Accurate modeling of replication rates in genome-wide association studies by accounting for winner's curse and study-specific heterogeneity

Jennifer Zou[1], Jinjing Zhou[1], Sarah Faller[2], Robert Brown[1], and Eleazar Eskin[1,3*]

[1]Computer Science Department, University of California Los Angeles, CA, USA

[2]Computer Science Department, Duke University, Durham, NC, USA

[3]Human Genetics Department, University of California Los Angeles, CA, USA

Abstract

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex human traits, but only a fraction of variants identified in discovery studies achieve significance in replication studies. Replication in GWAS studies has been well-studied in the context of winner's curse, which is the inflation of effect size estimates for significant variants in a study. Multiple methods have been proposed to correct for the effects of winner's curse. However, winner's curse is often not sufficient to explain lack of replication. Another reason why studies fail to replicate is that there are fundamental differences between the discovery and replication studies. A confounding factor can create the appearance of a significant finding while actually being an artifact that will not replicate in future studies. We propose a statistical framework that utilizes GWAS replication studies to model winner's curse and study-specific heterogeneity due to confounders and correct for these effects. We show through simulations and application to 100 human GWAS data sets that modeling both winner's curse and study-specific heterogeneity explains observed patterns of replication in GWAS studies better than modeling winner's curse alone.

Introduction

Replication is a gold standard in scientific discovery. Consensus emerges when a result has been replicated repeatedly by multiple researchers. Recently, a vigorous discussion has emerged of how often replication of an initial study fails across all fields of science, including genomics [1, 2, 3, 4, 5].

Genome-wide association studies (GWAS) are an ideal model to study replication because there are a large number of GWAS data sets with replication studies publicly available. GWAS replication studies are typically conducted in an independent cohort and on a smaller set of variants than the discovery study. In the National Human Genome Research Institute Catalog of Published GWAS, thousands of genetic variants have been associated with complex human traits but not all associated variants achieve significance in follow up replication studies [6, 4, 5, 7].

There are several reasons why associations do not replicate. The first is simply statistical. It is possible that the association is not observed in the replication study by chance. However, if the p-value from the original finding is highly significant and the replication studies have similar experimental designs, this scenario is unlikely. A second reason why studies can fail to replicate is winner's curse, which is the inflation of effect size estimates for significant variants in a study. This phenomenon occurs because the reported findings are a small fraction of many possible findings. In the case of GWAS, the significant associations are discovered after examining millions of variants and pass a stringent genome-wide significance threshold. This can result in inflated effect size estimates of significant variants in a study, especially when studies are underpowered. Winner's curse has been studied extensively in GWAS, and multiple methods have been proposed to correct for its effects [9, 10, 11, 12, 4]. However, winner's curse is often not sufficient to explain lack of replication. A third reason why studies fail to replicate is that there are fundamental differences between the original and the replication study. An effect present in one study but not present in other studies can create the appearance of a significant finding that is not replicated in future studies [13]. This can either occur because of an underlying biological difference or a technical difference between the two studies. We refer to the cause of these differences as confounders.

Current methods for modeling confounders fall into two broad categories. The first class of methods attempt to model the effect of confounders before the association statistic is calculated in order to remove their effects from the association statistic. While these methods are widely used, they have several fundamental limitations. Methods that account for known covariates may not correct for all potential confounders. Confounding correction methods that use unsupervised learning to learn principal components or other global patterns in the data can incorrectly model the true signal as a confounder, which would remove true biological signal from the data [15, 14]. Similarly, when using unsupervised methods, it is unclear when there is residual confounding that remains in the data. The second class of methods attempt to directly adjust p-values by a constant factor to remove inflation. An example of such a method is genomic control [16]. In genomic control, there is an assumption that relatively few variants affect the trait and the vast majority do not. The implication of this assumption is that if the association statistics are ranked, then the variant corresponding to the median statistic will not affect the trait, and the value of this statistic will represent only the effect of the confounders. Genomic control scales all of the p-values using this statistic. Recently it has been observed that due to polygenicity and linkage disequilibrium (LD) structure in the genome, the majority of variants (including the one corresponding to the median statistic) either affect the trait or are correlated with a variants that affect the trait. This breaks the genomic control assumption. While LD-score regression has been shown to distinguish polygenicity and confounding [17], it has been shown that this approach can also result in inflated SNP-based heritability estimates under strong stratification [18].

In this paper, we present a novel approach for characterizing study-specific heterogeneity due to confounders using replication studies. The key insight in our approach is that we can use replications to identify the presence of confounders and then use this information to correct the studies. Since replication studies are performed on the same phenotype, utilizing replication studies to estimate the effect of confounders does not reply on assumptions to distinguish between polygenicity and confounding. Furthermore, we can apply our approach in combination with traditional techniques like linear mixed models and principal component analysis. Our approach can be used to model any residual confounding effects after application of these methods.

In our framework, we use a random effects model to jointly model the effect of both winner's curse and study-specific confounders on GWAS summary statistics. We show through simulations that we can accurately estimate the contribution of confounders on a study by using the existing findings of the study and a replication. We apply this framework to 100 GWAS studies from the Human GWAS Catalog and observe a surprising amount of confounding in GWAS studies. We validate our approach by comparing the predicted replication rate under our model with both the true replication rate and the predicted replication rate under a naive model that only accounts for winner's curse. We show that modeling both winner's curse and study-specific heterogeneity due to confounders explains observed patterns of replication in GWAS studies better than modeling winner's curse alone.

Results

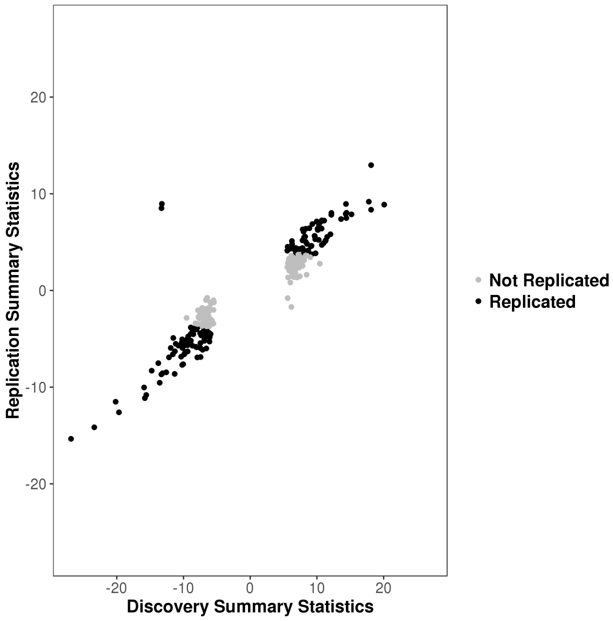
Method overview

The main goal of this framework is to account for winner's curse and confounding between discovery and replicate GWAS studies of the same phenotype. We compare the predicted replication rate of two random effects models — one that corrects for only winner's curse and one that corrects for both winner's curse and confounding. Through this comparison, we show that jointly modeling both winner's curse and study-specific confounders explains observed patterns of replication better than the naive approach that only models winner's curse. We introduce these models without accounting for difference in sample size for clarity, but we relax this constraint in the Methods section.

In GWAS, winner's curse is the phenomenon where the association statistics for variants meeting a genome-wide threshold tend to be overestimated. The effect of winner's curse can be observed in Figure 1, where the summary statistics for the significant variants in the discovery study are substantially lower in the replication study. Due to this phenomenon, not all of the significant variants in the initial discovery study replicate. Winner's curse is widely observed in GWAS due to lack of statistical power in initial discovery studies. When power is low, the variants that are most significant in a study are likely to have inflated effect sizes due to random noise.

To model random noise contributing to winner's curse, we model the statistics for each variant k from the initial and discovery studies as normally distributed random variables ($s_k^{(1)}$ and $s_k^{(2)}$, respectively). We assume that there is a shared genetic effect λ that is responsible for the observed association signal. Thus, the distribution of the statistic for variant k in study i is $s_k^{(i)} \sim \mathcal{N}(\lambda, 1)$. We define the prior probability of the true genetic effect to be $\lambda \sim \mathcal{N}(0, \sigma_g^2)$, where the variance in the true genetic effect is learned from the data. Then, we model the joint distribution of the

Figure 1: **Winner's curse.** Not all GWAS variants replicate in followup studies. The significant variants in a discovery GWAS study on height (PMID 25282103) are shown. The variants that replicated successfully are shown in black, and the ones that did not replicate are shown in grey. The summary statistics for the variants that did not replicate are lower than expected in the replication study. This phenomenon can be partially explained by winner's curse.



summary statistics from the two studies (Equation 1).

$$\begin{pmatrix} s_k^{(1)} \\ s_k^{(2)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_g^2 + 1 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 + 1 \end{pmatrix}\right) \tag{1}$$

We correct for winner's curse by computing the conditional distribution of the replication summary statistic given the initial summary statistic (Equation 2). Using this conditional distribution, we can account for winner's curse and compute the expected value of the summary statistic in the replication study, along with confidence intervals. This framework accurately models the data in cases where winner's curse is the only source of inflation. Figure 2A shows a GWAS on height, where most of the variants fall within the 95% confidence intervals of the model accounting for winner's curse [19]. This shows that in studies without substantial confounding effects, winner's curse can adequately explain the replication rate.

$$(s_k^{(2)} | s_k^{(1)} = x) \sim \mathcal{N}\left(\frac{\sigma_g^2}{\sigma_g^2 + 1}x, 1 + \sigma_g^2 - \frac{\sigma_g^2}{\sigma_g^2 + 1}\right) \tag{2}$$

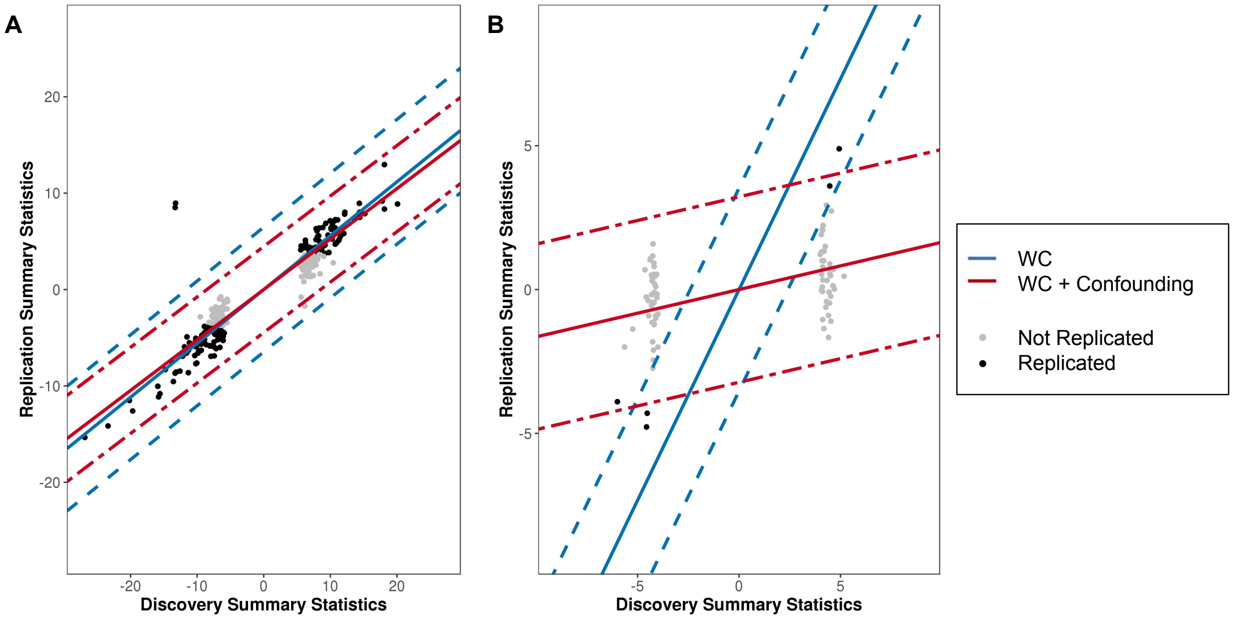


Figure 2: **Correcting for winner's curse and confounding** The x-axis for each plot is the value of the discovery summary statistic, and the y-axis is the value of the replication summary statistic. The solid lines correspond to the expected values of the replication summary statistics given the initial summary statistics. The dotted lines represent confidence intervals in the estimates. The blue lines correspond to the model that only accounts for winner's curse, and the red lines correspond to the model that accounts for winner's curse and confounding. A) In this GWAS on height (PMID 25282103), there is very little confounding, and a model that accounts for winner's curse explains the majority of the data. B. In this GWAS on height in African American women (PMID 22021425), there is substantial confounding. The model accounting for only winner's curse (blue) does not explain the observed data well, whereas the model with winner's curse and confounding (red) does explain the data well.

However, there is often additional heterogeneity due to confounding, and a framework that only accounts for winner's curse is inadequate. Figure 2B shows an example of a GWAS on height in African American women [20]. In this study there was substantial confounding, and only 5/84 (6%) of variants replicated. Using a model that only accounts for winner's curse, most variants are outside of the 95% confidence intervals, indicating that there is additional heterogeneity that is not modeled. To account for study-specific confounding, we decompose the effect size of the summary statistics into a genetic effect (λ) and study-specific confounding effects ($\delta^{(i)}$). The distribution of the statistic for variant k in study i is $s_k^{(i)} \sim \mathcal{N}\left(\lambda + \delta^{(i)}, 1\right)$. In addition to the prior on the genetic effect, we introduce priors on the study-specific confounders ($\delta^{(i)} \sim \mathcal{N}(0, \sigma^2_{c_i})$).We incorporate both of these priors into the joint distribution of the summary statistics (Equation 3).

$$\begin{pmatrix} s_k^{(1)} \\ s_k^{(2)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_g^2 + \sigma_{c_1}^2 + 1 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 + \sigma_{c_2}^2 + 1 \end{pmatrix}\right) \tag{3}$$

We correct for both winner's curse and confounding by computing the conditional distribution of the replication summary statistic given the initial summary statistic (Equation 4). By taking into account the additional variance in the statistics from confounders, we are able to more accurately model the summary statistic data from the two studies (Figure 2B). The model that only accounts for winner's curse predicted that 84 variants would replicate, whereas our model that also accounts for confounding predicted that only 4 variants would replicate, which is substantially closer to the true value of 5 variants. This difference in predictions is due to the study-specific confounding effects estimated in the second model, which both decreases the expected value of the statistics in the replication study and increases the variance of the statistics in the replication study. After correcting for winner's curse and confounding, most variants are within the 95% confidence intervals for the model.

$$(s_k^{(2)} | s_k^{(1)} = x) \sim \mathcal{N}\left(\frac{\sigma_g^2}{\sigma_g^2 + \sigma_{c_1}^2 + 1} x, \sigma_g^2 + \sigma_{c_2}^2 + 1 - \frac{\sigma_g^4}{\sigma_g^2 + \sigma_{c_1}^2 + 1}\right) \tag{4}$$

For each data set, we compute estimates of the summary statistics that we would expect using a framework that only accounts for winner's curse and a framework that accounts for winner's curse and confounding. We also compute the expected replication rate under the two models. We apply this framework to simulated data and 100 human GWAS in the GWAS catalog. We compare the predicted replication rates under the two models with the true replication rate.

Confounding explains low replication in simulated data

We generated simulated data to demonstrate that our approach accurately models the effects of winners' curse and confounding to explain low replication in GWAS studies (See Methods). We set the variance for the genetic and confounding effects to a range of values from 0.0 to 3.0. Using multiple combinations of fixed parameters, we simulated summary statistics for 1000 variants by drawing the shared genetic effect, study-specific confounders, and study-specific error from normal distributions. We then computed the summary statistics for each variant as the sum of the genetic effect, the study-specific confounder, and the study-specific error. We define the true replication rate to be the percentage of variants that are significant in the discovery study that are also significant in the replication study using a Bonferonni correction for multiple testing.

We directly compared our method with a simplified model that only takes into account winner's curse. When only accounting for winner's curse, the predicted replication rate was often much higher than the true replication rate (Figure 3). The winner's curse model only accurately predicted the replication rate when the confounding for both studies is set to zero (i.e, $\sigma_{c_1}^2 = 0$ and $\sigma_{c_2}^2 = 0$). This indicates that when confounding exists between two GWAS studies, the two studies may have different effect sizes. Thus, a model that only accounts for winner's curse may overestimate the expected replication in the presence of confounding.

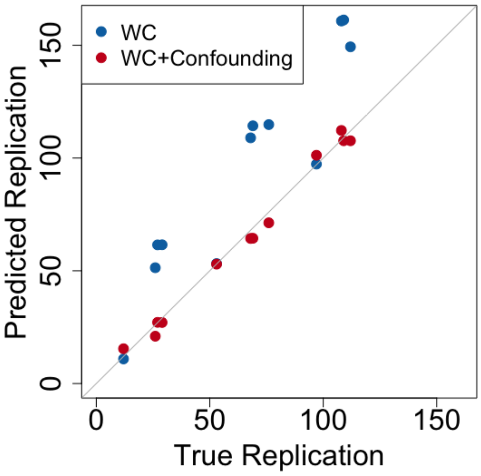


Figure 3: **Prediction of replication rate in simulated data.** We predicted the replication rate under a model accounting for winner's curse (WC) and a model accounting for winner's curse and confounding (WC+Confounding). The model accounting for winner's curse and confounding explains replication rates substantially better than the model that only accounts for winner's curse.

We then applied our method that takes into account both winner's curse and study-specific confounding. For simulations where the confounding is greater than zero, the predicted replication rate under this model was closer to the true replication rate than the simplified model that only accounts for winner's curse (Figure 3). To ensure that our maximum likelihood estimates of the variance parameters were accurate, we compared the estimates with the true values. For all simulations, the maximum likelihood estimates of the variance parameters are close to the true values (Figure 4).

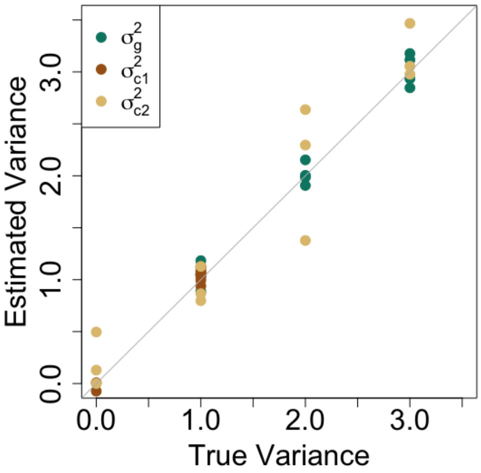


Figure 4: **Prediction of variance components in simulated data.** We fixed the values of the variance components to a range of values between 0.0 and 3.0 and simulated data according to our framework. We estimated the parameters from the simulated data using our maximum likelihood estimates.

Accounting for confounding better explains replication rate in 100 human GWAS datasets

We then apply our framework to 100 human GWAS studies previously curated to require summary statistic data availability, a focus on human genetics, and other consistency criteria [4]. All studies have a discovery and replication design, where only the significant SNPs in the discovery study are tested in the replication study. We use the summary statistics from these discovery and replication studies to test our framework's ability to capture the effects of winners' curse and confounding. After learning the variance parameters for the genetic and confounding effects, we calculate the predicted replication rate under the model accounting for winner's curse and the model accounting for winner's curse and confounding (See Methods). We compare these predicted replication rates to the true replication rates to assess which model explains the observed replication better.

We define the true replication rate to be the percentage of variants that are significant in the discovery study that are also significant in the replication study. We use a Bonferonni adjusted p-value threshold of $\alpha = \frac{0.05}{M}$ for each study, where M is the number of SNPs tested in the study. Of the 1652 reported GWAS variants, only 519 (31%) replicate. Using the simplified model that does not account for confounding, we would expect 1552 (94%) of the variants to replicate. However,

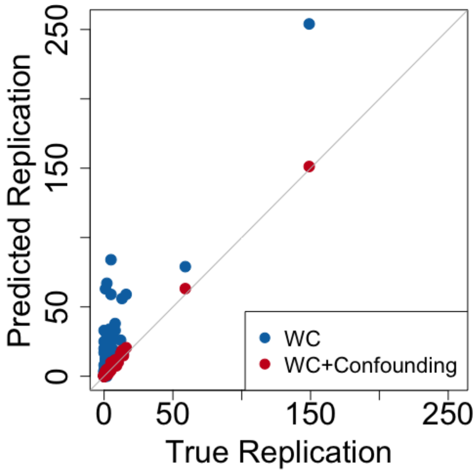


Figure 5: **Prediction of replication rate by study in 100 human GWAS.** The x-axis is the true number of variants that replicate. The y-axis is the predicted number of variants that replicate. Each dot represents one GWAS study, and the color indicates which model was used to predict the number of variants that replicate. Accounting for winner's curse and confounding yields more accurate estimates of the replication than the model that only accounts for winner's curse.

when we estimate the effect of confounding in our framework, we expect 548 (33%) of the variants to replicate, which is very close to the observed value thus giving evidence that we do observe a substantial bias beyond what we would expect from winner's curse alone. Our study-specific replication rates also show that accounting for confounding improves prediction of replication rate, indicating that accounting for confounding is important for understanding patterns of replication across studies (Figure 5).

We compare our predicted replication rates with those previously reported by Pe'er et al., which corrects for the expected bias in observed effect due to winner's curse in the same 100 GWAS studies [4]. At a Bonferonni adjusted significance level of 0.05, Palmer et al. predicts that 610 loci will replicate, which is more than both the true replication rate (519) and the predicted replication rate using our method when accounting for confounders (548). This suggests that by utilizing replication studies, we can account for more heterogeneity due to confounding and explain replication better than adjusting for winner's curse alone.

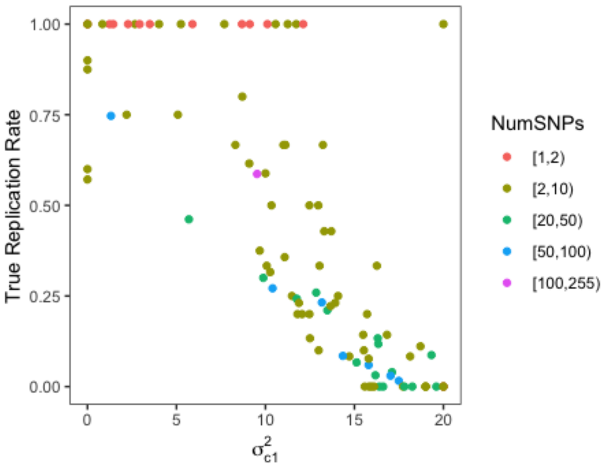


Figure 6: **Estimated level of confounding in discovery study strongly associated with true replication rate.** The x-axis is the MLE of $\sigma^2_{c_1}$, and the y-axis is the true replication rate. Each dot represents a single GWAS study. The Spearman correlation between the estimated variance of confounding and true replication rate is -0.84. The colors correspond to the number of significant SNPs in the discovery GWAS.

Estimates of study-specific confounding elucidates lack of replication

Our framework models additional variation in the summary statistics that is due to study-specific confounders. To further assess the effect of our estimated levels of confounding on replication rate, we analyzed the distribution of estimated confounding in all studies. The estimated value of $\sigma^2_{c_1}$ was negatively correlated (Spearman $\rho = -0.84$) with the true replication rate in these studies (Figure 6). In many studies, the level of confounding estimated was substantial. In many studies we quantified the variance due to confounding to be an order of magnitude larger than the variance due to noise (WC). We stratified the studies by the total number of significant variants in the discovery study since our estimates of $\sigma^2_{c_1}$ may be less robust for studies with only one significant variant. In studies with at least 50 significant variants, the correlation between confounding and true replication rate is strongest (Spearman $\rho = -0.95$). For subsequent analyses, we focused only on studies that have at least 50 significant SNPs in the discovery GWAS (8 studies).

In order to understand why some studies replicate poorly, we analyzed the ancestry of the discovery and replication studies. When GWAS are performed in populations with different ancestries, differences in the true effect sizes between populations can contribute to lack of replication. Thus, we expect that studies using homogenous populations would replicate better than studies using heterogenous populations. However, we observed a range of confounding and replication for both

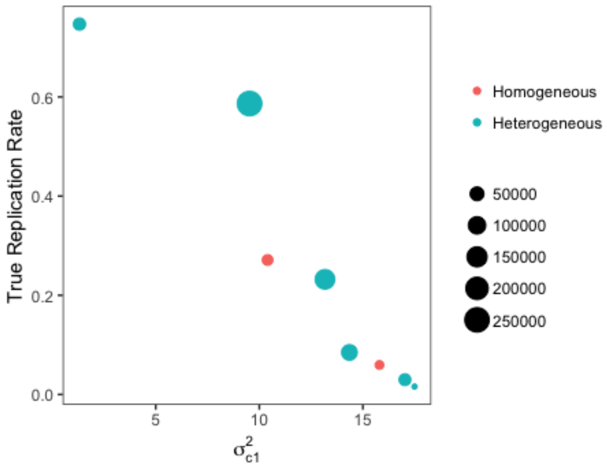


Figure 7: **Ancestry differences and sample size explain replication inconsistently.** The x-axis is the MLE of $\sigma^2_{c_1}$, and the y-axis is the true replication rate. Each dot represents a single GWAS study. The Spearman correlation between the estimated variance of confounding and true replication rate is -0.84. The colors correspond to whether the studies were conducted in a homogeneous or heterogeneous population. The size corresponds to the number of individuals in the discovery study. While the estimated confounding in the discovery study explains the replication rate well, ancestry and size do not explain the replication consistently.

types of studies (Figure 6). For instance, the studies using heterogeneous populations had replication rates ranging from 2% to 75% [22, 23, 24, 25, 19, 26]. Of the two studies from homogeneous populations, one study had a replication rate of 27% [27], while the other had a replication rate of only 2% [20]. While ancestry explains replication inconsistently, our estimates of confounding can distinguish between studies where ancestry is correctly accounted for and studies where it is not (Figure 6).

Another potential cause of poor replication is sample size. When sample sizes are small, winner's curse may contribute to lack of replication in GWAS studies. The study with the smallest sample size (176 individuals) also had the lowest replication rate (1%) and highest amount of confounding ($\sigma^2_{c_1} = 17.5$) [22]. While the correlation between sample size and true replication is quite high (Spearman $\rho = .46$), there are some studies where smaller sample sizes have higher replication rates and vice versa. Our model can be used to identify when small sample sizes negatively affect replication rate.

Discussion

We developed a novel statistical framework to correct for winner's curse and study-specific confounding in GWAS data. This framework utilizes GWAS replications to identify the presence of confounders without replying on assumptions to distinguish between polygenicity and confounding.

We showed through simulations that our model accurately estimates the variance of the genetic and confounding effects and that our model can be used to explain replication rates. When applying our method to 100 human GWAS studies, we showed that a model that accounts for winner's curse and confounding explains replication rates more accurately than a model that only accounts for winner's curse. While the estimated confounding in the discovery study explains the replication rate well, ancestry and size do not explain the replication consistently. We also showed that confounding is highly prevalent in GWAS studies. This indicates that modeling residual confounding is necessary for understanding lack of replication in GWAS studies.

One of the difficulties in our analyses is that some GWAS studies have very few significant variants, making the maximum likelihood estimates of the variance parameters unstable. Theoretically, it is possible to compute the variance parameters using additional variants that were not significant in the initial discovery study. However, summary statistics for these variants are often not computed to decrease the multiple testing burden for replication studies. Nevertheless, as GWAS studies have increasingly larger sample sizes, we expect that the number of GWAS variants will increase and make our estimated parameters increasingly robust.

Methods

GWAS overview

In GWAS studies, an association study is performed between each genetic variant and the phenotype. The effect size of each variant (k) is determined by estimating the maximum likelihood parameters of Equation 5, where y_j is the phenotype or individual j, μ is the phenotypic mean, x_{kj} is the normalized genotype of individual j, β_k is the effect size of the variant k, e_j is the error, and N is the number of individuals.

$$y_j = \mu + \beta_k x_{ij} + e_j \tag{5}$$

In vector notation, Equation 5 becomes the following.

$$y = \mu\mathbf{1} + \beta_k X_k + \mathbf{e} \tag{6}$$

The resulting maximum likelihood estimates are $\hat{\mu} = \frac{1}{N}\mathbf{1}^T y$ and $\hat{\beta_k} = \frac{X_k^T y}{N}$. The residuals $\hat{\mathbf{e}} = y - \hat{\mu}\mathbf{1} - \hat{\beta}_k X_k$ can be used to estimate the standard error $\hat{\sigma_e} = \sqrt{\frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{N-2}}$. The standard error of the estimator is $\sigma_{\hat{\beta_k}} = \frac{\hat{\sigma_e}}{\sqrt{N}}$. Since the sample sizes for GWAS studies are large, the association statistic $s_k = \frac{\hat{\beta_k}}{\hat{\sigma_e}}\sqrt{N}$ follows an approximately normal distribution (Equation 8).

$$s_k \sim \mathcal{N}\left(\frac{\beta_k}{\sigma_e}\sqrt{N}, 1\right) \tag{7}$$

Under the null hypothesis, S_k will follow the standard normal distribution, which can be used to compute the significance of association. In the standard GWAS framework, we assume that the standardized effect size is caused by a true genetic effect $\lambda = \frac{\beta_k}{\sigma_{\sigma_e}}$. Thus, Equation 8 can be rewritten as the following.

$$s_k|\lambda \sim \mathcal{N}\left(\lambda\sqrt{N}, 1\right) \tag{8}$$

Correcting GWAS summary statistics for winner's curse

Given Equation 8, we can write the distributions of summary statistics for a initial discovery study and a replication study as $s_k^{(1)}|\lambda \sim \mathcal{N}\left(\lambda\sqrt{N_1}, 1\right)$ and $s_k^{(2)}|\lambda \sim \mathcal{N}\left(\lambda\sqrt{N_2}, 1\right)$, respectively.

We assume that λ is the same across multiple studies on the same trait. We define the prior distribution of λ as $\lambda \sim \mathcal{N}(0, \sigma_g^2)$, where σ_g^2 is the variance in the true effect size. Thus, the posterior distributions of $s_k^{(1)}$ and $s_k^{(2)}$ are also normally distributed.

$$s_k^{(1)} \sim \mathcal{N}(0, N_1\sigma_g^2 + 1) \tag{9}$$

$$s_k^{(2)} \sim \mathcal{N}(0, N_2\sigma_g^2 + 1) \tag{10}$$

We correct for winner's curse by computing the conditional distribution of the replication statistic ($s_k^{(2)}$) given the discovery statistic ($s_k^{(1)}$). We derive the conditional distribution from the joint distribution as follows.

The covariance between $s_k^{(1)}$ and $s_k^{(2)}$ is computed as follows.

$$\begin{aligned} cov(s_k^{(1)}, s_k^{(2)}) &= \mathbb{E}\left[(\lambda\sqrt{N_1} - \mathbb{E}(\lambda\sqrt{N_1}))(\lambda\sqrt{N_2} - \mathbb{E}(\lambda\sqrt{N_2}))\right] \\ &= \mathbb{E}\left[\lambda^2\sqrt{N_1N_2}\right] \\ &= \sqrt{N_1N_2\sigma_g^2} \end{aligned}$$

Therefore, the joint distribution of $s_k^{(1)}$ and $s_k^{(2)}$ is Equation 11.

$$\begin{pmatrix} s_k^{(1)} \\ s_k^{(2)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} N_1\sigma_g^2 + 1 & \sqrt{N_1N_2}\sigma_g^2 \\ \sqrt{N_1N_2}\sigma_g^2 & N_2\sigma_g^2 + 1 \end{pmatrix}\right), \tag{11}$$

Conditioning on $s_k^{(1)}$, we obtain Equation 12.

$$(s_k^{(2)}|s_k^{(1)} = x) \sim \mathcal{N}\left(\frac{\sqrt{N_1N_2}\sigma_g^2}{N_1\sigma_g^2 + 1}x, 1N_2\sigma_g^2 - +\frac{N_2\sigma_g^2}{N_1\sigma_g^2 + 1}\right) \tag{12}$$

For each value of $s_k^{(1)}$, the mean of the conditional distribution gives the expected summary statistic in a replication study, correcting for winner's curse. This distribution can also be used to create a confidence interval on the replication sample statistics.

Correcting GWAS summary statistics for winner's curse and confounding

Suppose in addition to study-specific environmental effects, there are also study-specific confounders. We model these confounders in the discovery study and replication study as $\delta^{(1)} \sim \mathcal{N}(0, \sigma_{c_1}^2)$ and $\delta^{(2)} \sim \mathcal{N}(0, \sigma_{c_2}^2)$ respectively. We decompose the effect size into the sum of a genetic component (λ) and a confounding component $\delta^{(i)}$.

$$s_k^{(1)}|\lambda \sim \mathcal{N}\left((\lambda + \delta^{(1)})\sqrt{N_1}, 1\right) \tag{13}$$

$$s_k^{(2)}|\lambda \sim \mathcal{N}\left((\lambda + \delta^{(2)})\sqrt{N_2}, 1\right) \tag{14}$$

Similar to the case without confounding, the posterior distributions of $s_k^{(1)}$ and $s_k^{(2)}$ are normally distributed (Equations 15 and 16).

$$s_k^{(1)} \sim \mathcal{N}(0, N_1\sigma_g^2 + N_1\sigma_{c_1}^2 + 1) \tag{15}$$

$$s_k^{(2)} \sim \mathcal{N}(0, N_2\sigma_g^2 + N_2\sigma_{c_2}^2 + 1) \tag{16}$$

Therefore, the joint distribution is Equation 17

$$\begin{pmatrix} s_k^{(1)} \\ s_k^{(2)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} N_1\sigma_g^2 + N_1\sigma_{c_1}^2 + 1 & \sqrt{N_1N_2}\sigma_g^2 \\ \sqrt{N_1N_2}\sigma_g^2 & N_2\sigma_g^2 + N_2\sigma_{c_2}^2 + 1 \end{pmatrix}\right), \tag{17}$$

Similar to the winner's curse only model, we can find the expected summary statistic in a replication study correcting for winner's curse by computing the conditional distribution of the replication statistic ($s_k^{(2)}$) given the discovery statistic ($s_k^{(1)}$) (Equation 18).

$$(s_k^{(2)}|s_k^{(1)} = x) \sim \mathcal{N}\left(\frac{\sqrt{N_1N_2}\sigma_g^2}{N_1\sigma_g^2 + N_1\sigma_{c_1}^2 + 1}x, N_2\sigma_g^2 + N_2\sigma_{c_2}^2 + 1 - \frac{N_1N_2\sigma_g^4}{N_1\sigma_g^2 + N_1\sigma_{c_1}^2 + 1}\right) \tag{18}$$

Predicting the replication rate

The conditional distribution ($s_k^{(2)}|s_k^{(1)}$) can also be used to predict the replication rate of an initial discovery study. For a genetic variant k with association statistic $s_k^{(1)} = x$ in an initial discovery study, the probability of replication is $Pr\left(abs(s_k^{(2)}) > z|s_k^{(1)} = x\right)$, where z is the z-score thresholds corresponding to a specific significance threshold t for the replication study.

The predicted replication rate can be calculated as the average probability of replication across all significant variants in the discovery study. Let $\mathcal{A}$ be the set of variants found to be significant in the discovery study. The predicted replication rate (r) is defined as

$$r = \frac{1}{|\mathcal{A}|}\sum_{s_k \in \mathcal{A}} P\left(abs(s_k^{(2)}) > z|s_k^{(1)} = x\right) \tag{19}$$

Estimating the variance components from data

The genetic variance (σ_g^2) and confounding variance ($\sigma_{c_1}^2$, $\sigma_{c_2}^2$) are not known a priori. We estimate these parameters from the data using the following procedures. Since in many cases, the replication study only tests variants that are significant in the initial study, we calculate the variance components using only data from variants that are significant in the initial study and tested in both studies. Let the total number of variants studied be M and the total number of significant variants in the initial study be M'.

We first calculate the maximum likelihood estimate (MLE) for the total variance of the statistics in the discovery study $s^{(1)}$, which we denote as $\hat{\sigma}^2_{s^{(1)}}$ (Equation 20). We include the unobserved variants that are not significant in the first study in the likelihood by integrating over all of their possible values.

$$\underset{\sigma^2_{s^{(1)}}}{arg\,max}\left(\sum_{i=0}^{M'}(log\left[P(s_i^{(1)}|0,\sigma^2_{s^{(1)}}\right]\right)+\left(\sum_{i=0}^{M-M'}log\left[P(-z\leq s_i^{(1)}\leq z)|0,\sigma^2_{s^{(1)}}\right]\right) \tag{20}$$

We then use this estimate of the total variance to compute the expected value of the replication statistics $s^{(2)}$ for different values of σ_g^2. We select the value of σ_g^2 that minimizes the residual sum of squares between the predicted value of $s^{(2)}$ and the true value (Equation 21).

$$\underset{\sigma_g^2}{arg\,min}\sqrt{\sum_{i}^{M'}\left(\frac{\sigma_g^2}{\sigma^2_{s^{(1)}}}s_i^{(1)}-s_i^{(2)}\right)^2} \tag{21}$$

We can then decompose total variance in $s^{(1)}$ and estimated $\sigma_{c_1}^2$ using the previously estimated total variance ($\hat{\sigma}^2_{s^{(1)}}$) and genetic variance ($\hat{\sigma}_g^2$). We solve for $\hat{\sigma}_{c_1}$ as follows.

$$\hat{\sigma}^2_{s^{(1)}} = 1 + N_1\sigma_g^2 + N_1\sigma_{c_1}^2$$
$$\hat{\sigma}^2_{c_1} = \frac{\hat{\sigma}^2_{s^{(1)}} - N_1\hat{\sigma}_g^2 - 1}{N_1}$$

Finally, we use the joint distribution of $s_k^{(1)}$ and $s_k^{(2)}$ (Equation 17) to compute the MLE estimate of $\sigma_{c_2}^2$, using the previously estimated $\hat{\sigma}_g^2$ and $\hat{\sigma}_{c_1}$.

Data generating model

We generated simulated data to demonstrate that our approach can capture the effects of winners' curse and confounding to explain low replication in GWAS studies. To show that our model is more effective at explaining low replication than a method that only takes into account winners' curse, we directly compare our method with a simplified model that only takes into account winners' curse.

We set variance of the shared genetic variance to be $\sigma_g^2 = 1$. We then set the variance of study-specific confounders to be $\sigma_{c_1}^2 = 1$ and $\sigma_{c_2}^2 = 3$. For each variant, we sampled from the following distributions.

$$\begin{aligned}
\lambda &\sim N(0, \sigma_g^2) \\
\delta^{(1)} &\sim N(0, \sigma_{c_1}^2) \\
\delta^{(2)} &\sim N(0, \sigma_{c_2}^2) \\
\epsilon^{(1)} &\sim N(0, 1) \\
\epsilon^{(2)} &\sim N(0, 1)
\end{aligned}$$

We assumed that sample sizes for the discovery and replication studies were 5000 and 1000, respectively. We computed the summary statistics for each study as $s_k^{(i)} = \sqrt{N_i}(\lambda + \delta^{(i)}) + \epsilon^{(i)}$. Using our framework, we estimate the variance components and compare these estimates with the true values.

Acknowledgment

J.Z., J.Z, S.F., R.B., and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589, 1331176, and 1815624, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. J.Z. is supported by National Institutes of Health Award Number T32MH073526 and National Science Foundation Graduate Research Fellowship under Grant DGE-1650604.

References

[1] John P A Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8), 2005.

[2] John P A Ioannidis. Why Most Clinical Research Is Not Useful. *PLoS Medicine*, 13(6):1–10, 2016.

[3] Prasad Patil, et al. Test set bias affects reproducibility of gene signatures. *Bioinformatics*, 31(March):2318–2323, 2015.

[4] Cameron Palmer and Itsik Pe. Statistical correction of the Winner ' s Curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genetics*, pages 1–18, 2017.

[5] Urko M Marigorta, et al. Replicability and Prediction : Lessons and Challenges from GWAS. *Trends in Genetics*, 34(7):504–517, 2018.

[6] Danielle Welter, et al. The NHGRI GWAS Catalog , a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(December 2013):1001–1006, 2014.

[7] Casey S Greene, et al. Failure to Replicate a Genetic Association May Provide Important Clues About Genetic Architecture. *PLoS ONE*, 4(6), 2009.

[8] Kerry Dwan, et al. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias An Updated Review. *PLoS ONE*, 8(7), 2013.

[9] R. Prentice H. Zhong. Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genet Epidemiol*, 34(1):78–91, 2010.

[10] Lei Sun, et al. BR-squared : a practical solution to the winner ' s curse in genome-wide scans. *Hum Genet*, 129(5):545–552, 2011.

[11] M. Boehnke R. Xiao. Quantifying and correcting for the winner's curse in quantitative trait association studies. *Genet Epidemiol.*, 35(3):133–138, 2012.

[12] M. Boehnke R. Xiao. Quantifying and correcting for the winner's curse in genetic association studies. *Genet Epidemiol.*, 33(5):453–462, 2010.

[13] Hyun Min Kang, et al. Accurate Discovery of Expression Quantitative Trait Loci Under Confounding From Spurious and Genuine Regulatory Hotspots. *Genetics*, 1925(December):1909–1925, 2008.

[14] Alkes L Price, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[15] Jong Wha J Joo, et al. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. *GenomeBiology*, 15:1–15, 2014.

[16] B Devlin and Kathryn Roeder. Genomic Control for Association Studies. *Biometrics*, 55(December):997–1004, 1999.

[17] Brendan K Bulik-sullivan, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015.

[18] Ronald De Vlaming, et al. Equivalence of LD-Score Regression and Individual-Level-Data Methods. *bioRxiv*, 2017.

[19] Results The, et al. Articles Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11), 2014.

[20] Cara L Carty, et al. Genome-wide association study of body height in African Americans : the Women ' s Health Initiative SNP Health Association Resource (SHARe). *Human Molecular Genetics*, 21(3), 2012.

[21] Ross L Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9(4):621–634, 2008.

[22] Sophie I Candille, et al. Genome-Wide Association Studies of Quantitatively Measured Skin , Hair , and Eye Pigmentation in Four European Populations. *PLoS ONE*, 7(10), 2012.

[23] Mariaelisa Graff, et al. Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. *Human Molecular Genetics*, 22(17), 2013.

[24] CHARGE and Global BPgen consortia. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nature Genetics*, 43(10):1005–1012, 2011.

[25] Global Urate, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics*, 45(2), 2013.

[26] Fourteen Bmd-associated, et al. Articles density loci and reveals 14 loci associated with risk of fracture. *Nature Genetics*, 44(5), 2012.

[27] Amidou N Diaye, et al. Identification , Replication , and Fine-Mapping of Loci Associated with Adult Height in Individuals of African Ancestry. *PLoS Genetics*, 7(10), 2011.

[28] Jonathan Flint and Eleazar Eskin. Genome-wide association studies in mice. *Nature Reviews Genetics*, 13(11):807–817, 2012.

[29] Clarissa C Parker, et al. Genome-wide association study of behavioral , physiological and gene expression traits in outbred CFW mice. *Nat Genet*, 48(8), 2016.

[30] Jérôme Nicod, et al. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. *Nat Genet*, 48(8), 2016.