


## RESEARCH ARTICLE

# Beyond the traditional simulation design for evaluating type 1 error control: From the “theoretical” null to “empirical” null

Ting Zhang<sup>1</sup> | Lei Sun<sup>1,2</sup> 
<sup>1</sup>Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

<sup>2</sup>Department of Statistical Sciences, Faculty of Arts and Science, University of Toronto, Toronto, Ontario, Canada
**Correspondence**

Lei Sun, Department of Statistical Sciences, Faculty of Arts and Science, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada. Email: sun@utstat.toronto.edu

**Funding information**

Canadian Institutes of Health Research, Grant/Award Number: MOP-310732-G-CEAA-117978; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-250053

**Abstract**

When evaluating a newly developed statistical test, an important step is to check its type 1 error (T1E) control using simulations. This is often achieved by the standard simulation design S0 under the so-called “theoretical” null of no association. In practice, the whole-genome association analyses scan through a large number of genetic markers (Gs) for the ones associated with an outcome of interest (Y), where Y comes from an alternative while the majority of Gs are not associated with Y; the Y – G relationships are under the “empirical” null. This reality can be better represented by two other simulation designs, where design S1.1 simulates Y from an *alternative* model based on G, then evaluates its association with independently generated  $G_{\text{new}}$ ; while design S1.2 evaluates the association between permuted Y and G. More than a decade ago, Efron (2004) has noted the important distinction between the “theoretical” and “empirical” null in false discovery rate control. Using scale tests for variance heterogeneity, direct univariate, and multivariate interaction tests as examples, here we show that not all null simulation designs are equal. In examining the accuracy of a likelihood ratio test, while simulation design S0 suggested the method being accurate, designs S1.1 and S1.2 revealed its increased empirical T1E rate if applied in real data setting. The inflation becomes more severe at the tail and does not diminish as sample size increases. This is an important observation that calls for new practices for methods evaluation and T1E control interpretation.

**KEYWORDS**

interaction, simulation, type 1 error, variance heterogeneity, whole-genome association scans

## 1 | INTRODUCTION

Type 1 error (T1E) control evaluation using simulations is always the first step in understanding the performance of any newly developed statistical test. To formulate the problem more precisely, let us consider the current large-scale

genome-wide association studies (GWAS) or next-generation sequencing (NGS) studies of complex and heritable traits. These studies scan through millions or more genetic markers (Gs) across the genome for the ones associated with a trait of interest (Y), while accounting for environmental effects; without loss of generality we assume these genetic markers

are single-nucleotide polymorphisms (SNPs). Many  $Y - G$  association tests have been developed, and they often require the assumption of (approximately) normally distributed errors to maintain an accurate T1E, with some tests being more robust than others. For example, Bartlett test for variance heterogeneity has been shown to have large inflated T1E rates when the error term  $e$  follows a  $t$ - or  $\chi^2$ -distribution (Struchalin, Dehghan, Witteman, van Duijn, & Aulchenko, 2010), and the likelihood ratio test (LRT) is similarly sensitive to nonnormal errors (Cao, Wei, Bailey, Kauwe, & Maxwell, 2014), while Levene's test appears to be more robust (Soave & Sun, 2017; Soave et al., 2015).

Let  $T$  be the associate test statistic to be evaluated. Standard T1E simulation design, denoted as S0, first generates phenotype data  $Y_0 = e$  under the “theoretical” null of no association. It then independently generates genotype data  $G_{\text{new}}$ , applies a fitted model of  $Y_0 = b_G G_{\text{new}} + \epsilon$  to test the  $H_0 : b_G = 0$ , derives the asymptotic distribution of  $T$  denoted as  $f_0(T)$ , obtains the corresponding  $p$ -value for each simulated replicate, and finally estimates the empirical T1E rate of  $T$ . (For notation simplicity and without loss of generality, for the moment we omit the intercept and additional covariates  $Z$ s from the regression models.)  $T$  is considered sound if the T1E rate is well controlled under the  $e \sim N(0, \sigma^2)$  assumption, and its robustness is then evaluated by assuming other distributional forms for  $e$ . Given well-controlled T1E under normality, power will then be studied by generating phenotype under an alternative,  $Y_1 = \beta_G G + e$ , where typically  $e \sim N(0, \sigma^2)$ . In that case, one applies the fitted model of  $Y_1 = b_G G + \epsilon$  to obtain the test statistic  $T$  under the alternative, calculates the corresponding  $p$ -value based on  $f_0(T)$ , and finally estimates the empirical power of  $T$ . Combining the two, one would then expect that, in practice,  $T$  maintains good T1E control for a null SNP and has certain amount of power for an alternative one. However, for a real GWAS, the relationship between the phenotype and a null SNP is under the “empirical” null, which we describe below. This inconsistency in T1E evaluation and interpretation is the focus of our study.

In practice, a whole-genome association scan receives an *empirical*  $Y_1$  that comes from an *alternative*, influenced by one or more  $G$ s. Among the million or more SNPs to be analyzed, most are not associated with  $Y_1$ . However, the corresponding no phenotype-genotype association is not accurately reflected by the “theoretical” null simulation design S0 as described above. Now consider two alternative simulation designs to evaluate T1E control of the test statistic  $T$ . Design S1.1 first generates  $Y_1 = \beta_G G + e$  from an alternative, where  $e \sim N(0, \sigma^2)$ . It then independently generates  $G_{\text{new}}$  for a new SNP, fits the model of  $Y_1 = b_G G_{\text{new}} + \epsilon$ , calculates  $T$  and its corresponding  $p$ -value based on  $f_0(T)$ , and finally estimates the T1E rate. Design S1.2 permutes the simulated  $Y_1$

and evaluates T1E from  $Y_1^{\text{perm}} = b_G G + \epsilon$ . Clearly, the “empirical” null simulation designs S1.1 and S1.2 mimic the real data far better than the “theoretical” null S0. Thus, an important question can be asked as to whether the S1.1 and S1.2 designs lead to similar T1E conclusion for  $T$  as the S0 design. In particular, even if the  $e \sim N(0, \sigma^2)$  assumption is true in the generating model and  $T$  appears to be accurate based on the S0 evaluation, do we expect  $T$  to perform equally well when applied to real data? The answer would depend on the sensitivity of the test statistic  $T$  used.

Efron (2004) has brought up the discussion of the “theoretical” versus “empirical” null more than a decade ago. Focusing on controlling the false discovery rate, Efron (2004) outlined several possible sources of nonnormality including unobserved covariates and hidden correlation, and he proposed an empirical Bayes approach to the problem. Here, we study the practical implications of T1E evaluation based on S0, the commonly used “theoretical” null simulation design, in the context of whole-genome association scans. We show that while a test  $T$  may appear to be accurate under S0 and assuming normality, it can have incorrect T1E rates under the “empirical” null of S1.1 or S1.2, also “assuming normality.” The fundamental cause of the discrepancy is that, even if the error term in the generating model of  $Y_1 = \beta_G G + e$  is normal,  $e \sim N(0, \sigma^2)$ , marginally  $Y_1$  may not be normal. Thus, in evaluating the null  $Y_1 - G_{\text{new}}$  relationship using the fitted model of  $Y_1 = b_G G_{\text{new}} + \epsilon$  (or  $Y_1^{\text{perm}} - G$  using  $Y_1^{\text{perm}} = b_G G + \epsilon$ ), the true null distribution of  $T$  may not be  $f_0(T)$  which was derived for  $Y_0 - G_{\text{new}}$  under the “theoretical” null. Essentially, the  $Y_1 = b_G G_{\text{new}} + \epsilon$  model is misspecified if  $\epsilon$  was assumed to be normal. Inference of the location parameter  $b_G$ , the main effect of a SNP is generally quite robust to model assumptions (Khan & Rayner, 2003). However, for emerging association tests that are designed to improve power by going beyond the first moment, the distinction between the “theoretical” and “empirical” null in T1E evaluation can be consequential.

As a proof-of-principle, we will focus on testing gene-environment ( $G \times E$ ) effects; testing gene-gene ( $G \times G$ ) effects is similar. Such interaction effects are expected for complex traits,  $Y = \beta_G G + \beta_E E + \beta_{GE}(G \times E) + e$ , and we study three scenarios for testing  $G \times E$  interaction.

The first scenario assumes that the data on  $E$  are not available in practice. Thus, direct  $G \times E$  interaction analysis is not possible. In that case, because the unmodeled interaction induces variance heterogeneity in  $Y$  when conditional only on  $G$ , scale tests such as Levene's test, originally developed for model diagnostics, can be used to *indirectly* test for the interaction effect. We will investigate the several scale tests recently proposed for this purpose (Aschard, Zaitlen, Tamimi, Lindström, & Kraft, 2013;

Cao et al., 2014; Paré, Cook, Ridker, & Chasman, 2010; Soave et al., 2015). It is worth noting that the causes of variance heterogeneity are multifaceted beyond potential interactions (Dudbridge & Fletcher, 2014; Sun, Elston, Morris, & Zhu, 2013; Wood et al., 2014). We show that, depending on the robustness of a test, T1E conclusion may differ between the “theoretical” and “empirical” null.

The second scenario assumes  $E$  was available for direct modeling of the  $G \times E$  interaction effect. Previously, Voorman, Lumley, McKnight, and Rice (2011) and Rao and Province (2016) showed that T1E of testing for interaction effect in a whole-genome scan can be more variable than testing for main effect. The authors examined  $SNP_{\text{nonrepeating}} \times SNP_{\text{repeating}}$  analysis, where  $SNP_{\text{repeating}}$  represents a fixed SNP and its interactions with all other SNPs are of primary interest. Statistically, this is similar to  $G \times E$  that we will be studying here, because  $E$  does not vary between SNPs in a genome-wide  $G \times E$  interaction scan. Focusing on inflated or deflated genomic inflation factor  $\lambda_{GC}$  (Devlin & Roeder, 1999), Rao and Province (2016) demonstrated a larger variation in  $\lambda_{GC}$  (similar to a larger variation in T1E rates between different whole-genome association scans), when testing the interaction effect as compared to the main effect under the “theoretical” null. They attributed this to dependency between the interaction test statistics, because  $SNP_{\text{repeating}}$  (or the fixed  $E$  in our setting) is not changing between the tests. They also noted that increasing sample size mitigates the problem. Here, we use this opportunity to revisit direct  $G \times E$  interaction testing in a GWAS setting. We show that, while T1E rates are indeed more variable between simulation replicates (i.e., between genome-wide interaction scans) under the conventional “theoretical” null, due to dependency between the tests as in Rao and Province (2016), the average T1E rate is correct regardless of the sample size. However, under the “empirical” null, a different picture emerges as in the scale test setting.

The third scenario extends the above to a multivariate setting, where the fixed  $E$  interacts with multiple different  $G$ s as in gene-based interaction studies. We will examine sequence kernel association test (SKAT)-type of variance component test (Wu et al., 2011), together with burden-type of sum test (Madsen & Browning, 2009), and the classical  $F$ -test, that jointly evaluate multiple interaction effects. Generally speaking, departure from normality is of a lesser concern when an outcome  $Y$  is influenced by multiple genetic and environmental factors (Falconer, 1960; Mackay, 2009). However, we will show that the distinction between the “theoretical” and “empirical” null remains relevant for multivariate models.

In Section 2, we first describe the three scenarios for interaction testing, including when  $E$  is missing, and when  $E$  is known and interacts with one  $G$  or multiple

$G$ s. For clarification, we call these *interaction scenarios*. Under each interaction scenario, we briefly review all the statistical tests to be investigated. We then describe the three simulation designs for evaluating T1E control, namely, S0 for the “theoretical” null design, and S1.1 and S1.2 for the “empirical” null designs; we call these *T1E simulation designs*. The implementation of the different T1E simulation designs depends on the test of interest. Thus, we also describe in details how the data are being generated for each of the three interaction scenarios, and for each of the three T1E simulation designs; we call these *data-generating models*. In Section 3, we conduct simulation studies, provide the corresponding numerical results, and reveal the existing problems in T1E evaluation based on the “theoretical” null simulation design S0. We show that (a) a T1E conclusion drawn from S0 can be different from that based on the two alternative “empirical” null simulation designs S1.1 and S1.2, (b) the T1E discrepancy can remain as sample size increases, and (c) the T1E issue may be more severe at the tail. The root cause of the issue demonstrated in this study is due to subtle model misspecifications. A test might be shown to be accurate under the idealistic “theoretical” null S0. However, its true T1E behavior, when applied to real whole-genome association scans, is only uncovered through the “empirical” null designs S1.1 and S1.2. We make additional remarks in Section 4.

## 2 | METHODS AND MATERIALS

### 2.1 | Three $G \times E$ interaction scenarios and corresponding statistical tests

For association analysis of a complex trait  $Y$  using a sample of size  $n$ , we first define genotype data  $G_i$  for individual  $i$  at each SNP under the study. As in convention,  $G_i$  denotes the number of copies of the minor allele, coded additively as  $G_i = 0, 1$ , and  $2$ . Also as in convention,  $G_i$  is assumed to come from a binomial distribution,  $G_i \sim \text{Binomial}(2, f)$ , where  $f$  is the minor allele frequency (MAF).

In the simplest case, we might assume the true generating model for the trait to be

$$Y = \beta_G G + \beta_E E + \beta_{GE} (G \times E) + e, \quad e \sim N(0, \sigma^2), \quad (1)$$

where  $\beta_G$  is the main effect of a causal SNP,  $\beta_E$  is the environmental effect, and  $\beta_{GE}$  is the gene–environment interaction effect. Note that the error term in the true data-generating model is denoted as  $e$  and assumed to be normal. We wish to identify the SNP whose genotype  $G$  influences  $Y$ .

### 2.1.1 | $G \times E$ interaction scenario 1: Single $G$ , and $E$ missing

Suppose information regarding  $E$  was not available, then the working or fitted model can only account for the main effect of  $G$ ,

$$Y = b_G G + \epsilon. \quad (2)$$

However, it is straightforward to show that variances of  $Y$  stratified by the three genotype groups of  $G$  differ, if  $\beta_{GE} \neq 0$ ,

$$\text{Var}(Y|G) = (\beta_E + \beta_{GE} G)^2 \text{Var}(E) + \sigma^2 = \sigma_G^2 = \text{Var}(\epsilon). \quad (3)$$

That is,  $\epsilon$  in the fitted model (2) can behave quite differently from  $e$ , the error term in the generating model (1). Thus, when  $E$  is missing and direct interaction modeling is not feasible, scale tests for heteroscedasticity can be utilized to identify  $G$  associated with variance of  $Y$  (Paré et al., 2010). A joint location-scale testing framework can provide robustness against either  $\beta_G = 0$  or  $\beta_{GE} = 0$ , and it can improve power if both main and interaction effects are present (Soave & Sun, 2017; Soave et al., 2015). Here we focus on studying the more sensitive scale test, because the power of the joint test depends on the individual components.

Different scale tests have been studied in this context, and chief among them are the Levene's test (Levene, 1960) considered by Paré et al. (2010) and Soave et al. (2015), and the LRT considered by Cao et al. (2014). Levene's test for variance heterogeneity between  $k$  groups is an analysis of variance of the absolute deviation of each observation  $y_i$  from its group mean or median. Under the null of variance homogeneity and assuming normality, the resulting test statistic *Levene* follows a  $F(k-1, n-k)$  distribution, and it is asymptotically  $\chi^2_{k-1}/(k-1)$  distributed,  $k=3$  in our case. Using median instead of mean to measure the spread within each group is more robust to nonnormality, particularly for  $t$ -distributed or skewed data (Brown & Forsythe, 1974; Soave & Sun, 2017). And we will be using the median version of *Levene* in the remaining paper.

The variance LRT considered by Cao et al. (2014) contrasts the null model of no variance difference with the alternative model,

$$Y = b_G G + \epsilon, \epsilon \sim N(0, \sigma^2) \quad \text{versus} \\ Y = b_G G + \epsilon, \epsilon \sim N(0, \sigma_G^2), \quad (4)$$

and conduct the corresponding LRT for  $H_0: \sigma_{G=0}^2 = \sigma_{G=1}^2 = \sigma_{G=2}^2$ . The corresponding test statistic  $LRT_v$  is asymptotically  $\chi^2_{(2)}$  distributed, under the null of

homoscedasticity and normality. That is,  $f_0(T) = \chi^2_{(2)}$  where the test statistic  $T$  is  $LRT_v$ . For the purpose of comparison, we will also examine the LRT ( $LRT_m$ ) and score test ( $Score_m$ ) for testing the main effect,  $H_0: b_G = 0$ .

Cao et al. (2014) has pointed out that  $LRT_v$  is sensitive to the normality assumption, but under normality they have demonstrated that  $LRT_v$  has good T1E control. However, they implicitly assumed that the test would work well as long as the error term  $e$  in the phenotype-generating model is normally distributed, regardless of the “theoretical” or “empirical” null. The work here is to show why  $LRT_v$  may have T1E issue in practice. Indeed, Soave et al. (2015) applied  $LRT_v$  to a GWAS of lung function measures in 1,409 individuals with cystic fibrosis. Despite the fact that the lung measures were approximately normally distributed and permuted before the variance association analysis, the histogram of GWAS  $p$ -values clearly showed an increased T1E (Supporting Information Figure S2.G of Soave et al., 2015); the actual application was a joint  $LRT_m$  and  $LRT_v$  test, but the T1E issue was due to the  $LRT_v$  component.

### 2.1.2 | $G \times E$ interaction scenario 2: Single $G$ , and $E$ known

When  $E$  is known and its data were collected, we can directly test for the  $G \times E$  interaction effect by contrasting the following two fitted models:

$$Y = b_G G + b_E E + \epsilon \quad \text{versus} \\ Y = b_G G + b_E E + b_{GE}(G \times E) + \epsilon, \quad (5)$$

where  $\epsilon \sim N(0, \sigma^2)$ . The corresponding  $LRT_{GE}$  and  $Score_{GE}$  tests are both asymptotically  $\chi^2_{(1)}$  distributed under the null of no interaction effect. That is, under the “theoretical” null that  $\beta_{GE} = 0$  in the true phenotype-generating model (1),  $f_0(T) = \chi^2_{(1)}$ , where the test statistic  $T$  is either  $LRT_{GE}$  or  $Score_{GE}$ . This is an important note for our study here. We will show that if  $\beta_{GE} \neq 0$  but the association test is conducted for  $G_{\text{new}}$  (of a new SNP generated independently of  $Y$ ) under the “empirical” null, continuing using  $\chi^2_{(1)}$  to obtain  $p$ -value can lead to T1E result that is quite different from that obtained under the “theoretical” null.

### 2.1.3 | $G \times E$ interaction scenario 3: Multiple $G$ s, and $E$ known

A complex trait is influenced by multiple factors. For example, intelligence has been found to be associated with more than a hundred of SNPs (Dadaev et al., 2018; Hill et al., 2018). Without loss of generality, the simple phenotype-generating model (1) can be extended to include  $J$  SNPs,



$$Y = \sum_{j=1}^J \beta_{G_j} G_j + \beta_E E + \sum_{j=1}^J \beta_{GE_j} (G_j \times E) + e, e \sim N(0, \sigma^2), \quad (6)$$

where  $\beta_{G_j}$  is the main effect of SNP  $j$ ,  $\beta_E$  is the environmental effect, and  $\beta_{GE_j}$  is the interaction effect of  $G_j \times E$ .

To detect any of the  $J$  interaction effects, the classical  $F$ -test, denoted as  $F_{GE}$ , can be applied to the following fitted model:

$$Y = \sum_{j=1}^J b_{G_j} G_j + b_E E + \sum_{j=1}^J b_{GE_j} (G_j \times E) + \epsilon, \quad (7)$$

and test  $H_0 : b_{GE_1} = \dots = b_{GE_J} = 0$  simultaneously. Under the “theoretical” null that  $\beta_{GE_j} = 0 \forall j$  in the phenotype-generating model (6), the  $F_{GE}$  test statistic is  $F(J, n - 2J - 1)$  distributed and is known to have good T1E control. However, it is not clear how the test would behave in practice under the “empirical” null. In that case, a set of  $G_{\text{new}}$ ’s under the study do not interact with  $E$  to influence  $Y$ , but  $\beta_{GE_j} \neq 0$  and  $e \sim N(0, \sigma^2)$  in the true phenotype-generating model (6).

For rare variants with low MAFs, multivariate testing is common and different methods have been proposed, such as the burden-type of sum test (Li & Leal, 2008; Madsen & Browning, 2009; Price et al., 2010), and SKAT-type of variance component test (Wu et al., 2011); see Derkach, Lawless and Sun (2014) for a review. Although these tests were originally developed for main effects of rare variants, they can be applied to common variants (Lin, Lee, Christiani, & Lin, 2013) and used for interaction effects (Section of Lin et al., 2016).

Briefly, the  $Burden_{GE}$  interaction test first aggregates the allele counts across the  $J$  SNPs to obtain  $G^* = \sum_{j=1}^J G_j$ . It then tests  $H_0 : b_{G^*E} = 0$ , using the fitted model of

$$Y = b_{G^*} G^* + b_E E + b_{G^*E} (G^* \times E) + \epsilon. \quad (8)$$

However, the  $Burden_{GE}$  test has T1E issue even under the “theoretical” null; this was studied in Section 3 of Lin et al. (2016). For completeness of method evaluation, we include  $Burden_{GE}$  in our study of “theoretical” versus “empirical” null.

Extending the earlier SKAT work for main effects, Lin et al. (2016) then used it to study interaction effects. Without going to the technical details, the main component of the  $SKAT_{GE}$  interaction test is  $\sum_{j=1}^J w_j \text{Score}^2(b_{GE_j})$ , where  $\text{Score}(b_{GE_j})$  is the score test statistic for each  $b_{GE_j}$  in the fitted model (7), and  $w_j$

depends on the MAF of SNP  $j$ . We will study the  $SKAT_{GE}$ , as well as the  $Burden_{GE}$  and  $F_{GE}$  for gene-based interaction studies of both common and rare variants.

## 2.2 | Three T1E simulation designs: The “theoretical” null S0, and the “empirical” null S1.1 and S1.2

The three simulation designs for evaluate T1E can be conceptualized as follows:

|                             |   |
|-----------------------------|---|
| The ‘theoretical’ null’ S0: | simulates $Y_0$ under the null, independently generates $G_{\text{new}}$ , and evaluates T1E from the $Y_0 - G_{\text{new}}$ null relationship.   |
|                             | The ‘empirical’ null’ S1.1: simulates $Y_1$ under an alternative based on $G$ , independently generates $G_{\text{new}}$ , and evaluates T1E from the $Y_1 - G_{\text{new}}$ null relationship. |
|                             | The ‘empirical’ null’ S1.2: permutes the $Y_1$ , and evaluates T1E from the $Y_1^{\text{perm}} - G$ null relationship.  |

(9)

The exact implementation depends on the test to be evaluated. Thus, we describe below, in detail, how data are being generated for the three T1E simulation designs (S0, S1.1, and S1.2), and under the different interaction testing scenarios ( $E$  missing or not, and single or multiple  $G$ s).

## 2.3 | Data-generating models for the three T1E simulation designs and under each of the three interaction scenarios

### 2.3.1 | Scenario 1: Single $G$ , and $E$ missing

Consider the true phenotype-generating model (1), the “theoretical” null simulation design S0 assumes  $\beta_{GE} = 0$ ; for simplicity but without loss of generality we also assume  $\beta_G = 0$ . Thus, S0 simulates phenotype data using

$$Y_0 = \beta_E E + e, e \sim N(0, \sigma^2). \quad (10)$$

It then independently simulates genotype data for an nonassociated SNP,

$$G_{\text{new}} \sim \text{Binomial}(2, f), \quad (11)$$

where  $f$  is the MAF. Finally, because  $E$  was assumed missing in practice, S0 uses the following fitted model:

$$Y_0 = b_G G_{\text{new}} + \epsilon, \epsilon \sim N(0, \sigma_G^2), \quad (12)$$

to detect the variance heterogeneity present in  $\epsilon$ , based on the *Levene* and *LRT<sub>v</sub>* tests as described in Section 2.1.1.

The “empirical” null simulation design S1.1, however, first simulates phenotype data under an alternative. That is,

$$Y_1 = \beta_G G + \beta_E E + \beta_{GE}(G \times E) + e, e \sim N(0, \sigma^2). \quad (13)$$

Note that  $G$  is a truly associated SNP, and  $G \sim \text{Binomial}(2, f)$ , where the MAF  $f$  does not have to be the same as that of  $G_{\text{new}}$  above. S1.1 then independently simulates genotype data for a nonassociated SNP  $G_{\text{new}}$  as in (11). Similarly, because  $E$  was assumed missing in practice, S1.1 uses the fitted model

$$Y_1 = b_G G_{\text{new}} + \epsilon, \epsilon \sim N(0, \sigma_G^2), \quad (14)$$

to conduct the *Levene* and *LRT<sub>v</sub>* variance tests.

The “empirical” null simulation design S1.2 first permutes the  $Y_1$  generated above. Because  $Y_1^{\text{perm}}$  is no longer associated with the  $Y_1$ -generating  $G$ , it then uses the fitted model

$$Y_1^{\text{perm}} = b_G G + \epsilon, \epsilon \sim N(0, \sigma_G^2), \quad (15)$$

to assess T1E control of a test.

In summary, the true phenotype-generating models are  $Y_0 = \beta_E E + e$  or  $Y_1 = \beta_G G + \beta_E E + \beta_{GE}(G \times E) + e$ , where  $e \sim N(0, \sigma^2)$  in both cases.  $G_{\text{new}}$  is genotype data of a nonassociated SNP. The three T1E simulation designs estimate T1E rate of a heteroscedasticity test for  $\text{Var}(\epsilon)$  using, respectively, the fitted models of

$$\begin{cases} \text{S0: } Y_0 = b_G G_{\text{new}} + \epsilon \\ \text{S1.1: } Y_1 = b_G G_{\text{new}} + \epsilon \\ \text{S1.2: } Y_1^{\text{perm}} = b_G G + \epsilon. \end{cases} \quad (16)$$

### 2.3.2 | Scenario 2: Single $G$ , and $E$ known

In this case, the data-generating model is the same as above, namely,

$$Y_0 = \beta_E E + e, e \sim N(0, \sigma^2), \text{ or } Y_1 = \beta_G G + \beta_E E + \beta_{GE}(G \times E) + e, e \sim N(0, \sigma^2).$$

Because  $E$  is known in this second interaction scenario, the three T1E simulation designs estimate the T1E rate by testing  $H_0: b_{GE} = 0$  using, respectively, the fitted models of

$$\begin{cases} \text{S0: } Y_0 = b_G G_{\text{new}} + b_E E + b_{GE}(G_{\text{new}} \times E) + \epsilon \\ \text{S1.1: } Y_1 = b_G G_{\text{new}} + b_E E + b_{GE}(G_{\text{new}} \times E) + \epsilon \\ \text{S1.2: } Y_1^{\text{perm}} = b_G G + b_E E^{\text{perm}} + b_{GE}(G \times E^{\text{perm}}) + \epsilon. \end{cases} \quad (17)$$

Note that  $E^{\text{perm}}$  represents the fact that the permutation must be performed jointly for  $Y_1$  and  $E$ . This is to maintain the  $Y_1 - E$  relationship while breaking the  $Y_1 - G$  association.

For the “theoretical” null S0, we assumed  $Y_0 = \beta_E E + e$  without the main  $G$  effect in the true generating model (similar to Model I of Rao & Province, 2016). Alternatively, we could consider  $Y_0 = \beta_G G + \beta_E E + e$  with the main effect (Model II of Rao & Province, 2016). In that case, the fitted model would be  $Y_0 = b_G G + b_E E + b_{GE}(G \times E) + \epsilon$ , and  $b_{GE}$  is expected to be zero. However, this difference in the “theoretical” null design regarding the main effect does not affect our study of the interaction effect.

The work of Rao and Province (2016) studied the effect of dependency in  $G_{\text{nonrepeating}} \times G_{\text{repeating}}$  interaction analysis on T1E control. Similarly, we can assume  $E$  is fixed to represent the fact that, in a real genome-wide  $G \times E$  interaction scan, the  $E$  does not change from SNP to SNP. However, as demonstrated below, we show that this dependency is not the source of the T1E issue addressed here.

### 2.3.3 | Scenario 3: Multiple $G$ s, and $E$ known

By now, it should be clear how the three T1E simulation designs would be implemented in this setting. The true phenotype-generating models are

$$Y_0 = \beta_E E + e, e \sim N(0, \sigma^2), \text{ or} \quad (18)$$

$$Y_1 = \sum_{j=1}^J \beta_{G_j} G_j + \beta_E E + \sum_{j=1}^J \beta_{GE_j}(G_j \times E) + e, e \sim N(0, \sigma^2). \quad (19)$$

For the  $SKAT_{GE}$  and  $F_{GE}$  tests, the three T1E simulation designs estimate the T1E rate of jointly testing  $H_0: b_{GE_j} = 0, \forall j = 1, \dots, J$ , using, respectively, the fitted models of

**TABLE 1** Summary of the two data-generating models for indirect and direct  $G \times E$  interaction testing, and evaluating the “theoretical” null simulation designs S0 versus the two “empirical” null simulation designs S1.1 and S1.2, as described in Sections 2.3.1 and 2.3.2

|                                      | Introduce variance heterogeneity by $\sigma_G^2$<br>(Cao et al., 2014)   | Introduce variance heterogeneity by $G \times E$<br>(Aschard et al., 2013)<br>Or, directly test $\beta_{GE}$ (assuming $E$ was available)         |
|--------------------------------------|--|---|
| Null model for S0                    | $Y_0 = e, e \sim N(0, \sigma^2)$   | $Y_0 = \beta_E E + e, e \sim N(0, \sigma^2)$  |
| Alternative models for S1.1 and S1.2 | $Y_1 = \beta_G G + e, e \sim N(0, \sigma_G^2)$   | $Y_1 = \beta_G G + \beta_E E + \beta_{GE} G \times E + e, e \sim N(0, \sigma^2)$  |
| Parameters                           | MAF = 0.4 for $G$ and $G_{\text{new}}$<br>$\beta_G = 0.3, \sigma_0^2 = 0.23, \sigma_1^2 = 0.25, \sigma_2^2 = 0.29$ | MAF = 0.4 for $G$ and $G_{\text{new}}$<br>$\mathbb{P}(E = 1) = 0.3, \beta_G = 0.01, \beta_E = 0.3, \beta_{GE} = 0.1, 0.2, \dots, 1, \sigma^2 = 1$ |
| Sample size                          | $n = 10^3$ or $10^4$   | $n = 10^3$ or $10^4$  |
| Nominal T1E level                    | $\alpha = 0.05$  | $\alpha = 0.05, 0.01, 0.001, 10^{-5}$   |
| Replications                         | nrep.in = $10^5$<br>nrep.out = 100   | nrep.in = $10^5$ , or $10^7$ for $\alpha = 10^{-5}$<br>nrep.out = 100   |

If  $E$  was available for direct  $G \times E$  testing, the Aschard et al. (2013) model coincides with Model I of Rao and Province (2016), except  $E$  was  $G_{\text{nonrepeating}}$ . T1E rate is first estimated from nrep.in simulation replicates in an inner loop in which  $E$  is fixed (similar to one whole-genome  $G \times E$  interaction scan), then averaged over nrep.out simulation replicates in an outer loop in which  $E$  varies.

MAF: minor allele frequency.

$$\left\{ \begin{array}{l} \text{S0: } Y_0 = \sum_{j=1}^J b_{G_j} G_{\text{new}_j} + b_E E \\ \quad + \sum_{j=1}^J b_{GE_j} (G_{\text{new}_j} \times E) + \epsilon \\ \text{S1.1: } Y_1 = \sum_{j=1}^J b_{G_j} G_{\text{new}_j} + b_E E \\ \quad + \sum_{j=1}^J b_{GE_j} (G_{\text{new}_j} \times E) + \epsilon \\ \text{S1.2: } Y_1^{\text{perm}} = \sum_{j=1}^J b_{G_j} G_j + b_E E^{\text{perm}} \\ \quad + \sum_{j=1}^J b_{GE_j} (G_j \times E^{\text{perm}}) + \epsilon. \end{array} \right. \quad (20)$$

For the  $Burden_{GE}$  test based on  $G^* = \sum_{j=1}^J G_j$ , the three T1E simulation designs estimate the T1E rate of testing  $H_0 : b_{G^*E} = 0$ , using, respectively, the fitted models of

$$\left\{ \begin{array}{l} \text{S0: } Y_0 = b_{G^*} G_{\text{new}}^* + b_E E + b_{G^*E} (G_{\text{new}}^* \times E) + \epsilon \\ \text{S1.1: } Y_1 = b_{G^*} G_{\text{new}}^* + b_E E + b_{G^*E} (G_{\text{new}}^* \times E) + \epsilon \\ \text{S1.2: } Y_1^{\text{perm}} = b_{G^*} G^* + b_E E^{\text{perm}} \\ \quad + b_{G^*E} (G^* \times E^{\text{perm}}) + \epsilon. \end{array} \right. \quad (21)$$

### 3 | SIMULATION STUDIES

#### 3.1 | Simulation models and parameter values

For indirect or direct  $G \times E$  interaction study of a single SNP (i.e., interaction scenarios 1 and 2), Table 1 provides the details of the data-generating models and parameter

values used. To evaluate the *Levene* and  $LRT_v$  tests for variance heterogeneity, besides the data-generating model as described in Section 2.3.1 and as used by Aschard et al. (2013), we also considered the model adopted by Cao et al. (2014) for a more extensive comparison. Cao et al. (2014) used Model (4) to directly simulate variance homogeneity or heterogeneity in  $Y$  stratified by  $G$ . In contrast, Aschard et al. (2013) used Model (1) to indirectly simulate variance heterogeneity that has better genetic epidemiology interpretation, because the size of  $\beta_{GE}$  corresponds to power of scale tests under alternatives. For direct testing of the interaction effect, conveniently, the data-generating model of Aschard et al. (2013) in Table 1 is conceptually the same as the simulation model I of Rao and Province (2016), except that  $SNP_{\text{repeating}}$  is a fixed  $E$  here.

To best mimic real data conditions, we also used a “double-loop” simulation design. Consider the “theoretical” null of  $Y_0 = \beta_E E + e$  for direct  $G \times E$  interaction analysis. Within each of nrep.out replicates (e.g., 100) in an outer simulation loop, we simulate  $Y_0$  based on a fixed  $E$ ,  $Y_0 = \beta_E E_{\text{fixed}} + e$ . We then simulate nrep.in replicates (e.g.,  $10^4$ ) of  $G_{\text{new}}$ , test  $b_{GE}$  based on the fitted model of  $Y_0 = b_G G_{\text{new}} + b_E E_{\text{fixed}} + b_{GE} (G_{\text{new}} \times E_{\text{fixed}}) + \epsilon$ , and estimate the T1E rate using the nrep.in replicates. This is similar to one single whole-genome  $G \times E$  interaction scan. Finally, we average the T1E values over nrep.out replicates to account for sampling variation inherent in the simulation of a  $E_{\text{fixed}}$  for one single whole-genome interaction scan. This can be done similarly for the “empirical” null.

**TABLE 2** Summary of the data-generating models for direct multivariate  $G \times E$  interaction testing, and evaluating the “theoretical” null simulation designs S0 versus the two “empirical” null simulation designs S1.1 and S1.2, as described in Section 2.3.3

|   |   |
|---|---|
| Null model for S0                         | $Y_0 = \sum_{j=1}^J \beta_{G_j} G_j + \beta_E E + e, e \sim N(0, \sigma^2)$   |
| Alternative models for S1.1 and S1.2      | $Y_1 = \sum_{j=1}^J \beta_{G_j} G_j + \beta_E E + \sum_{j=1}^J \beta_{GE_j} (G_j \times E) + e, e \sim N(0, \sigma^2)$  |
| MAF for $\vec{G}_j$ and $\vec{G}_{new_j}$ | Large: (2.15, 2.58, 2.58, 4.16, 2.57, 2.61, 4.95, 2.58, 2.57, 2.58, 3.68) $\times 10^{-1}$<br>Small: (2.15, 2.58, 2.58, 4.16, 2.57, 2.61, 4.95, 2.58, 2.57, 2.58, 3.68) $\times 10^{-2}$  |
| Parameters                                | $J = 11; \mathbb{P}(E = 1) = 0.3,$<br>$\vec{\beta}_G = (-0.218, 0, 0, -0.476, 0, 0, -0.151, -0.845, 0.0945, 0, -0.133),$<br>$\beta_E = 0.3, \vec{\beta}_{GE} = (0.1, -0.1, 0, 0, 0.1, 0.1, 0, -0.1, 0, -0.1, 0), \sigma^2 = 0.27$ |
| Sample size                               | $n = 10^3$  |
| Nominal T1E level                         | $\alpha = 0.05, 0.01, 0.001$  |
| Replications                              | nrep.in = $10^5$ ; nrep.out = 100   |

T1E rate is first estimated from nrep.in simulation replicates in an inner loop in which  $E$  is fixed (similar to one whole-genome gene-based  $G \times E$  interaction scan), then averaged over nrep.out simulation replicates in an outer loop in which  $E$  varies.

MAF: minor allele frequency.

For each combination of parameter values in Table 1 that generates  $Y_1$  under an alternative, instead of studying power (of  $Y_1 - G$ ), we focused on evaluating T1E control (of  $Y_1 - G_{new}$  or  $Y_1^{\text{perm}} - G$ ) contrasting the proposed “empirical” null simulation designs (S1.1 and S1.2) with the previously considered “theoretical” null simulation design (S0), as described in Section 2. Table 2 provides the parameter values for gene-based interaction analysis (i.e., interaction scenario 3). The number of SNPs was chosen to be  $J = 11$  as in Lin et al. (2016), among which only six interact with  $E$ . That is,  $\beta_{GE_j} \neq 0$  only for some  $j = 1, \dots, J$  as detailed in Table 2. Two sets of MAF levels were considered, with one presents gene-based interaction studies of common variants and the other of rare variants.

### 3.2 | Simulation results

For each of the three  $G \times E$  interaction scenarios, for each of the statistical tests under the study, and for each of the three T1E simulation designs, we recorded the corresponding T1E rate. We bold the ones that exceed the  $\alpha \pm 3\sqrt{\alpha \times (1 - \alpha)/\text{nrep.in}}$  range, where  $\alpha$  is the nominal T1E rate, and rep.in is the number of simulation replicates used to estimate the empirical T1E rate for each of the nrep.out replicates; each replicate in an outer loop represents a whole-genome association scan. Thus,  $\alpha \pm 3\sqrt{\alpha \times (1 - \alpha)/\text{nrep.in}}$  is a conservative interval.

#### 3.2.1 | Scenario 1: Single $G$ , and $E$ missing

Results in Table 3 show that, while location tests for phenotypical mean differences across genotype groups ( $LRT_m$  and  $Score_m$ ) are generally robust to the choice of

“theoretical” null S0 versus “empirical” null S1.1 or S1.2, it is not the case for the  $LRT_v$  test for variance heterogeneity; the empirical T1E rates of the *Levene* test were slightly deflated but not significantly. Different choices of the null simulation designs lead to different conclusions regarding the accuracy of  $LRT_v$ . For example, simulation design S0 showed that  $LRT_v$  has the correct T1E control across the parameter values considered. However, designs S1.1 and S1.2 revealed its inflated empirical T1E rate, for example, 0.07 for the nominal  $\alpha = 0.05$  level for some settings.

**TABLE 3** Simulation results of interaction scenario 1: Single  $G$ , and  $E$  missing

|          |           | $\alpha = 5 \times 10^{-2}, n = 10^3$ |       |       |              |              |              |              |
|----------|-----------|---------------------------------------|-------|-------|--------------|--------------|--------------|--------------|
|          |           | $\beta_{GE}$                          | 0.0   | 0.2   | 0.4          | 0.6          | 0.8          | 1            |
| Location | $LRT_m$   | S0                                    | 5.029 | 5.027 | 5.026        | 5.027        | 5.027        | 5.026        |
|          |           | S1.1                                  | 5.039 | 5.021 | 5.021        | 5.019        | 5.014        | 5.010        |
|          |           | S1.2                                  | 4.997 | 5.023 | 5.022        | 5.020        | 5.017        | 5.011        |
|          | $Score_m$ | S0                                    | 5.002 | 4.998 | 4.997        | 4.998        | 4.998        | 4.997        |
|          |           | S1.1                                  | 5.014 | 4.992 | 4.993        | 4.991        | 4.985        | 4.981        |
|          |           | S1.2                                  | 4.974 | 4.994 | 4.993        | 4.990        | 4.988        | 4.983        |
| Scale    | $LRT_v$   | S0                                    | 5.035 | 5.083 | 5.081        | 5.081        | 5.081        | 5.079        |
|          |           | S1.1                                  | 5.029 | 5.188 | <b>5.262</b> | <b>5.507</b> | <b>5.979</b> | <b>6.757</b> |
|          |           | S1.2                                  | 5.031 | 5.198 | <b>5.274</b> | <b>5.519</b> | <b>5.994</b> | <b>6.756</b> |
|          | $Levene$  | S0                                    | 4.956 | 4.898 | 4.898        | 4.898        | 4.898        | 4.896        |
|          |           | S1.1                                  | 4.989 | 4.901 | 4.904        | 4.905        | 4.909        | 4.907        |
|          |           | S1.2                                  | 4.906 | 4.922 | 4.912        | 4.911        | 4.915        | 4.908        |

Empirical T1E rates of  $LRT_m$  and  $Score_m$  location tests for mean difference in  $Y$  across the three  $G$  groups, and of  $LRT_v$  and *Levene* scale tests for variance difference in  $Y$ , based on the “theoretical” null design of S0 and the alternative “empirical” null designs of S1.1 and S1.2. The alternative  $Y_1$  data were generated using the *Aschard's genetic model* as described in Table 1. Empirical T1E rates outside  $\alpha \pm 3\sqrt{\alpha \times (1 - \alpha)/\text{nrep.in}}$  are bolded.



**TABLE 4** Simulation results of interaction scenario 1: Single  $G$ , and  $E$  missing; effect of the nominal  $\alpha$  level

|          | $n = 10^3$ | $\alpha$ | $5 \times 10^{-2}$ | $1 \times 10^{-2}$ | $1 \times 10^{-3}$ | $1 \times 10^{-5}$ |
|----------|------------|----------|--------------------|--------------------|--------------------|--------------------|
| Location | $LRT_m$    | S0       | 5.016              | 0.999              | 0.985              | 0.988              |
|          |            | S1.1     | 5.009              | 1.007              | 0.988              | 0.990              |
|          |            | S1.2     | 5.011              | 1.009              | 1.033              | 1.013              |
|          | $Score_m$  | S0       | 5.008              | 0.998              | 0.989              | 0.990              |
|          |            | S1.1     | 4.982              | 0.998              | 0.982              | 0.991              |
|          |            | S1.2     | 4.982              | 0.999              | 0.983              | 0.997              |
| Scale    | $LRT_v$    | S0       | 5.009              | 1.002              | 1.024              | 1.033              |
|          |            | S1.1     | <b>6.923</b>       | <b>1.636</b>       | <b>2.059</b>       | <b>11.599</b>      |
|          |            | S1.2     | <b>6.920</b>       | <b>1.639</b>       | <b>2.042</b>       | <b>11.624</b>      |
|          | $Levene$   | S0       | 4.955              | 0.961              | 0.938              | 0.971              |
|          |            | S1.1     | 4.964              | 0.978              | 0.932              | 0.952              |
|          |            | S1.2     | 4.962              | 0.970              | 0.953              | 0.958              |

Empirical T1E rates of  $LRT_m$  and  $Score_m$  location tests for mean difference in  $Y$  across the three  $G$  groups, and of  $LRT_v$  and  $Levene$  scale tests for variance difference in  $Y$ , based on the “theoretical” null design of S0 and the alternative “empirical” null designs of S1.1 and S1.2. The alternative  $Y_i$  data were generated using the *Aschard’s genetic model* as described in Table 1, focusing on the extreme case of large interaction effect,  $\beta_{GE} = 1$ . Empirical T1E rates outside  $\alpha \pm 3\sqrt{\alpha \times (1 - \alpha)/nrep.in}$  are boded.

While the increased T1E rates under the S1.1 and S1.2 “empirical” null designs appeared to be mild and occurred only in extreme models (i.e., large  $G \times E$  interaction effect), results in Table 4 demonstrate that the T1E issue of  $LRT_v$ , revealed under the “empirical” null simulation designs of S1.1 and S1.2, can be more severe at the tail. For example, for the nominal  $\alpha = 1 \times 10^{-5}$  level, the empirical T1E rate of  $LRT_v$  can be as high as  $11.5 \times 10^{-5}$ . Because the genome-wide significance level for GWAS is  $\alpha = 5 \times 10^{-8}$  (Dudbridge & Gusnanto, 2008), an inflation of false positive findings can be of a real problem in practice. Further, results in Table 5 confirm that increasing sample size  $n$  (from  $10^3$  to  $10^4$ ) does not mitigate the discrepancy in T1E conclusion drawn from the “theoretical” versus “empirical” null.

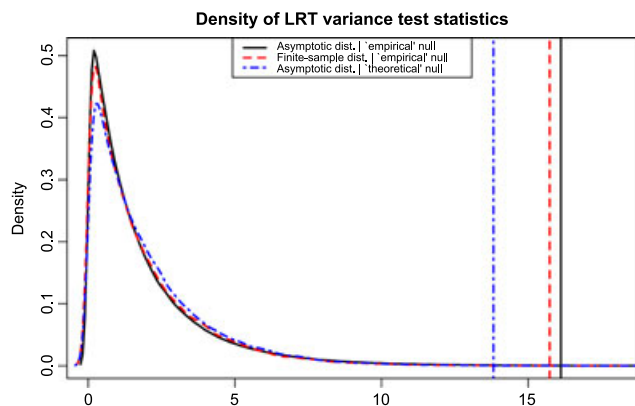
The root cause of this discrepancy is unsuspected model misspecification. Under the S1.1 and S1.2 designs,  $Y_i$  marginally is not normally distributed, even if it was generated (conditional on the true  $G$ ) using a normally distributed error  $e$  term. Using  $T = LRT_v$  as an example, Cao et al. (2014) has shown that  $LRT_v$  is accurate based on  $f_0(T)$ , which is  $\chi_{(2)}^2$  derived for  $Y_0 - G_{new}$  under the “theoretical” null S0 condition and assuming  $e \sim N(0, \sigma^2)$ . However, when  $LRT_v$  is applied to  $Y_1 - G_{new}$ , even if  $e \sim N(0, \sigma^2)$  for generating  $Y_i$  conditional on the true  $G$ , the correct asymptotic distribution of  $LRT_v$  under the “empirical” null is a weighted sum of independent  $\chi_{(1)}^2$  (Supporting Information Materials and Theorem 3.4.1(1) of Yanagihara, Tonda, & Matsumoto, 2005; Tonda, & Wakaki, 2003; Gomes-Sanchez-Manzano et al, 2006; Eicker, 1969). Thus, assessing  $LRT_v$  using  $\chi_{(2)}^2$  can have T1E issues if the data were generated under the “empirical” null S1.1 condition as in real data, similarly for S1.2.

Figure 1 compares the asymptotic distribution (black solid curve) with the finite-sample distribution (red dashed curve) of  $LRT_v$  under the “empirical” null, as well as with the asymptotic distribution ( $\chi_{(2)}^2$ , blue dot-dashed curve) of  $LRT_v$  under the “theoretical” null. While the finite-sample distribution approximates well the asymptotic distribution derived under the “empirical” null, it is clear that the distributions of  $LRT_v$  differ between the “empirical” and “theoretical” null; the difference is more visible on the

**TABLE 5** Simulation results of interaction scenario 1: Single  $G$ , and  $E$  missing; effect of sample size  $n$ 

|           |      | $\alpha = 5 \times 10^{-2}$ |              |
|-----------|------|-----------------------------|--------------|
|           |      | $n = 10^3$                  | $n = 10^4$   |
| Location  |      |                             |              |
| $LRT_m$   | S0   | 5.011                       | 5.012        |
|           | S1.1 | 4.992                       | 5.003        |
|           | S1.2 | 4.934                       | 4.989        |
| $Score_m$ | S0   | 5.011                       | 5.012        |
|           | S1.1 | 4.993                       | 5.003        |
|           | S1.2 | 4.934                       | 4.989        |
| Scale     |      |                             |              |
| $LRT_v$   | S0   | 5.103                       | 5.165        |
|           | S1.1 | <b>7.034</b>                | <b>7.125</b> |
|           | S1.2 | <b>7.007</b>                | <b>7.020</b> |
| $Levene$  | S0   | 4.924                       | 4.945        |
|           | S1.1 | 4.905                       | 4.965        |
|           | S1.2 | 4.874                       | 4.825        |

Empirical T1E rates of  $LRT_m$  and  $Score_m$  location tests for mean difference in  $Y$  across the three  $G$  groups, and of  $LRT_v$  and  $Levene$  scale tests for variance difference in  $Y$ , based on the “theoretical” null design of S0 and the alternative “empirical” null designs of S1.1 and S1.2. The alternative  $Y_i$  data were generated using the *Cao’s genetic model* as described in Table 1, and using two difference sample sizes of  $n = 10^3$  and  $10^4$ . Empirical T1E rates outside  $\alpha \pm 3\sqrt{\alpha \times (1 - \alpha)/nrep.in}$  are boded.



**FIGURE 1** Comparison of the asymptotic distribution (black solid) and finite-sample distribution (red dashed) of  $LRT_v$  under the “empirical” null, with the asymptotic distribution ( $\chi^2_2$ , blue dot-dashed) of  $LRT_v$  under the “theoretical” null. Vertical lines correspond the 99.9% quantile cutoffs for  $\alpha = 0.001$

scale of critical value for statistical significance (the vertical lines). The true significance threshold for  $LRT_v$  under the “empirical” null is further away to the tail, as compared with the threshold of  $\chi^2_{(2)}$  under the “theoretical” null. Thus, applying  $LRT_v$  to real GWAS or NGS while using the significance threshold of  $\chi^2_{(2)}$  can lead to inflated T1E rate. Deriving the correct distribution corresponding to the “empirical” null, unfortunately, requires the knowledge of the alternative model which is unknown in practice. Permutation-based method can provide reasonable estimates which we discuss later.

In practice, it is routine (and recommended) to display and examine the empirical distribution of a trait under the study. However, Supporting Information Figure S1 shows that even under the most extreme setting where  $\beta_{GE} = 1$ , the marginal histogram of  $Y_1$  appears to be approximately normal visually, unless a formal diagnostic test for normality was conducted. The slightly right-skewed empirical distribution of  $Y_1$  is the result of mixing six conditional distributions of  $Y_1$ , each perfectly normally distributed conditional on the causal  $G$  and  $E$ . This is the key difference between the “theoretical” and “empirical” null simulation designs, regardless of the sample size. For a less extreme case where  $\beta_{GE} = 0.2$ , although both the histogram and Q-Q plot (Supporting Information Figure S2) suggest that normal distribution is a good fit (passing the Shapiro–Wilk normality test), the T1E discrepancy between the “theoretical” and “empirical” null remains, although less severe, as shown in Table 3 (column  $\beta_{GE} = 0.2$ ) and Supporting Information Figure S3 (middle row). Tables 3, 4, and 5 also provide T1E results for testing phenotypic mean (as opposed to variance) differences across the genotype groups. Although location testing for main effect is generally quite robust to the assumption of normality, problem can arise when testing for interaction

effects, beyond any apparent model misspecifications (Rao & Province, 2016).

### 3.2.2 | Scenario 2: Single $G$ , and $E$ known

In direct testing for the  $G \times E$  interaction effect ( $SNP_{\text{repeating}} \times SNP_{\text{nonrepeating}}$  to be more precise), Rao and Province (2016) used the classical “theoretical” null simulation design S0, with or without the main  $G$  effect. Regardless, figures 1b,c of Rao and Province (2016) showed that the variation in the resulting  $\lambda_{GC}$  was substantially bigger when testing  $b_{GE}$  than testing  $b_G$ . Their figures 1d,e also demonstrated that the variation diminishes as sample size increases. However, we note that this observation was made before averaging across the 414 simulated interaction scans/datasets; each scan contained 20,000 SNPs from which a  $\lambda_{GC}$  value was estimated.

The results of Rao and Province (2016) are consistent with ours shown in Supporting Information Figure S6. Supporting Information Figure S6 shows that scan-specific estimated T1E rates are indeed variable (due to dependency between the tests) and become less so as the sample size increases; 100  $G \times E$  whole-genome interaction scans with  $10^5$  SNPs in each scan and  $E$  being fixed within a scan. However, it is important to note that the average T1E rate across nrep.out simulated scans reflects better the long-run behavior of a method. Indeed, results in Table 6 show that the T1E rate of testing  $b_{GE}$ , estimated from  $10^5 \times 100$  (nrep.in  $\times$  nrep.out) simulated replicates, is well controlled under the conventional “theoretical” null simulation design of S0. However, this is not the case under the “empirical” null simulation designs of S1.1 or S1.2. Similar to the  $LRT_v$  scale test for variance heterogeneity, the discrepancy in T1E control, between the two types of null simulation designs, becomes more prominent at the tail and persists as sample increases (Table 6).

### 3.2.3 | Scenario 3: Multiple $G$ s, and $E$ known

Table 7 contains T1E results for gene-based interaction studies of both common and rare variants. When the MAFs are between 0.2 and 0.5 (common), we observed that the T1E rates of all tests considered are significantly different between the “theoretical” null and “empirical” null. This result is consistent with that of the interaction scenarios 1 and 2 above. Our simulation results also confirmed that the  $Burden_{GE}$  interaction test has T1E issue even under the “theoretical” null, as previously noted in Lin et al. (2016). When the MAFs are reduced 10-fold to be between 0.02 and 0.05 (rare), although  $SKAT_{GE}$  appears to be robust, the  $Burden_{GE}$  and  $F_{GE}$  tests

**TABLE 6** Simulation results of interaction scenario 2: Single  $G$ , and  $E$  known

|              |            | $n = 10^3$         |                    |                    | $n = 10^4$         |                    |                    |
|--------------|------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|              | $\alpha =$ | $5 \times 10^{-2}$ | $1 \times 10^{-2}$ | $1 \times 10^{-3}$ | $5 \times 10^{-2}$ | $1 \times 10^{-2}$ | $1 \times 10^{-3}$ |
| $LRT_{GE}$   | S0         | 5.034              | 1.017              | 1.029              | 5.033              | 1.007              | 0.979              |
|              | S1.1       | <b>7.091</b>       | <b>1.763</b>       | <b>2.422</b>       | <b>6.886</b>       | <b>1.709</b>       | <b>2.410</b>       |
|              | S1.2       | <b>7.100</b>       | <b>1.771</b>       | <b>2.389</b>       | <b>6.972</b>       | <b>1.707</b>       | <b>2.339</b>       |
| $Score_{GE}$ | S0         | 4.982              | 1.003              | 1.002              | 5.021              | 1.004              | 0.972              |
|              | S1.1       | <b>7.028</b>       | <b>1.738</b>       | <b>2.363</b>       | <b>6.874</b>       | <b>1.705</b>       | <b>2.400</b>       |
|              | S1.2       | <b>7.040</b>       | <b>1.747</b>       | <b>2.339</b>       | <b>6.960</b>       | <b>1.703</b>       | <b>2.326</b>       |

Empirical T1E rates of the  $LRT_{GE}$  and  $Score_{GE}$  tests, based on the ‘theoretical’ null design of S0 and the alternative ‘empirical’ null designs of S1.1 and S1.2. The alternative  $Y_1$  data were generated using the *Aschard’s genetic model* as described in Tables 2 when  $\beta_{GE} = 1$ , but  $E$  was assumed to be known in this case and direction interaction testing was possible. Empirical T1E rates outside  $\alpha \pm 3\sqrt{\alpha \times (1 - \alpha)/nrep.in}$  are bolded.

continue to display significant differences between the “theoretical” and “empirical” null simulation designs for evaluating their T1E control.

## 4 | DISCUSSION

In this study, we highlight the importance of distinguishing the “theoretical” and “empirical” null distributions, first noted by Efron (2004), in a different application context. Starting with scale tests for variance heterogeneity and through simulation studies, we showed that conclusions of T1E control of a statistical test could differ depending on the choice of the null simulation designs. For example, the  $LRT_v$  variance test appears to be accurate under the “theoretical” null design, but it can have inflated T1E under the “empirical” null design that better mimics real data conditions (Tables 3, 4, and 5, and Supporting Information Figure S3).

Cao et al. (2014) has pointed out the sensitivity and limitation of  $LRT_v$  when the error term  $e$  is not normally distributed. However, they implicitly assumed that the test

would work well as long as  $e \sim N(0, \sigma^2)$ , regardless of “theoretical” null or “empirical” null. In our simulation study, although all the  $e$ ’s for generating the phenotype data were normal, the increased T1E under the “empirical” null are, fundamentally, attributed to the sensitivity of  $LRT_v$  to subtle model misspecifications. This is because the marginal distribution of the empirical outcome data are in fact not normal (Supporting Information Figures S1 and S2). Thus, tests shown to be extremely sensitive to model assumptions are particularly vulnerable when applied to real data. Because empirical data are better represented by the “empirical” null than the “theoretical” null, evaluating T1E control using the “empirical” null design can expose the true behavior of a test when applied to real whole-genome association scans.

Conversely, power calculation for  $LRT_v$  using the significance threshold of the “theoretical” null distribution can be too optimistic. Indeed, this study was motivated by the observation made in Supporting Information Table S7 of Soave et al. (2015) that, “*further analysis using permutation estimation of p-values showed that power of the LRT under asymptotic*

**TABLE 7** Simulation results of interaction scenario 3: Multiple  $G$ s, and  $E$  known

|               |            | Small MAF (rare)   |                    |                    | large MAF (common) |                    |                    |
|---------------|------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|               | $\alpha =$ | $5 \times 10^{-2}$ | $1 \times 10^{-2}$ | $1 \times 10^{-3}$ | $5 \times 10^{-2}$ | $1 \times 10^{-2}$ | $1 \times 10^{-3}$ |
| $F_{GE}$      | S0         | 4.991              | 0.982              | 0.985              | 4.983              | 1.001              | 0.992              |
|               | S1.1       | <b>5.448</b>       | <b>1.130</b>       | <b>2.363</b>       | <b>6.905</b>       | <b>1.591</b>       | <b>2.043</b>       |
|               | S1.2       | <b>5.448</b>       | <b>1.130</b>       | <b>1.238</b>       | <b>6.818</b>       | <b>1.596</b>       | <b>1.945</b>       |
| $SKAT_{GE}$   | S0         | 4.904              | 0.949              | 0.895              | 4.934              | 0.976              | 0.917              |
|               | S1.1       | 5.038              | 1.074              | 1.059              | <b>6.460</b>       | <b>1.411</b>       | <b>1.640</b>       |
|               | S1.2       | 5.034              | 1.065              | 1.110              | <b>6.378</b>       | <b>1.394</b>       | <b>1.651</b>       |
| $Burden_{GE}$ | S0         | <b>6.169</b>       | <b>3.751</b>       | <b>15.049</b>      | <b>13.709</b>      | <b>4.280</b>       | <b>7.440</b>       |
|               | S1.1       | <b>8.432</b>       | <b>2.213</b>       | <b>3.144</b>       | <b>5.666</b>       | <b>1.232</b>       | <b>1.410</b>       |
|               | S1.2       | <b>8.408</b>       | <b>2.208</b>       | <b>3.148</b>       | <b>5.712</b>       | <b>1.237</b>       | <b>1.420</b>       |

Empirical T1E rates of the  $F_{GE}$ ,  $SKAT_{GE}$ , and  $Burden_{GE}$  tests of jointly testing for multiple interaction effects, based on the “theoretical” null design of S0 and the alternative “empirical” null designs of S1.1 and S1.2. Models and parameters values are given in Table 2. Empirical T1E rates outside  $\alpha \pm 3\sqrt{\alpha \times (1 - \alpha)/nrep.in}$  are bolded.

MAF: minor allele frequency.

*analysis was greatly inflated.*” The work here provides analytical insights on why this is the case. The asymptotic power was obtained using the  $\chi^2_{(2)}$  distribution derived under the “theoretical” null while controlling T1E at  $\alpha$ , as in Cao et al. (2014). The permutation-based power was obtained by using the empirical  $1 - \alpha$  quantile cutoff of the  $LRT_v$  statistic applied to permuted phenotype data. This is equivalent to controlling T1E at  $\alpha$  under the “empirical” null of S1.2. Under the “empirical” null, however, we showed that  $LRT_v$  follows a different distribution and the corresponding significance quantile cutoff is further to the tail, as compared with the  $f_0(T) = \chi^2_{(2)}$  distribution under the “theoretical” null (Figure 1). This leads to (correct) smaller permutation-based power. The (incorrect) higher asymptotic-based power is a result of increased T1E rate, when the data were generated under the “empirical” null condition but the test statistic was evaluated using  $\chi^2_{(2)}$  derived under the “theoretical” null condition. Permutation-based S1.2 null design can estimate the true distribution of a test statistic  $T$  when applied to a real data set, and identify the correct significance threshold for  $T$  under the “empirical” null. But, permutation must be carried out carefully in practice, for example, in the presence of sample correlation (Abney, 2015).

In practice, investigators often rely on visual inspection of histograms of outcome data as illustrated in Supporting Information Figures S1 and S2. We have noted that the departure from normality does not have to be severe to have an effect on tests such as  $LRT_v$ , as observed in a  $LRT_v$  scan of lung function in cystic fibrosis subjects by Soave et al. (2015). In that case, the lung phenotype is called SaKnorm, defined as the forced expiratory volume in one second, adjusted for sex, age, height, and mortality, and normalized. In the sample analyzed, the distribution of the phenotype indeed appeared to be normal, but the application of  $LRT_v$  to permuted SaKnorm showed inflated T1E even for common SNPs: Supporting Information Figure S2.H of Soave et al. (2015) for 454,764 SNPs with  $MAF \geq 0.1$ , Supporting Information Figure S2.I for 111,120 SNPs with  $MAF < 0.1$ , and Supporting Information Figure S2.G for all 565,884 GWAS SNPs.

Furthermore, for data appear to deviate from normal such as those displayed in Supporting Information Figure S1, even if investigators chose to perform some standard normal transformations, the T1E issue can persist. For example, let us consider the phenotype data simulated based on Aschard’s genetic model, as described in Table 1 where  $\beta_{GE} = 1$  (Supporting Information Figure S1). After square-root or log transformations (Goh & Yap, 2009), although the empirical marginal distribution of the phenotype data improved as expected (Supporting Infor-

mation Figure S4), the severity of T1E inflation of  $LRT_v$  in fact worsened under the “empirical” null simulation designs S1.1 and S1.2 (Supporting Information Figure S5).

Voorman et al. (2011) showed that spurious false positives can occur in genome-wide scans for  $G \times E$  interactions, particularly in the presence of model misspecification. Rao and Province (2016) also presented inflated or deflated genomic inflation factor  $\lambda_{GC}$  in a  $G \times G$  interaction scan when one SNP is anchored (i.e.,  $SNP_{\text{repeating}} \times SNP_{\text{nonrepeating}}$ ), using the conventional “theoretical” null simulation design without any apparent model misspecifications. Based on our  $G \times E$  simulation studies where  $E$  was fixed within each scan, we note that the large variation in  $\lambda_{GC}$  estimate demonstrated by Rao and Province (2016) corresponds to the sampling variation inherent in estimating T1E rate from  $n_{\text{rep.in}}$  replicates (or SNPs) within each of the  $n_{\text{rep.out}}$  replicates (or scans). This, however, does not translate to T1E issue based on the classical frequentist interpretation. Results in Table 6 show that, under the “theoretical” null of S0, there is no T1E issue in  $G \times E$  interaction studies even if  $E$  was fixed within a genome-wide scan. But, similar to scale test of variance, T1E results differ using the “empirical” S1.1 or S1.2 null simulation designs. Additional theoretical insights are provided in Section 3 of the Supporting Information Materials.

Departure from normality is generally weakened under multivariate assumptions. However, the topic addressed here remains relevant. To demonstrate this, in addition to studying interaction between  $E$  and  $G$  of a single SNP, we also examined testing for interaction effects between  $E$  and multiple  $G$ s as in gene-based interaction studies. We reached the same conclusion that T1E conclusions could differ between the “theoretical” and “empirical” null simulation designs.

To conclude, although we only presented three examples (i.e., scale tests for variance heterogeneity, and direct tests for one or multiple interaction effects jointly), the findings here have important implications for future evaluation of T1E control and interpretation. The newer tests being developed often go beyond the first moment such as the scale tests studied here, and they are increasingly complex and possibly more sensitive to subtle model misspecifications. The conventional “theoretical” null simulation design (S0) is unrealistic and can lead to misleading conclusion regarding T1E control, which in turn affects power study. The alternative “empirical” null simulation designs (S1.1 and S1.2) can reveal the true behavior of a test when applied to real data.

## ACKNOWLEDGMENTS

The authors have no conflict of interest to declare. The authors thank Dr. David Soave and Dr. Jerry Lawless for



helpful discussions, and the two external reviewers for valuable suggestions that substantially improved the presentation of the paper. This study was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-250053) and the Canadian Institutes of Health Research (CIHR; MOP-310732-G-CEAA-117978) and to L. S.

## ORCID

Lei Sun  <http://orcid.org/0000-0002-5640-937X>

## REFERENCES

- Abney, M. (2015). Permutation testing in the presence of polygenic variation. *Genetic Epidemiology*, 39(4), 249–258.
- Aschard, H., Zaitlen, N., Tamimi, R. M., Lindström, S., & Kraft, P. (2013). A nonparametric test to detect quantitative trait loci where the pheno-typic distribution differs by genotypes. *Genetic Epidemiology*, 37(4), 323–333.
- Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346), 364–367.
- Cao, Y., Wei, P., Bailey, M., Kauwe, J. S. K., & Maxwell, T. J. (2014). A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology*, 38(1), 51–59.
- Dadaev, T., Saunders, E. J., Newcombe, P. J., Anokian, E., Leongamornlert, D. A., Brook, M. N., & Kote-Jarai, Z. (2018). Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nature Communications*, 9(1), 2256.
- Derkach, A., Lawless, J. F., & Sun, L. (2014). Pooled association tests for rare genetic variants: A review and some new results. *Statistical Science*, 29(2), 302–321.
- Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4), 997–1004.
- Dudbridge, F., & Fletcher, O. (2014). Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, 95(3), 301–307.
- Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3), 227–234.
- Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465), 96–104.
- Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, 1(1), 59–82.
- Falconer, D. S. (1960). *Introduction to quantitative genetics*. Edinburgh, London: Oliver and Boyd.
- Goh, L., & Yap, V. B. (2009). Effects of normalization on quantitative traits in association test. *BMC Bioinformatics*, 10(1), 415.
- Gómez-Sánchez-Manzano, E., Gómez-Villegas, M., & Marín, J. (2006). Sequences of elliptical distributions and mixtures of normal distributions. *Journal of Multivariate Analysis*, 97(2), 295–310.
- Hill, W., Marioni, R. E., Maghziian, O., Ritchie, S. J., Hagenaars, S. P., McIntosh, A. M., & Deary, J. (2018). A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Molecular Psychiatry*, 23(1).
- Khan, A., & Rayner, G. D. (2003). Robustness to nonnormality of common tests for the many-sample location problem. *Journal of Applied Mathematics and Decision Sciences*, 7(4), 187–206.
- Levene, H., et al. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics*, 1, 278–292.
- Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3), 311–321.
- Lin, X., Lee, S., Christiani, D. C., & Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4), 667–681.
- Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., & Li, Z. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, 72(1), 156–164.
- Mackay, T. F. (2009). Q&A: Genetic analysis of quantitative traits. *Journal of Biology*, 8(3), 23.
- Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2), e1000384.
- 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.
- Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., & Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6), 832–838.
- Rao, T. J., & Province, M. A. (2016). A framework for interpreting type I error rates from a product-term model of interaction applied to quantitative traits. *Genetic Epidemiology*, 40(2), 144–153.
- Soave, D., Corvol, H., Panjwani, N., Gong, J., Li, W., Boëlle, P. Y., & Sun, L. (2015). A joint location-scale test improves power to detect associated SNPs, gene sets, and pathways. *The American Journal of Human Genetics*, 97(1), 125–138.
- Soave, D., & Sun, L. (2017). A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics*, 73(3), 960–971.
- Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C., & Aulchenko, Y. S. (2010). Variance heterogeneity analysis for detection of potentially interacting genetic loci: Method and its limitations. *BMC Genetics*, 11(1), 92.
- Sun, X., Elston, R., Morris, N., & Zhu, X. (2013). What is the significance of difference in phenotypic variability across SNP genotypes? *The American Journal of Human Genetics*, 93(2), 390–397.
- Tonda, T., & Wakaki, H. (2003). Asymptotic expansion of the null distribution of the likelihood ratio statistic for testing the equality of variances in a nonnormal one-way {ANOVA} model. *Hiroshima Mathematical Journal*, 33(1), 113–126.



- Voorman, A., Lumley, T., McKnight, B., & Rice, K. (2011). Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One*, 6(5), e19416.
- Wood, A. R., Tuke, M. A., Nalls, M. A., Hernandez, D. G., Bandinelli, S., Singleton, A. B., & Weedon, M. N. (2014). Another explanation for apparent epistasis. *Nature*, 514(7520), E3.
- Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, N. (2011). Rare-variant association testing for sequencing data with the sequencekernel association test. *The American Journal of Human Genetics*, 89(1), 82–93.
- Yanagihara, H., Tonda, T., & Matsumoto, C. (2005). The effects of nonnormality on asymptoticdistributions of some likelihood ratio criteria for testing covariance structures under normal-assumption. *Journal of Multivariate Analysis*, 96(2), 237–264.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Zhang T, Sun L. Beyond the traditional simulation design for evaluating type 1 error control: From the “theoretical” null to “empirical” null. *Genet. Epidemiol.* 2019;43: 166–179. <https://doi.org/10.1002/gepi.22172>