frontiers | Frontiers in Genetics

BRIEF RESEARCH REPORT
published: 12 April 2022
doi: 10.3389/fgene.2022.856872

Linear Mixed-Effect Models Through the Lens of Hardy–Weinberg Disequilibrium

Lin Zhang [1] *and Lei Sun* [1,2]*

[1]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada, [2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

For genetic association studies with related individuals, the linear mixed-effect model is the most commonly used method. In this report, we show that contrary to the popular belief, this standard method can be sensitive to departure from Hardy–Weinberg equilibrium (i.e., Hardy–Weinberg disequilibrium) at the causal SNPs in two ways. First, when the trait heritability is treated as a nuisance parameter, although the association test has correct type I error control, the resulting heritability estimate can be biased, often upward, in the presence of Hardy–Weinberg disequilibrium. Second, if the true heritability is used in the linear mixed-effect model, then the corresponding association test can be biased in the presence of Hardy–Weinberg disequilibrium. We provide some analytical insights along with supporting empirical results from simulation and application studies.

Keywords: genome-wide association study, dependent sample, robust association analysis, heritability estimate, Hardy–Weinberg equilibrium

OPEN ACCESS

Edited by:
Lide Han,
Vanderbilt University Medical Center, United States

Reviewed by:
Guo-Bo Chen,
Zhejiang Provincial People's Hospital, China
Chengsong Zhu,
University of Texas Southwestern Medical Center, United States

***Correspondence:**
Lei Sun
sun@utstat.toronto.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *17 January 2022*
Accepted: *09 March 2022*
Published: *12 April 2022*

Citation:
Zhang L and Sun L (2022) Linear Mixed-Effect Models Through the Lens of Hardy–Weinberg Disequilibrium. Front. Genet. 13:856872. doi: 10.3389/fgene.2022.856872

1 INTRODUCTION

Genetic association tests are often derived from a regression model, regressing the phenotypic data of a complex trait (Y) on the genotypic data of a single-nucleotide polymorphism (SNP; G), as well as on the covariate data of important environmental factors (Z). When individuals in a sample are genetically related with each other, the linear mixed-effect model (LMM) is the most commonly used method for genome-wide association studies (GWAS) (Eu-Ahsunthornwattana et al., 2014). The variance–covariance matrix of the regression model is partitioned into a weighted sum of the genetic correlation matrix and the correlation matrix due to shared environmental effects. The genetic correlation matrix is typically represented by the kinship coefficient matrix, which is either inferred from the (correctly) known pedigree structure or estimated based on the available genome-wide genetic data (Yang et al., 2011; Dimitromanolakis et al., 2019). The weight for the genetic correlation matrix is referred to as the heritability of the trait (Visscher et al., 2006; Visscher et al., 2008); Falconer (1985) gave a theoretical modeling of the variance partition, which sets the foundation for heritability.

It is commonly assumed that these regression-based association tests are robust to departure from Hardy–Weinberg equilibrium (HWE) (Sasieni, 1997). HWE states that the two alleles in a genotype are independent draws from the same Bernoulli distribution, or, equivalently, genotype frequencies depend solely on the allele frequencies (Hardy et al., 1908; Weinberg, 1908). For a biallelic SNP with two possible alleles A and a, let p and $1 - p$ be the population allele frequencies, respectively. Under HWE, $p_{aa} = (1 - p)^2$, $p_{Aa} = 2p(1 - p)$, and $p_{AA} = p^2$, where p_{aa}, p_{Aa}, and p_{AA} are the population genotype frequencies of genotypes aa, Aa, and AA, respectively. To quantify the departure from HWE or the amount of Hardy–Weinberg disequilibrium (HWD),

$$\delta = p_{AA} - p^2 \quad (1)$$

is a widely used measure (Weir, 1996), and $\delta = 0$ indicates HWE holds. We note that a) HWE is also known as Hardy–Weinberg proportion and b) δ is also known as $p(1-p)F$, where F is the inbreeding coefficient (Powell et al., 2010). Equivalently, instead of quantifying the genotype frequencies as $p_{aa} = (1-p)^2 + \delta$, $p_{Aa} = 2p(1-p) - 2\delta$, and $p_{AA} = p^2 + \delta$ based on δ (Weir, 1996), we can define them based on F as $p_{aa} = (1-p)^2 + p(1-p)F$, $p_{Aa} = 2p(1-p)(1-F)$, and $p_{AA} = p^2 + p(1-p)F$ (Powell et al., 2010). As the classical Pearson χ^2 HWE testing is based on comparing the observed genotype counts with the expected under HWE (Zhang and Sun, 2021); we thus chose δ for this work to be consistent with the GWAS literature.

A truly associated or causal SNP can be out of HWE (Wittke-Thompson et al., 2005; Ryckman and Williams, 2008; Turner et al., 2011), which is often overlooked but an important consideration when studying a method's robustness to HWD. Note that the HWD attributed to true association is typically not as extreme as the HWD caused by genotyping errors (Zhang and Sun, 2020). Thus, true HWD can remain in a "cleaned" dataset after applying the standard HWD-based quality control screening using a stringent p-value threshold [e.g., 10^{-12} for an application of the UK Biobank data by Bycroft et al. (2018)]. With a sample of independent individuals, both theoretical and empirical results support that genotype-based association tests are robust to HWD (Sasieni, 1997; Schaid and Jacobsen, 1999; Zhang, 2021). However, in the presence of sample dependency, little has been discussed.

In this report, we first provide some analytical insights on why the standard LMM can be sensitive to HWD in pedigree data in contrast to when analyzing a sample of unrelated individuals. We then demonstrate with a simple sib-pair design that 1) when the heritability is estimated from the data as in practice, although the empirical type I error rate of the LMM is well controlled, the estimated heritability is biased, often upward biased; 2) when the true heritability is known and used, the empirical type I error rate of the LMM is then inflated when $\delta > 0$, and deflated if $\delta < 0$. The result of 2) is novel, but it is mostly of an academic interest as the true heritability of a trait is often unknown in practice. On the other hand, the result of 1) has important practical implications because if the estimate of a trait heritability is larger than the true value, then it helps explain some of the "missing heritability" (Manolio et al., 2009); the insightful work of Chen (2014) "discuss [es] the circumstances in which the HE [Haseman-Elston] regression and the mixed linear model are equivalent."

2 METHODS

2.1 Traditional $Y \sim G$ Model With Independent Samples, T_{Indep}, Is Robust to HWD

Let Y be a (continuous) trait of interest, and $G = 0$, 1, and 2, respectively, for the genotypes *aa*, *Aa*, and *AA* of a SNP. Additionally, for notation simplicity but without loss of generality, we assume that there is only one additional covariate, denoted by Z. With a sample of n unrelated individuals, the traditional genotype-based association analysis assumes that

$$y = \alpha^* 1 + \beta^* \mathrm{g} + \gamma^* \mathrm{z} + \epsilon^*, \quad \epsilon^* \sim N(0, \sigma^{*2} I), \quad (2)$$

where $\mathrm{y} = (y_1, y_2, \ldots, y_n)$ is a $n \times 1$ vector for the phenotypic values, 1 is a $n \times 1$ vector of 1's, $\mathrm{g} = (g_1, g_2, \ldots, g_n)$ is a $n \times 1$ vector for the genotypes of the SNP, $\mathrm{z} = (z_1, z_2, \ldots, z_n)$ is a $n \times 1$ vector for the covariate values, ϵ^* is the error term with variance σ^{*2}, and I is the identity matrix.

Score-based tests are often used for genetic association analyses (Derkach et al., 2015). In this case, the score statistic of testing H_0: $\beta^* = 0$ can be easily derived as

$$T_{\text{indep}} = n \cdot \frac{\left\{(g-\bar{g}1)^T(\mathrm{y}-\bar{y}1) - \frac{(g-\bar{g}1)^T(\mathrm{z}-\bar{z}1)(\mathrm{y}-\bar{y}1)^T(\mathrm{z}-\bar{z}1)}{(\mathrm{z}-\bar{z}1)^T(\mathrm{z}-\bar{z}1)}\right\}^2}{\left[(g-\bar{g}1)^T(g-\bar{g}1) - \frac{\left\{(g-\bar{g}1)^T(\mathrm{z}-\bar{z}1)\right\}^2}{(\mathrm{z}-\bar{z}1)^T(\mathrm{z}-\bar{z}1)}\right]\left[(\mathrm{y}-\bar{y}1)^T(\mathrm{y}-\bar{y}1) - \frac{\left\{(\mathrm{y}-\bar{y}1)^T(\mathrm{z}-\bar{z}1)\right\}^2}{(\mathrm{z}-\bar{z}1)^T(\mathrm{z}-\bar{z}1)}\right]}. \quad (3)$$

To observe T_{indep}'s connection with Hardy–Weinberg disequilibrium, it is instructive to employ some algebraic tricks and show that

$$\frac{1}{n}(g-\bar{g}1)^T(g-\bar{g}1) = \widehat{\mathrm{var}}(G) = 2\left(\hat{p}(1-\hat{p}) + \left(\hat{p}_{AA} - \hat{p}^2\right)\right) = 2\left(\hat{p}(1-\hat{p}) + \hat{\delta}\right).$$

Because $\hat{\delta} = \hat{p}_{AA} - \hat{p}^2$ measures the amount of HWD present in the data (Weir, 1996), T_{indep} inherently adjusts for departure from HWE through $\widehat{\mathrm{var}}(G) = 2(\hat{p}(1-\hat{p}) + \hat{\delta})$. As a result, the traditional genotype-based association test is robust to HWD in independent samples.

When Y is binary, the classic logistic regression is commonly used. However, Chen (1983) showed that under some regularity conditions, the score test statistics have an identical form for the exponential family in independent samples, which was recently validated by Zhang and Sun (2021) for genetic association studies. Additionally, Derkach et al. (2015) showed that for Y-dependent sampling, "the score statistics are identical for conditional and full likelihood approaches, and are of the same form as those for ordinary random sampling." Thus, in terms of association testing (not genetic effect estimation), we can conclude that genotype-based association studies of binary traits in independent samples are also robust to HWD.

2.2 Linear Mixed-Effect Model With Dependent Samples, T_{LMM}, Can Be Sensitive to HWD

Although a pedigree-based study design is rare for genome-wide association studies, individuals can be (cryptically) related with each other even in population-based GWAS (Sun et al., 2017). Omitting related individuals simplifies the association analysis but reduces the sample size and thus power. Instead, Σ_Φ, the kinship coefficient matrix, can be estimated using the available genome-wide data to capture the sample relatedness between the n individuals (Visscher et al., 2006; Yang et al., 2011). The

association analysis using the full sample can be conducted using the linear mixed-effect model.

$$\begin{aligned} &y = \alpha^* 1 + \beta^* g + \gamma^* z + \epsilon^*, \\ &\text{where } \epsilon^* \sim N\left(0, \sigma_y^2 \Sigma_y\right) \text{ and } \Sigma_y = h^2 \Sigma_\Phi + (1 - h^2) I. \end{aligned} \tag{4}$$

Compared with the linear model used for independent samples, $\text{var}(\epsilon^*) = \sigma^{*2} I$ in **Eq. 2** is replaced by $\sigma_y^2 \Sigma_y$ to reflect the sample dependence. The matrix Σ_y is a weighted average of two components, where Σ_Φ reflects the sample relatedness; naturally, the model is reduced to the linear model of **Eq. 2** for independent samples when $\Sigma_\Phi = I$. The weight h^2 is interpreted as the heritability of the trait (Visscher et al., 2008), $h^2\sigma_y^2$ as the phenotypic variation due to (additive) genetic variation, and $(1 - h^2)\sigma_y^2 = \sigma_e^2$ as the phenotypic variation due to environmental variation. The matrix Σ_Φ is the kinship matrix, where $\Sigma_\Phi(i, j) = 2\phi_{i,j}$ and $\phi_{i,j}$ is the kinship coefficient between the ith and jth samples.

By convention, h^2 is defined as

$$h^2 = \frac{\sum \beta_k^2 \text{var}(G_k)}{\text{var}(Y)} = \frac{\sum \beta_k^2 2p_k(1 - p_k)}{\sum \beta_k^2 2p_k(1 - p_k) + \sigma_e^2},$$

where there could be multiple causal SNPs, $k = 1,\ldots, S$. In reality, h^2 is estimated by the correlation between phenotypes of related individuals. Consider the simple case of sibling pairs, and let Y_1 and Y_2 be the phenotypes for sib 1 and sib 2, respectively. Allowing for HWD and adjusting for the kinship coefficient ϕ, the estimated h^2 is

$$\hat{h}^2 = \widehat{\text{corr}}(Y_1, Y_2) \Big/ 2\phi,$$

where $\text{corr}(Y_1, Y_2)$ depends on the correlation between G_{1k} and G_{2k} between the siblings; see Zhang and Sun (2021) for the derivation of $\text{corr}(G_{1k}, G_{2k})$ accounting for kinship coefficient and HWD. Thus,

$$\frac{E\left(\hat{h}^2\right)}{h^2} = \frac{\sum \beta_k^2 2\left(p_k(1 - p_k) + \delta_k\right)}{\sum \beta_k^2 2p_k(1 - p_k)},$$

and the bias of the h^2 estimate is

$$E\left(\hat{h}^2\right) - h^2 = h^2 \cdot \frac{\sum \beta_k^2 \delta_k}{\sum \beta_k^2 p_k(1 - p_k)}. \tag{5}$$

Under the simple case of one causal SNP, the bias is simplified to $h^2 \cdot \delta/(p(1 - p))$.

Given the analytical insights provided so far, we then briefly examine the empirical properties of T_{LMM} through both application and simulation studies.

3 RESULTS

3.1 Cystic Fibrosis Sib-Pair Data Application: T_{LMM} Has Correct Type I Error but h^2 Appears to Be Overestimated

We extracted 65 sibling pairs from a cystic fibrosis (CF) gene modifier study (Wright et al., 2011; Sun et al., 2012). The phenotype Y of interest is the lung function measurements of the 130 related individuals with CF. In total, there were 570,539 SNPs genotyped using the Illumina 610-Quad Beadchip after applying the standard quality control, including minor allele frequency (MAF) greater than 2%. To stabilize the variance estimation, we additionally required SNPs to have MAF greater than 5%. We then applied T_{LMM} to the remaining 505,172 SNPs. In the application, we treated h^2 as unknown and estimated it based on the linear mixed-effect model of **Eq. 4** as in convention.

When h^2 was estimated from the data, our association testing based on T_{LMM} had good type I error control (results not shown), consistent with the empirical observations in the GWAS literature. However, the estimated h^2, obtained using the 65-pair sibling data, is $\hat{h}^2 = 0.82$. This value is substantially greater than 0.5, the commonly believed "true" heritability of lung function in CF obtained from the classic monozygous (MZ) vs. dizygous (DZ) twin-based estimation method (Vanscoy et al., 2007).

To verify if the large heritability estimate from the LMM method in our application was due to chance, we conducted a proof-of-principle simulation study. We assumed that only one causal SNP, G_{causal} with MAF of 0.2, affects Y with $h^2 = 0.5$. Genotype and phenotype values for 65 sibling pairs were then simulated under the assumption of HWE (i.e., without HWD). Among the 100,000 independently simulated replicates, only 4.24% of the heritability estimates were greater than $\hat{h}^2 = 0.82$. This suggests that $\hat{h}^2 = 0.82$, the value that was observed in the CF data application, was unlikely if the true heritability was 0.5 and without HWD at the causal SNP.

To verify if HWD at the causal SNP can lead to a biased heritability estimate, we then conducted additional simulation studies, following the same sib-pair design as mentioned previously. Our goal is to demonstrate that 1) when h^2 is treated as a nuisance parameter, its estimate based on model (**Eq. 4**) cross-reference can be biased in the presence of HWD; and 2) assuming the true h^2 is known, the empirical type I error rate of LMM (**Eq. 4**) cross-reference inflates when $\delta > 0$ and deflates when $\delta < 0$.

3.2 Simulated Sib-Pair Data in the Presence of HWD: h^2 Estimate Is Biased

Consider a continuous trait Y with $h^2 = 0.5$ and influenced by one causal SNP, G_{causal}, with minor allele frequency of 0.2 and with HWD factor, δ_{causal}, ranging from −0.04 to 0.16. A non-associated SNP, G_{tested}, also has an MAF of 0.2 but with its own δ_{tested}, which may not be the same as δ_{causal} in a specific simulation study. The sample size was 65 sibling pairs, chosen to match with the sample size of the cystic fibrosis application study in **Section 3.1**.

Most practical implementations of the linear mixed-effect model (**Eq. 4**) cross-reference treat h^2 as a nuisance parameter, and no type I error issue has been reported. Indeed, when h^2 was estimated in our simulation study conducted in **Section 3.3**, the test size of T_{LMM} was correct at the nominal level (black squares in **Figure 3** shown in **Section 3.3**) even if $\delta_{\text{tested}} \neq 0$ (i.e., out of HWE) and across the range of δ_{causal} values (from −0.04 to 0.16).

However, in this situation, when h^2 is treated as unknown, the impact of HWD is on the estimation of h^2. Specifically, **Figure 1**

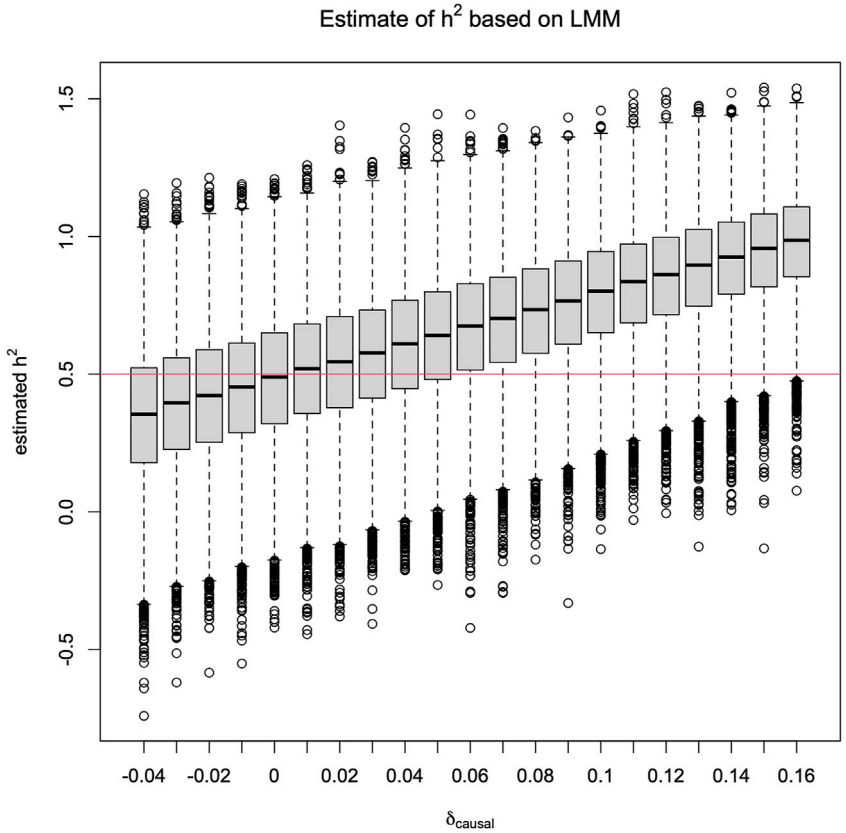

FIGURE 1 | Box plots of $\hat{h}^2$, estimated from the linear mixed-effect model (**Eq. 4**) against $\boldsymbol{\delta_{causal}}$. The true heritability of the phenotype is $h^2 = 0.5$. The minor allele frequencies $p_{causal} = p_{tested} = 0.2$ and 10,000 independent replicates of phenotypes and genotypes for 65 sibling pairs were simulated for each δ_{causal} value. The empirical type 1 error rates are shown in **Figure 3** as black squares.

shows that $\hat{h}^2$ is downward biased when $\delta_{causal} < 0$, and upward biased if $\delta_{causal} > 0$. The bias can be substantial. For example, when $\delta_{causal} = 0.10$, the estimated heritability $\hat{h}^2$ is centered at 0.80 as compared to the true value of 0.5, with a bias of 0.30. Indeed, based on our theoretical insight in **Section 2.2**, the expected bias is $h^2 \cdot \delta/(p(1-p)) = 0.5 \cdot 0.1/(0.2(1–0.2)) = 0.31$.

In **Figure 1**, it is notable that $\hat{h}^2$ can be greater than one. Since h^2 is the proportion of variance in Y explained by additive genetic variation, $0 \le h^2 \le 1$ by definition. However, if $\delta_{causal} \neq 0$, $\hat{h}^2$ based on the LMM, without additional truncation, is a biased estimate of h^2 with a bias of $h^2 \cdot \delta/(p(1-p))$ for this sib-pair design as shown in **Section 2.2**; the bias is 0 (i.e., no bias) under HWE when $\delta = 0$.

Additionally, although a larger sample that consists of 5,000 sibling pairs shrinks the variance of the h^2 estimate as expected, it does not shrink the bias, as shown in **Figure 2**. However, we also note that, in practice, it is unlikely to have so many sibling pairs.

3.3 Simulated Sib-Pair Data in the Presence of HWD: When Using the True h^2 Value T_{LMM} Has Incorrect Test Size

Here, we conducted the association analysis between Y and the non-associated SNP, G_{tested}, using the LMM model of **Eq. 4** but assuming $h^2 = 0.5$ is known.

Figure 3A plots the empirical type I error rates (blue circles) of T_{LMM} using the true $h^2 = 0.5$, for a nominal level of 0.05, estimated from independently simulated 10,000 replicates for each δ_{causal} value. (An empirical type I error greater than $0.05 + 3 \cdot 0.002 = 0.056$ can be considered inflated as the standard error of the empirical type I error rate can be estimated as $\sqrt{0.05 \cdot 0.95/10000} = 0.002$.) In **Figure 3A**, the trend of type I error inflation is clear as δ_{causal} increases.

In **Figure 3A**, we set $\delta_{tested} = 0.06$, but we note that the main cause of the type I error issue is $\delta_{causal} \neq 0$ when using the LMM of **Eq. 4** with $h^2 = 0.5$ plugged in. Indeed, **Figure 3B** shows that even if G_{tested} is in HWE (i.e., $\delta_{tested} = 0$), the problem remains, albeit less severe, as long as $\delta_{causal} \neq 0$.

4 DISCUSSION

We used a sib-pair design to demonstrate that the linear mixed-effect model can be problematic in the presence of Hardy–Weinberg disequilibrium at the causal SNP(s). To demonstrate that the LMM-based heritability estimate can be biased, as a proof-of-principle, our simulation study assumed that the phenotype Y has only one causal SNP, which is unrealistic for complex traits. However, the analytical insight shown in **Eq. 5**

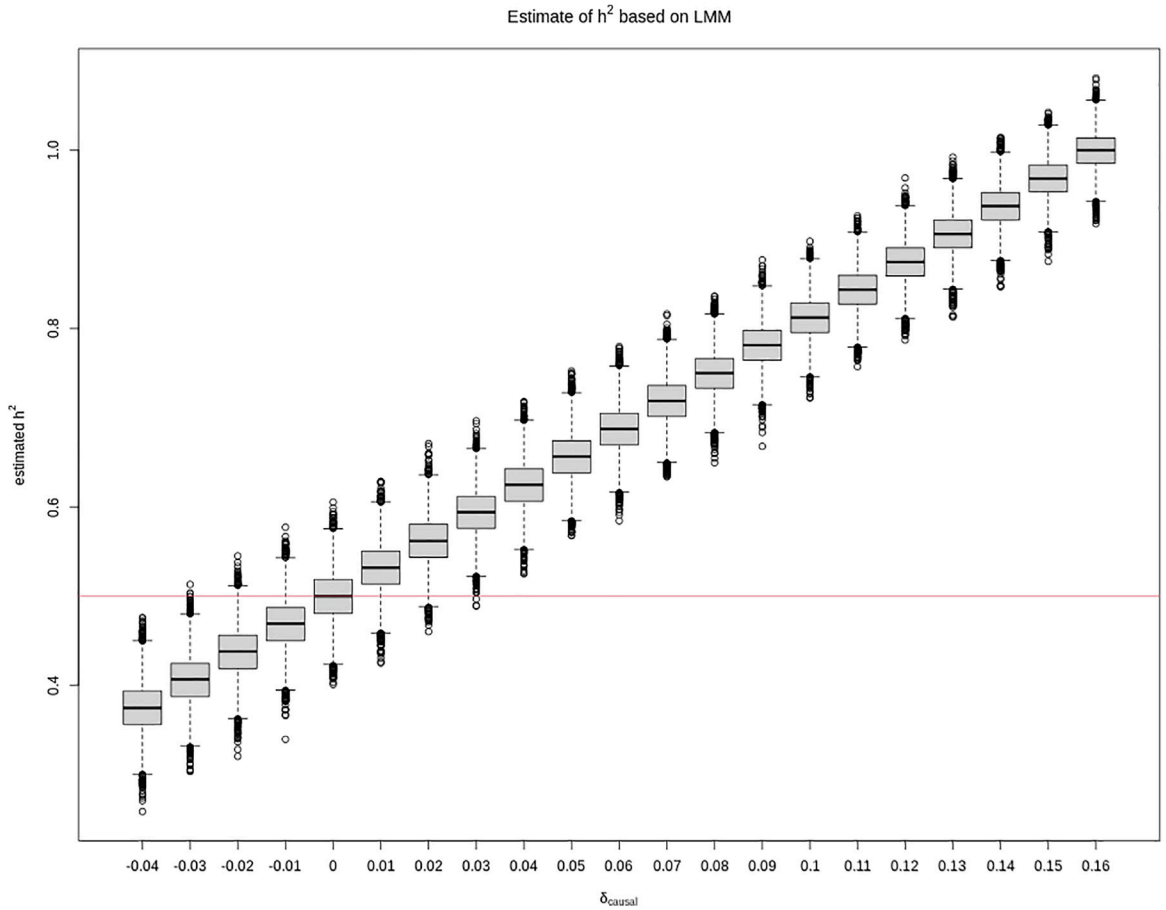

FIGURE 2 | Box plots of $\hat{h}^2$, estimated from the linear mixed-effect model (**Eq. 4**) against δ_{causal} using 5,000 sibling pairs. The other parameter values are the same as in **Figure 1**, where the true heritability of the phenotype is $h^2 = 0.5$, the minor allele frequencies $p_{\text{causal}} = p_{\text{tested}} = 0.2$, and 10,000 independent replicates were simulated for each δ_{causal} value.

(i.e., bias expected to be $h^2 \cdot \sum \beta_k^2 \delta_k / \sum \beta_k^2 p_k (1 - p_k)$) suggests that the issue discussed here remains relevant in the case of multiple causal SNPs as $\sum \beta_k^2 \delta_k$ is unlikely to be zero, even while allowing the signs of δ_k to differ.

Assuming the true heritability h^2 is known, we also demonstrated the potential type I error issue of the LMM in the presence of HWD using data that consist of related individuals only. In practice, this issue diminishes if the sample includes a large number of independent individuals or the magnitude of HWD at the causal SNP is small. Additionally, in practice, h^2 is treated as unknown, in which case, the type I error rate of the LMM is well controlled; indeed, no increased false positives of the LMM due to HWD have been reported in the literature to the best of our knowledge. However, the estimate of h^2 can be upward biased and upwardly so if $\delta_{\text{causal}} > 0$, as demonstrated in the simulation study in **Section 3.2** and seen in the cystic fibrosis application study in **Section 3.1**. This new observation offers a possible complementary explanation of the "missing heritability" discussed extensively in Maher (2008).

In practice, SNPs out of HWE are typically not analyzed due to concerns for low genotyping quality (Wellcome Trust Case Control Consortium, 2007; Bycroft et al., 2018; Marees et al., 2018). However, the observation made here remains relevant as the heritability estimates in LMM-based models are biased when the causal SNPs are in HWD (which is unknown in practice) but not the tested SNPs. This is also supported by **Figure 3B**. When there was HWD at the causal SNP (e.g., $\delta_{\text{causal}} = 0.10$ on the X-axis), there was a type I error issue even if there was no HWD at the tested SNP (i.e., $\delta_{\text{tested}} = 0$). Conversely, **Figure 3A** shows that if there was no HWD at the causal SNP (i.e., $\delta_{\text{causal}} = 0$ on the X-axis), then the test is accurate even if there was HWD at the tested SNP ($\delta_{\text{tested}} = 0.06$).

Additionally, the HWE-based screening practice itself can be called into question because a truly associated SNP is often in HWD (Wittke-Thompson et al., 2005; Ryckman and Williams, 2008; Turner et al., 2011). The potential of leveraging the HWD expected at a causal SNP to increase the power of association testing has been explored by several groups (Song and Elston, 2006; Wang and Shete, 2008; Zhang and Sun, 2020).

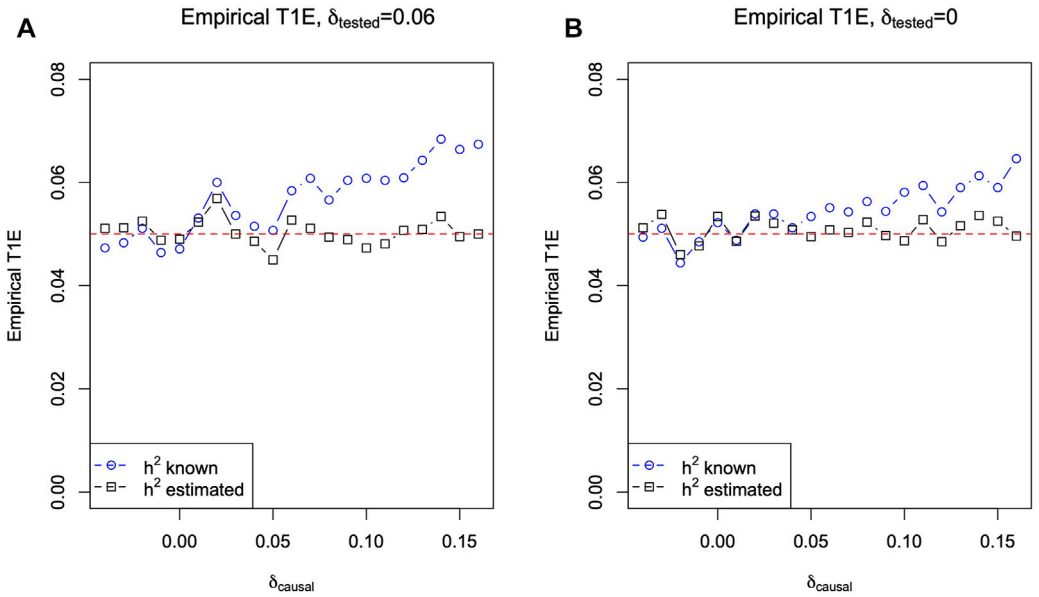

FIGURE 3 | Empirical type I error rate of T_{LMM} based on the linear mixed-effect model (**Eq. 4**) against $\boldsymbol{\delta_{causal}}$. **(A)** When G_{tested} of tested SNPs is in HWD with δ_{tested} = 0.06. **(B)** When G_{tested} of tested SNPs is in HWE with δ_{tested} = 0. The true heritability of the phenotype is h^2 = 0.5, the minor allele frequencies $p_{causal} = p_{tested}$ = 0.2, and 10,000 independent replicates of phenotypes and genotypes for 65 sibling pairs were simulated for each δ_{causal} value. The blue circles are for T_{LMM} using the true heritability h^2 = 0.5, and the black squares are for T_{LMM} while estimating h^2 (results of h^2 are shown in **Figure 1**).

We have not examined the implication of HWD combined with linkage disequilibrium (LD) (Weir, 2008) on the LMM, which is an important future research question. Additionally, recent work has shown that dominant genetic effect could complicate the LD measure and interpretation (Palmer et al., 2021), which in turn could affect our examination of the effect of HWD on the LMM.

Although the linear mixed-effect model is a popular and powerful method for GWAS, conceptually, the use of kinship coefficient matrix (i.e., Σ_Φ), derived from G, as part of the variance–covariance matrix (i.e., Σ_y) of the LMM can be problematic because the response variable Y is the phenotype of interest. An alternative approach is to reverse the roles of Y and G in the regression model. Indeed, O'Reilly et al. (2012) proposed MultiPhen, a method that treats the genotype G of an SNP as the response variable and phenotype values Y of multiple traits as predictors, and uses an ordinal logistic regression applicable to independent samples. More recently, Zhang (2021) (Chapter 2) proposed a generalized reverse (or retrospective) regression model that can be applied to dependent samples, which takes the form of

$$g = \alpha 1 + \beta y + \gamma z + \epsilon, \quad \epsilon \sim N\left(0, \sigma^2\Sigma_g\right), \quad \sigma^2\Sigma_g = \sigma^2\Sigma_\Phi + \Sigma_\delta, \tag{6}$$

where Σ_Φ is the kinship coefficient matrix as defined earlier and Σ_δ is a function of δ that explicitly models the amount of HWD; the use of a linear model for the discrete genotype data G is motivated by the work of Chen (1983).

Interestingly, if the variance and covariance matrices in **Eqs 4**, **6** of the LMM were the same, the resulting score test statistics are also the same. However, conceptually, the model **Eq. 6** correctly uses the kinship coefficient matrix to model the response variable G, in contrast to the LMM model of **Eq. 4**. Specifically, at a tested SNP, as the reverse regression is conditional on Y, the variance–covariance matrix only concerns G_{tested}, that is, Σ_g. The modeling and estimation of Σ_g can account for potential HWD through Σ_δ, in addition to the genetic correlation captured by the kinship coefficient matrix of Σ_Φ, resulting in a more robust association test for related individuals. Indeed, when the method was applied to the same simulated sib-pair data in **Section 3.3**, it had correct type I error control [results shown in Figure 2.2 of Chapter 2 of Zhang (2021)]. However, how to model gene–environment interaction through the reverse regression framework remains an open question.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are subject to the following licenses/restrictions: The CF application data are available by application to the Cystic Fibrosis Canada National data registry for researchers who meet the criteria for access to confidential clinical data for the purpose of CF research. Requests to access these datasets should be directed to cfregistry@cysticfibrosis.ca.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Canadian Gene Modifier Study (CGMS), the Research Ethics Board of the Hospital for Sick Children (# 0020020214 from 2012–2019 and #1000065760 from 2019-present), and all participating subsites. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

LZ and LS proposed the method and wrote the manuscript. LZ performed the analysis. LS obtained the application data and funding.

FUNDING

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-04934 and RGPAS-522594).

ACKNOWLEDGMENTS

We thank Dr. Lisa J. Strug and acknowledge her laboratory for providing the cystic fibrosis genotype data. LZ was a trainee and funding recipient of the CANSSI Ontario STAGE (Strategic Training for Advanced Genetic Epidemiology) program at the University of Toronto.

REFERENCES

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The uk Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* 562, 203–209. doi:10.1038/s41586-018-0579-z

Chen, C.-F. (1983). Score Tests for Regression Models. *J. Am. Stat. Assoc.* 78, 158–161. doi:10.1080/01621459.1983.10477945

Chen, G.-B. (2014). Estimating Heritability of Complex Traits from Genome-wide Association Studies Using IBS-Based Hasemanâ€"Elston Regression. *Front. Genet.* 5, 107. doi:10.3389/fgene.2014.00107

Derkach, A., Lawless, J. F., and Sun, L. (2015). Score Tests for Association under Response-dependent Sampling Designs for Expensive Covariates. *Biometrika* 102, 988–994. doi:10.1093/biomet/asv038

Dimitromanolakis, A., Paterson, A. D., and Sun, L. (2019). Fast and Accurate Shared Segment Detection and Relatedness Estimation in Un-phased Genetic Data via Truffle. *Am. J. Hum. Genet.* 105, 78–88. doi:10.1016/j.ajhg.2019.05.007

Eu-Ahsunthornwattana, J., Miller, E. N., Fakiola, M., Jeronimo, S. M. B., Blackwell, J. M., Cordell, H. J., et al. (2014). Comparison of Methods to Account for Relatedness in Genome-wide Association Studies with Family-Based Data. *Plos Genet.* 10, e1004445. doi:10.1371/journal.pgen.1004445

Falconer, D. S. (1985). A Note on Fisher's 'average Effect' and 'average Excess'. *Genet. Res.* 46, 337–347. doi:10.1017/s0016672300022825

Hardy, G. H. (1908). Mendelian Proportions in a Mixed Population. *Science* 28, 49–50. doi:10.1126/science.28.706.49

Maher, B. (2008). Personal Genomes: The Case of the Missing Heritability. *Nature* 456, 18–21. doi:10.1038/456018a

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature* 461, 747–753. doi:10.1038/nature08494

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et al. (2018). A Tutorial on Conducting Genome-wide Association Studies: Quality Control and Statistical Analysis. *Int. J. Methods Psychiatr. Res.* 27, e1608. doi:10.1002/mpr.1608

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M. R., et al. (2012). Multiphen: Joint Model of Multiple Phenotypes Can Increase Discovery in Gwas. *PLoS One* 7, e34861. doi:10.1371/journal.pone.0034861

Palmer, D. S., Zhou, W., Abbott, L., Baya, N., Churchhouse, C., Seed, C., et al. (2021). Analysis of Genetic Dominance in the uk Biobank. bioRxiv

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the Analysis of Ibd and Ibs in Complex Trait Studies. *Nat. Rev. Genet.* 11, 800–805. doi:10.1038/nrg2865

Ryckman, K., and Williams, S. M. (2008). Calculation and Use of the hardy-weinberg Model in Association Studies. *Curr. Protoc. Hum. Genet.* Chapter 1, Unit–18. doi:10.1002/0471142905.hg0118s57

Sasieni, P. D. (1997). From Genotypes to Genes: Doubling the Sample Size. *Biometrics* 53, 1253–1261. doi:10.2307/2533494

Schaid, D. J., and Jacobsen, S. J. (1999). Blased Tests of Association: Comparisons of Allele Frequencies when Departing from Hardy-Weinberg Proportions. *Am. J. Epidemiol.* 149, 706–711. doi:10.1093/oxfordjournals.aje.a009878

Song, K., and Elston, R. C. (2006). A Powerful Method of Combining Measures of Association and Hardy-Weinberg Disequilibrium for fine-mapping in Case-Control Studies. *Statist. Med.* 25, 105–126. doi:10.1002/sim.2350

Sun, L., Dimitromanolakis, A., and Chen, W.-M. (2017). "Identifying Cryptic Relationships," in *Statistical Human Genetics* (Springer), 45–60. doi:10.1007/978-1-4939-7274-6_4

Sun, L., Rommens, J. M., Corvol, H., Li, W., Li, X., Chiang, T. A., et al. (2012). Multiple Apical Plasma Membrane Constituents Are Associated with Susceptibility to Meconium Ileus in Individuals with Cystic Fibrosis. *Nat. Genet.* 44, 562–569. doi:10.1038/ng.2221

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011). Quality Control Procedures for Genome-wide Association Studies. *Curr. Protoc. Hum. Genet.* Chapter 1, Unit1–19. doi:10.1002/0471142905.hg0119s68

Vanscoy, L. L., Blackman, S. M., Collaco, J. M., Bowers, A., Lai, T., Naughton, K., et al. (2007). Heritability of Lung Disease Severity in Cystic Fibrosis. *Am. J. Respir. Crit. Care Med.* 175, 1036–1043. doi:10.1164/rccm.200608-1164oc

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the Genomics Era - Concepts and Misconceptions. *Nat. Rev. Genet.* 9, 255–266. doi:10.1038/nrg2322

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., et al. (2006). Assumption-free Estimation of Heritability from Genome-wide Identity-By-Descent Sharing between Full Siblings. *Plos Genet.* 2, e41. doi:10.1371/journal.pgen.0020041

Wang, J., and Shete, S. (2008). A Test for Genetic Association that Incorporates Information about Deviation from hardy-weinberg Proportions in Cases. *Am. J. Hum. Genet.* 83, 53–63. doi:10.1016/j.ajhg.2008.06.010

Weinberg, W. (1908). *On the Demonstration of Heredity in Man*. in *(1963) Papers on Human Genetics*.

Weir, B. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, Massachusetts: Sinauer Series (Sinauer).

Weir, B. S. (2008). Linkage Disequilibrium and Association Mapping. *Annu. Rev. Genom. Hum. Genet.* 9, 129–142. doi:10.1146/annurev.genom.9.081307.164347

Wellcome Trust Case Control Consortium (2007). Genome-wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature* 447, 661–678. doi:10.1038/nature05911

Wittke-Thompson, J. K., Pluzhnikov, A., and Cox, N. J. (2005). Rational Inferences about Departures from hardy-weinberg Equilibrium. *Am. J. Hum. Genet.* 76, 967–986. doi:10.1086/430507

Wright, F. A., Strug, L. J., Doshi, V. K., Commander, C. W., Blackman, S. M., Sun, L., et al. (2011). Genome-wide Association and Linkage Identify Modifier Loci of Lung Disease Severity in Cystic Fibrosis at 11p13 and 20q13.2. *Nat. Genet.* 43, 539–546. doi:10.1038/ng.838

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* 88, 76–82. doi:10.1016/j.ajhg.2010.11.011

Zhang, L. (2021). *A General Study of Genetic Association Tests and the Test of Hardy–Weinberg Equilibrium*. Ph.D. thesis. University of Toronto.

Zhang, L., and Sun, L. (2021). *A Generalized Robust Allele-Based Genetic Association Test*. Oxford, United Kingdom: Biometrics.

Zhang, L., and Sun, L. (2020). Leveraging hardy-weinberg Disequilibrium for Association Testing in Case-Control Studies. bioRxiv.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.