(j)$. Thus, the adjusted p-value becomes:

$$p_j^{\mathrm{adj}} = \inf\left\{0 < \alpha \leq 1 : p_j \leq \frac{\alpha}{M}\right\} = \min(M \cdot p_j, 1),$$

which is the familiar Bonferroni adjusted p-value.

Example: Two-step Subset adjustment. Based on the two-step critical values in (4) and using the general definition in (7), the adjusted p-value is:

$$p_j^{\mathrm{adj}} = \begin{cases} \min(m \cdot p_j, 1), & \text{if SNP } j \text{ passes the screening,} \\ 1, & \text{otherwise,} \end{cases}$$

where m is the number of SNPs tested in Step 2.

Example: Two-step Weighted Bonferroni Adjustment. Under the weighted Bonferroni scheme, the Step 2 p-value is compared to α_{k_j}/m_{k_j}. Hence, the adjusted p-value is:

$$p_j^{\mathrm{adj}} = \min\left(m_{k_j} \cdot p_j, 1\right).$$

For the binning scheme defined by (2) and (1), this simplifies to:

$$p_j^{\mathrm{adj}} = \begin{cases} \min\left(B_0 \cdot 2^{k_j - 1} \cdot p_j,\ 1\right), & \text{if } k_j < K, \\ \min\left(\left[M - B_0(2^{K-1} - 1)\right] \cdot p_j,\ 1\right), & \text{if } k_j = K, \end{cases}$$

This shows that SNPs assigned to lower-index bins—ranked as more promising in Step 1—receive smaller adjusted p-values for the same Step 2 test p-value.

This framework for computing multiple-testing-adjusted p-values is general and can be applied to any two-step testing strategy. In the next section, we demonstrate its application using real data and an enhanced version of the two-step weighted Bonferroni method introduced by Kawaguchi et al. (Kawaguchi et al., 2023).

3 Example

To illustrate the proposed multiple-testing p-value adjustment for two-step procedures, we consider the GWIS of colorectal cancer by alcohol intake from the FIGI consortium (Jordahl et al., 2022)(Figure 1). This GWIS applied a recent extension of the weighted-Bonferroni two-step procedure (Kawaguchi et al., 2023), in which SNPs are binned based on the range of their Step 1 p-values, rather than assigning a fixed number of SNPs as in Equation 1. Specifically, SNPs are assigned to bin B_k based on which range their Step 1 p-value falls into. Bin 1 includes all SNPs with:

$$0 < p \leq \frac{5}{1{,}000{,}000},$$

where 1,000,000 reflects the effective number of SNPs tested corresponding to a conventional genome-wide threshold of 5×10^{-8}. Bin 2 includes SNPs with:

$$\frac{5}{1{,}000{,}000} < p \leq \frac{15}{1{,}000{,}000},$$

and in general, bin B_k includes SNPs with Step 1 p-values in the interval:

$$\left(\sum_{j=1}^{k-1} \frac{5 \cdot 2^{j-1}}{1{,}000{,}000}, \sum_{j=1}^{k} \frac{5 \cdot 2^{j-1}}{1{,}000{,}000} \right] .$$

Under the null hypothesis of no marginal genetic effect and no gene–environment interaction—that is, assuming no SNP individually or in combination with alcohol affects colorectal cancer risk—the expected number of SNPs per bin is 5 in bin 1, 10 in bin 2, 20 in bin 3, and so on. In expectation, this mirrors the fixed bin assignments defined in Equation 1. A key advantage of this strategy is that it prevents "bin overcrowding" by SNPs with strong marginal effects but no interaction component (Kawaguchi et al., 2023). In addition, the bin-wise error rate (BWER) is controlled using a Bonferroni correction that accounts for the *effective* number of independent SNPs in each bin, adjusting for linkage disequilibrium. In their simulation study, Kawaguchi et al. also demonstrated improved power over the standard weighted-Bonferroni two-step procedure (Kawaguchi et al., 2023).

Figure 1 shows that one SNP in bin 3, two in bin 4, and two in bin 5 are significant (highlighted in red). For instance, the significant SNP in bin 3 has a Step 1 p-value of

3.43×10^{-5}, placing it in bin 3, and a Step 2 p-value of 2.14×10^{-6}. Bin 3 includes 1,421 SNPs, corresponding to 244 effectively independent tests after LD adjustment. The adjusted p-value is computed as:

$$p^{\text{adj}} = \inf\left\{0 < \alpha \leq 1 : 2.14 \times 10^{-6} < \frac{\alpha}{2^3 \cdot 244}\right\} = \min\left(2^3 \cdot 244 \cdot 2.14 \times 10^{-6},\ 1\right) = 0.0058.$$

This adjusted p-value can be directly compared to a conventional significance level such as $\alpha = 0.05$. Alternatively, dividing it by the conventional number of effective tests—assumed to be 1,000,000 SNPs, which gives rise to the genome-wide significance level of 5×10^{-8}—yields $0.0058/1{,}000{,}000 = 5.8 \times 10^{-9}$. This value can then be compared to the standard genome-wide significance threshold of 5×10^{-8}.

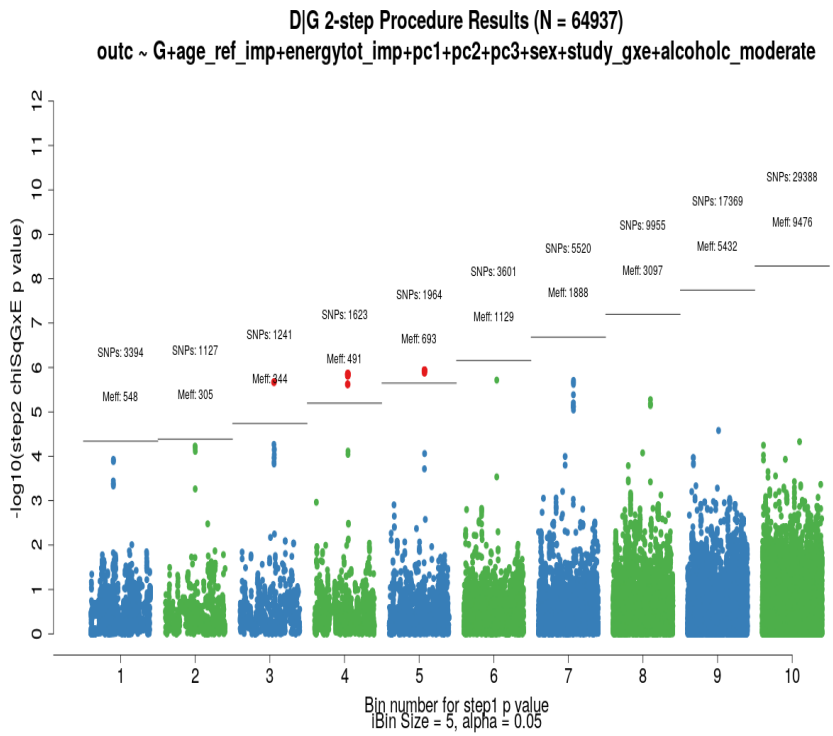

Figure 1: $G\times$ Alcohol GWIS

4 Discussion

In this paper, we presented a method for adjusted p-values for two-step tests of $G \times E$ interactions using standard multiple-testing theory. This approach enables valid comparisons with single-step tests and offers a principled and interpretable metric for assessing statistical significance in GWIS. In addition, the availability of valid adjusted p-values can facilitate downstream meta-analyses across studies using different GWIS strategies. The two-step

methods described above are all implemented in a suite of R packages. This suite includes GxEScanR (Morrison & Gauderman, 2025) to scan the genome and generate Step-1 and Step-2 statistics, and GxETest (Fu et al., 2025) to use these statistics to form 2-step tests and compute the adjusted p-values described in this paper. We anticipate that our methods and software framework will be a valuable resource for researchers conducting GWIS and will contribute to more consistent and transparent reporting of results.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability

The data and code that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval

This study used publicly available summary statistics and simulated data and did not involve human participants directly; therefore, ethics approval and consent were not required.

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

Carreras-Torres, R., Kim, A. E., Lin, Y., Díez-Obrero, V., Bien, S. A., Qu, C., Wang, J., Dimou, N., Aglago, E. K., Albanes, D., Arndt, V., Baurley, J. W., Berndt, S. I., Bézieau, S., Bishop, D. T., Bouras, E., Brenner, H., Budiarto, A., Campbell, P. T., … Gauderman, W. J. (2023). Genome-wide interaction study with smoking for colorectal cancer risk. *Cancer Epidemiology, Biomarkers & Prevention*, *32*(3), 315–329. https://doi.org/10.1158/1055-9965.EPI-22-0720

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd). Thomson Learning. http://catdir.loc.gov/catdir/enhancements/fy1302/2001025794-t.html

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Dudoit, S., & van der Laan, M. J. (2008). *Multiple testing procedures with applications to genomics*. Springer. http://swbplus.bsz-bw.de/bsz263134245cov.htm

Fu, Y., Kim, A., Morrison, J., & Gauderman, W. (2025). GxETest: An R package for powerful 2-step GxE tests [Accessed: 2025-07-30].

Gauderman, W. J., Zhang, P., Morrison, J. L., & Lewinger, J. P. (2013). Finding novel genes by testing g × e interactions in a genome-wide association study. *Genetic Epidemiology*, *37*(6), 603–613. https://doi.org/10.1002/gepi.21748

Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U., & Kooperberg, C. (2012). Powerful cocktail methods for detecting genome-wide gene–environment interaction. *Genetic Epidemiology*, *36*(3), 183–194. https://doi.org/10.1002/gepi.21610

Ionita-Laza, I., McQueen, M. B., Laird, N. M., & Lange, C. (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. https://doi.org/https://doi.org/10.1086/519748

Jordahl, K., Shcherbina, A., Kim, A., Su, Y., Lin, Y., Wang, J., Qu, C., Albanes, D., Arndt, V., Baurley, J., Berndt, S., Bien, S., Bishop, D., Bouras, E., Brenner, H., Buchanan, D., Budiarto, A., Campbell, P., Carreras-Torres, R., … Peters, U. (2022). Beyond gwas of colorectal cancer: Evidence of interaction with alcohol consumption and putative causal variant for the 10q24.2 region. *Cancer Epidemiology, Biomarkers & Prevention*, *31*(5), 1077–1089. https://doi.org/10.1158/1055-9965.EPI-21-1003

Kawaguchi, E. S., Kim, A. E., Lewinger, J. P., & Gauderman, W. J. (2023). Improved two-step testing of genome-wide gene–environment interactions. *Genetic Epidemiology*, *47*(2), 152–166. https://doi.org/10.1002/gepi.22509

Kawaguchi, E. S., Li, G., Lewinger, J. P., & Gauderman, W. J. (2022). Two-step hypothesis testing to detect gene–environment interactions in a genome-wide scan with a survival endpoint. *Statistics in Medicine*, *41*(7), 1214–1228. https://doi.org/10.1002/sim.9319

Kooperberg, C., & LeBlanc, M. (2008). Increasing the power of identifying gene × gene interactions in genome-wide association studies. *Genetic Epidemiology*, *32*(3), 255–263. https://doi.org/10.1002/gepi.20300

Morrison, J., & Gauderman, W. (2025). GxEScanR: An R package for genomewide GxE interaction scans [Accessed: 2025-07-30].

Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C., & Gauderman, W. J. (2011). Sample size requirements to detect gene–environment interactions in genome-wide association studies. *Genetic Epidemiology*, *35*(3), 201–210. https://doi.org/10.1002/gepi.20569

Murcray, C. E., Lewinger, J. P., & Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, *169*(2), 219–226. https://doi.org/10.1093/aje/kwn353

Paré, G., Cook, N. R., Ridker, P. M., & Chasman, D. I. (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 Genetics*, *6*(6), e1000981. https://doi.org/10.1371/journal.pgen.1000981

Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T., & Kottyan, L. C. (2023). Gene-environment interactions and their impact on human health. *Genes & Immunity*, *24*, 1–11. https://doi.org/10.1038/s41435-022-00192-6

Wang, J., Patel, A., Wason, J. M. S., & Newcombe, P. J. (2022). Two-stage penalized regression screening to detect biomarker–treatment interactions in randomized clinical trials. *Biometrics*, *78*(1), 141–150. https://doi.org/10.1111/biom.13424

Zhang, P., Lewinger, J. P., Conti, D. V., Morrison, J. L., & Gauderman, W. J. (2016). Detecting gene-environment interactions for a quantitative trait in a genome-wide association study. *Genetic Epidemiology*, *40*(5), 382–393. https://doi.org/10.1002/gepi.21977