Article

Not peer-reviewed version

Integrating External Controls by Regression Calibration for Genome-Wide Association Study

Lirong Zhu , Shijia Yan , Xuewei Cao , Shuanglin Zhang , Qiuying Sha *

Posted Date: 15 December 2023

doi: 10.20944/preprints202312.1184.v1

Keywords: Genome-wide association test; case-control study; batch effect; data integration

Article

Integrating External Controls by Regression Calibration for Genome-Wide Association Study

Lirong Zhu, Shijia Yan, Xuewei Cao, Shuanglin Zhang and Qiuying Sha *

Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, USA

* Correspondence: Qiuying Sha, Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan 49931, USA. E-mail: qsha@mtu.edu.

Abstract: Genome-wide association studies (GWAS) have successfully revealed many disease-associated genetic variants. For a case-control study, the adequate power of an association test can be achieved with a large sample size, although genotyping large samples is expensive. A cost-effective strategy to boost power is to integrate external control samples with publicly available genotyped data. However, the naïve integration of external controls may inflate the type I error rates if ignoring the systematic differences (batch effect) between studies, such as the differences in sequencing platforms, genotype calling procedures, population stratification, and so forth. To account for the batch effect, we propose an approach by integrating External Controls into the Association Test by Regression Calibration (iECAT-RC) in case-control association studies. Extensive simulation studies show that iECAT-RC not only can control type I error rates but also can boost statistical power in all models. We also apply iECAT-RC to the UK Biobank data for M72 Fibroblastic disorders by considering genotype calling as the batch effect. Four SNPs associated with Fibroblastic disorders have been detected by iECAT-RC and the other two comparison methods. However, our method has a higher probability of identifying these significant SNPs in the scenario of an unbalanced case-control association study.

Keywords: genome-wide association test; case-control study; batch effect; data integration

Introduction

Genome-wide association studies (GWAS) play a major role in associating specific genetic variants with common diseases and complex traits [1–3]. Sometimes, researchers may have limited access to individuals' genetic information with specific traits and large-scale genetic studies can be expensive and resource-intensive [4]. Thus, with a small sample size in GWAS, the association test could have low power and may also increase the possibility of false-positive findings, especially for infrequent variants (i.e., MAF < 5%) [5,6].

The rapid development of sequencing technologies has promoted substantial advancement in GWAS, particularly in obtaining comprehensive genetic information from limited samples [7,8]. The integration of sequenced samples provides a great opportunity for identifying novel genetic associations and increasing the statistical power of single-variant association tests [9]. Nevertheless, the challenges associated with integrating sequenced samples arise from various factors, such as the utilization of diverse sequencing platforms, variations in genotype calling procedures, the presence of population stratification, and so forth [10]. In a single study, by incorporating sequenced samples from other studies as an external control sample, the power of single-variant tests can be significantly increased without incurring additional sequencing costs. However, the systematic differences (batch effect) between studies could inflate the type I error rates and increase the possibility of false-positive findings in association studies [11].

Several methods have been proposed recently to address the systematic differences between genotyped data of internal and external sources using likelihood-based methods [12]. Liu and Leal proposed a method SEQCHIP to correct bias for integrating genotype data in rare variant association studies [13]. Derkach et al. proposed another method that substitutes the genotype calls by the expected values given observed sequence data to account for differential read depths between studies [14]. Motivated by Derkach et al., Chen and Lin proposed regression calibration (RC) methods to account for differential sequencing errors between cases and controls [15]. Although these methods are powerful, computing genotype probabilities and storing sequence reads data can be challenging and expensive for large-scale studies. Thus, ProxECAT incorporates external controls to estimate enrichment of rare variants using allele counts in case-control analysis [16]. However, nonuse of the internal control samples potentially limits the power of the association test. iECAT allowed the incorporation of external controls in single variant association tests [11]. And the batch effect between internal and external studies can be assessed by comparing odds ratio estimates of alleles using internal control samples and combined control samples from internal and external studies. Then an empirical Bayesian-type shrinkage estimator is constructed based on the degrees of odds ratios in the single-variant test. And it is demonstrated that this method can control type I error rates, as well as improve the power of the association test. However, this method cannot adjust for covariates such as age, gender, and so on [11]. Based on the aforementioned method, Li and Lee proposed a novel score based test, which constructs a shrinkage score statistic using exclusively internal samples and external control samples, allowing for covariate adjustment [17]. However, the power increase of this method in association testing by integrating external controls is limited for extremely unbalanced case-control studies.

In this study, we present a novel approach that integrates External Controls into Association Tests by Regression Calibration (iECAT-RC) to incorporate external control samples in case-control association studies. The objective of this research is to boost the statistical power of the single-variant association test by integrating external controls with the adjustment of batch effects. We propose an approach that adjusts the genotypes of an external control sample to approximate the same distribution as the genotypes in the internal control sample through regression calibration. Furthermore, we apply the Saddlepoint approximation [18] and efficient resampling [19] methods to control type I error rates with imbalanced case-control and low minor allele count (MAC) scenarios, respectively.

Materials and Methods

Consider a phenotype with case and control states. We code a case as 0 and control as 1. Assume that the internal study has the sample size n^I with n_0^I controls and n_1^I cases and $n_0^I + n_1^I = n^I$; the external study has n_0^E controls. For the i^{th} subject, let $y_i = 0/1$ be the dichotomous phenotype. Denote $G_1, G_2, ..., G_{n_0^I}, G_{n_0^I+1}, G_{n_0^I+2}, ..., G_{n^I}$, and $g_1, g_2, ..., g_{n_0^E}$ as the genotypes of the internal control sample, the internal case sample, and the external control sample at a genetic variant, respectively, with indicating the number of copies of the minor allele carried by the subject at that genetic variant. We denote $\mathbf{X}_i^I$ be the first p principal components of internal genotypes, and $\mathbf{X}_i^E$ be the first p principal components of external genotypes for the i^{th} subject.

Motivated by the novel method iECAT-Score [20], we propose a new method by integrating external controls into association tests to boost the statistical power. Our proposed method involves three steps. Step 1. adjusting the genotypes of external controls using regression calibration; Step 2. conducting single-variant association test; and Step 3. calibrating single-variant test using Saddlepoint approximation (SPA) [18] and efficient resampling (ER) methods [19], particular addressing scenarios of case-control imbalance and low minor allele count (MAC). By following these three steps, the iECAT-RC method effectively minimizes the impact of batch effects and improves the power of association testing.

Step 1. adjusting the genotypes of external controls by regression calibration

To adjust the genotype of external control samples for the batch effect, we propose to use the following procedure:

1). Without loss generality, we assume $n_0^E \geq n_0^I$. We randomly choose n_0^I individuals with genotypes $g_{k1},\ldots,g_{kn_0^I}$ from external control samples.

2). We assume a linear regression model $G_i=\beta_0^{(k)}+\beta_1^{(k)}g_{ki}+\boldsymbol{\alpha}_I^{(k)}\mathbf{X}_i^I+\boldsymbol{\alpha}_E^{(k)}\mathbf{X}_{ki}^E$ for $i=1,\ldots,n_0^I$, where $\hat{\boldsymbol{\beta}}^{(k)}=(\hat{\beta}_0^{(k)},\hat{\beta}_1^{(k)},\hat{\boldsymbol{\alpha}}_I^{(k)},\hat{\boldsymbol{\alpha}}_E^{(k)})^T$ is the least square estimate of $\boldsymbol{\beta}^{(k)}=(\beta_0^{(k)},\beta_1^{(k)},\boldsymbol{\alpha}_I^{(k)},\boldsymbol{\alpha}_E^{(k)})^T$.

3). We repeat 1) and 2) K times. We obtain $\hat{\boldsymbol{\beta}}^{(1)},\ldots,\hat{\boldsymbol{\beta}}^{(K)}$ and calculate the average value $\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1,\hat{\boldsymbol{\alpha}}_I,\hat{\boldsymbol{\alpha}}_E)^T=\sum_{k=1}^{K}\hat{\boldsymbol{\beta}}^{(k)}/K$. Let $G_{n^I+i}=\hat{\beta}_0+\hat{\beta}_1g_i+\hat{\boldsymbol{\alpha}}_I\mathbf{X}_i^I+\hat{\boldsymbol{\alpha}}_E\mathbf{X}_i^E$ for $i=1,\ldots,n_0^I$. When $G_{n^I+i}<a_0$, we let G_{n^I+i} take 0, where a_0 is determined such that the frequency of 0 in the internal control genotypes equals to the frequency of 0 in G_{n^I+i} for $i=1,\ldots,n_0^I$. When $a_0\le G_{n^I+i}<a_1$, we let G_{n^I+i} take 1, where a_1 is determined such that the frequency of 1 in the internal control genotypes equals to the frequency of 1 in G_{n^I+i} for $i=1,\ldots,n_0^I$. When $G_{n^I+i}>a_1$, we let G_{n^I+i} take 2.

We repeat the above procedure till we obtain G_{n^I+i} for $i=1,\ldots,n_0^E$. Then we perform the association test based on the internal case-control data and external control data with genotypes $G_1,G_2,\ldots,G_{n_0^I},G_{n_0^I+1},G_{n_0^I+2},\ldots,G_{n^I},G_{n^I+1},\ldots,G_{n^I+n_0^E}$.

Step 2. Single-variant association test

We combine the internal samples and external control samples with the adjusted genotypes. $\mathbf{G}=(G_1,G_2,\ldots,G_n)^T$ is the vector of genotypes at a variant for n subjects, where $n=n^I+n^E$. Assume that there are p covariates, then we relate the phenotype Y_i to the covariate $\mathbf{Z}_i$, and genotype G_i using the logistic regression model $\text{logit}[\text{P}(Y_i=1\,|\,\mathbf{Z}_i,G_i)]=\mathbf{Z}_i^T\boldsymbol{\alpha}+G_i\beta$, where the phenotype Y_i follows a Bernoulli distribution. In this equation, $\boldsymbol{\alpha}$ is a $p\times1$ vector of coefficients for p covariates including the intercept, and β is the genotype effect at the variant. Assessing whether the association exists between the phenotype and the genotype at a variant is equivalent to testing $H_0:\beta=0$.

Let $\boldsymbol{\mu}=\{\mu_i\}=\{P(Y_i=1\,|\,\mathbf{Z}_i)\}$ and $\hat{\mu}_i$ be the maximum-likelihood estimate of μ_i under H_0. In the score test, the score is given by $S=\tilde{\mathbf{G}}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})$. where $\mathbf{Y}=(Y_1,\ldots,Y_n)^T$, $\tilde{\mathbf{G}}=\{\tilde{G}_i\}=\mathbf{G}-\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{V}\mathbf{G}$ is the covariate adjusted genotype vector and $\mathbf{V}=diag\{\hat{\mu}_i(1-\hat{\mu}_i)\}$ [2]. Under the null hypothesis of no genetic effect, $E(S)=0$ and $Var(S)=\sum_{i=1}^{n}\tilde{G}_i^2\hat{\mu}_i(1-\hat{\mu}_i)$. Then the score test statistic $T_{Score}=S^2/Var(S)$ asymptotically follows the chi-square distribution with 1 degree of freedom, and the p-value can be obtained as $p=P(\chi_1^2>S^2/Var(S))$.

Step 3. Calibrating single-variant test using SPA and ER methods

The single-variant score test approximates the null hypothesis by normal distribution. The variance estimates based on such asymptotic test behaves well for common variants and balanced case-control studies. When allele frequency is extremely low resulting from low MAC, or when the case-control ratio is unbalanced, the underlying distribution of test statistic could be highly skewed. In such cases, the traditional asymptotic-based score test performs poorly with conservative or anticonservative results [21,22].

To account for scenarios of unbalanced case-control ratio, we apply the SPA method to obtain the p-value when the score estimates lie far from mean zero [18]. When the MAC is low ($MAC < 10$) either in the internal sample, combining sample, or external sample, we apply the ER method to obtain the p-values [19].

1). SPA method

SPA is an improvement over the normal approximation which only uses the mean and variance to approximate the underlying distribution. SPA uses the entire cumulant-generating function (CGF). Given the score test statistic $S = \sum_{i=1}^{n} \hat{G}_i (Y_i - \hat{\mu}_i)$, the estimation of the CGF of S is $K(t) = \log(E_{H_0}(e^{ts})) = \sum_{i=1}^{n} \log(1 - \hat{\mu}_i + \hat{\mu}_i e^{\hat{G}_i t}) - t \sum_{i=1}^{n} \hat{G}_i \hat{\mu}_i$. According to the SPA method, the distribution of S can be estimated by

$$\Pr(S < s) \approx \tilde{F}(s) = \Phi\left\{ w + \frac{1}{w} \log(\frac{v}{w}) \right\},$$

where $w = \operatorname{sgn}(\hat{t})\sqrt{2(\hat{t}s - K(\hat{t}))}$, $v = \hat{t}\sqrt{K''(\hat{t})}$, $K'(t)$ and $K''(t)$ are the estimations of the first- and second-order derivatives of K , $\hat{t}$ is the solution to the equation $K'(\hat{t}) = s$, and Φ is the distribution of a standard normal distribution. The p-value can be obtained using the R package SPAtest.

1). ER method

ER method is used for rare variant association test with binary trait. Given phenotypes $\mathbf{Y}$, genotypes $\mathbf{G}$, and covariates $\mathbf{Z}$, the p-value of ER method is defined as

$$\Pr(Q \geq \hat{Q} \mid Y, G, X) = \sum_{d=0}^{m} \Pr(Q \geq \hat{Q} \mid D = d, \mathbf{Y}, \mathbf{G}, \mathbf{Z}) \Pr(D = d \mid \mathbf{Y}, \mathbf{G}, \mathbf{Z})$$

where $\hat{Q}$ is the score test statistic from the original phenotype, m is the number of individuals with minor alleles, and D is the number of cases among m individuals carrying a minor allele. The p-value can be obtained using the R package SKAT.

Simulations

In order to evaluate the performance of the proposed method iECAT-RC related to the type I error rates and power, we carry out simulation studies under a series of scenarios. We generate the binary phenotypes with cases and controls from a logistic regression model: $\operatorname{logit}[P(Y = 1 \mid \mathbf{Z}, G)] = \alpha_0 + 0.5Z_1 + 0.5Z_2 + \beta G + \varepsilon$, where Z_1 is a continuous covariate generated from the standard normal distribution; Z_2 is a binary covariate taking values 0 and 1 with the probability of 0.5 ; α_0 is chosen such that the disease prevalence is 0.05 ; G is the genotype at a variant generated from a binomial distribution $BIN(2, MAF)$; β is the effect size of the variant; and ε follows a standard normal distribution. MAF is sampled from the empirical Mini-Exome genotype data provided by the GAW17, which includes $24,487$ variants in 3205 genes introduced in Sha et al [2].

To mimic the batch effect between internal and external control studies, we first define the differential variant size (DVS), that is the proportion of the variants subject to different MAFs between the internal and external control samples. For such variants, we set the MAFs of the external controls to be randomly generated based on the following two scenarios to mimic the level of batch effect: (1) $\text{Uniform}(0.1q, 4q)$ and (2) $2q$, where q is the MAF of the corresponding variants in the internal sample. Subsequently, we consider different numbers of cases and controls in the internal sample and the number of controls in the external controls. We set the following three ratios between the internal cases, internal controls and external controls $(n_1^I : n_0^I : n_0^E)$: (1) $5000:5000:10000$, (2) $6667:3333:10000$, and (3) $500:5000:10000$. Thus, we consider a total of six models. Model 1: the ratio $(n_1^I : n_0^I : n_0^E)$ is $5000:5000:10000$ and MAF of the external sample is from $2q$; Model 2: the ratio is $6667:3333:10000$ and MAF of external sample is from $2q$; Model 3: the ratio is $500:5000:10000$ and MAF of external sample is from $2q$; Model 4: the ratio is $5000:5000:10000$ and MAF of external sample is from $\text{Uniform}(0.1q, 4q)$; Model 5: the ratio is $6667:3333:10000$ and MAF of external sample is from $\text{Uniform}(0.1q, 4q)$; and Model 6: the ratio is $500:5000:10000$ and MAF of external sample is from $\text{Uniform}(0.1q, 4q)$.

We compare our proposed method, iECAT-RC, with other three approaches for a single-variant association test: iECAT-N based on the naïve integrating the internal and external control samples; Internal using only the internal sample; and iECAT-Score proposed by Li and Lee [20]. If the case-control ratio of the combined sample is unbalanced or MAC is low (< 10 is used in the simulation studies), iECAT-RC, iECAT-N, and Internal use SPA or ER to obtain the corresponding p-values, respectively.

To evaluate type I error rates, phenotypes are generated with $\beta = 0$. For each simulation, we generate 5×10^5 data sets and use different significance levels 0.05, 0.01, 10^{-3}, and 10^{-4} for single-variant tests. To save computation time, we generate 5×10^3 genotypes, then resample the disease phenotypes of internal samples 100 times for each set, while keeping other data fixed in the type I error rate evaluation.

To evaluate power, the effect size β in Model 3 and Model 6 is set to be $\log(2), \log(2.4), \log(2.8), \text{and } \log(3.2)$. The effect size β for other models is set to be $\log(1.6), \log(1.8), \log(2.0), \text{and } \log(2.2)$. We generate 5×10^3 data sets for each model to evaluate the empirical power at the significance level of 5×10^{-8}.

Result

Type I error rates

To evaluate the Type I error rates, we simulate 5×10^5 data sets under the null hypothesis of no association. **Table 1** and **Table S1** provide a summary of the type I error rates of the four methods, iECAT-RC, iECAT-N, Internal, and iECAT-Score, at different significance levels under $DVS = 0.03, \text{ and } 0.5$, respectively. From these two tables, we can see that iECAT-RC, Internal, and iECAT-Score control Type I error rates very well. However, the Type I error rates of iECAT-N are significantly inflated when the internal samples and external control samples are naively integrated without adjusting the batch effect. For instance, as shown in **Table 1**, the empirical Type I error rates of iECAT-N exceed the nominal significance level $\alpha = 10^{-4}$ by approximately 1000-fold when the internal and external samples are combined naively. Furthermore, we examine scenarios when the case, control, and external control ratio remains the same but the batch effect levels differ (Model 1 and Model 4). The performance of the four methods under Model 4 is consistent with those in Model 1. Under both models, the results show well-controlled Type I error rates across all methods except iECAT-N. Additionally, we consider scenarios with varying case, control, and external control ratio but the same batch effect level (Model 1-3). In these cases, iECAT-RC effectively controls the Type I error rates, even under extremely unbalanced case-control samples.

Table 1. Empirical Type I error rates of iECAT-RC, compared with other three methods iECAT-N, Internal, and iECAT-Score when DVS is 0.03 at different significance levels, 0.05, 0.01, 10^{-3}, and 10^{-4}.

Model	Significance level	iECAT-RC	iECAT-N	Internal	iECAT-Score
Model 1	0.05	0.0382	**0.3956**	0.0512	0.0482
	0.01	0.0057	**0.3352**	0.0102	0.0096
	0.001	3.00E-04	**0.2771**	0.001	0.001
	1E-04	1.00E-04	**0.2429**	1.00E-04	0
Model 2	0.05	0.0397	**0.4163**	0.0348	0.0394
	0.01	0.0078	**0.3685**	0.0087	0.0089
	0.001	9.00E-04	**0.3263**	4.00E-04	0.0013
	1E-04	1.00E-04	**0.2919**	0	2.00E-04
Model 3	0.05	0.0457	**0.113**	0.0136	0.0357
	0.01	0.0111	**0.0628**	0.004	0.0081
	0.001	6.00E-04	**0.0345**	5.00E-04	3.00E-04
	1E-04	0	**0.0223**	0	0
Model 4	0.05	0.0372	**0.4269**	0.0511	0.0475
	0.01	0.0065	**0.3513**	0.0105	0.0101
	0.001	4.00E-04	**0.2804**	9.00E-04	0.001
	1E-04	0	**0.2359**	3.00E-04	1.00E-04
Model 5	0.05	0.0494	**0.457**	0.0335	0.0446
	0.01	0.0107	**0.3876**	0.0079	0.0096
	0.001	0.0017	**0.3244**	9.00E-04	0.001
	1E-04	4.00E-04	**0.2806**	0	1.00E-04
Model 6	0.05	0.0467	**0.1013**	0.0133	0.0342
	0.01	0.011	**0.0569**	0.0042	0.007
	0.001	0.0012	**0.0291**	9.00E-04	7.00E-04
	1E-04	1.00E-04	**0.0169**	0	0

Note: The bold-faced values indicate the type I error rates beyond the upbound of the corresponding 95% confidence interval.

Power

To evaluate the performance of our proposed method, we consider different batch effect levels, different values of DVS, and different values of $n_1^I : n_0^I : n_0^E$. We compare the power of the three methods iECAT-RC, Internal, and iECAT-Score at an empirical significance level 5×10^{-8}. iECAT-N is ignored in the power comparison since this method inflates type I error rates. **Figure 1** shows the power comparison of these three tests (iECAT-RC, Internal, and iECAT-Score) for different values of $n_1^I : n_0^I : n_0^E$ when DVS is 0.03. As shown in the figure, in the case of both balance (Model 1 and model 4) and slightly unbalanced (Model 2 and Model 5) case control ratio in the internal samples, iECAT-RC is more powerful than the other two tests; Internal is the least powerful one due to a smaller sample size compared with other two methods. For the extremely unbalanced internal case-control ratio (Model 3 and Model 6), these three methods have a similar power performance. This is reasonable because there is slight inflation in the p-value for the extremely unbalanced case-control ratio after calibrating the score test by SPA [18].

Power comparison of the three tests for DVS = 0.5 is showed in Figure S1. The power patterns of the three methods are very similar between these two different DVS settings for Models 1, 2, 4, and 5. iECAT-RC is more powerful than the other two methods, iECAT is the second powerful method, and Internal is the least powerful method. For models 3 and 6, similar to the pattern for DVS = 0.03, iECAT-RC and Internal have similar power, but iECAT-Score has lower power than iECAT-RC and Internal.

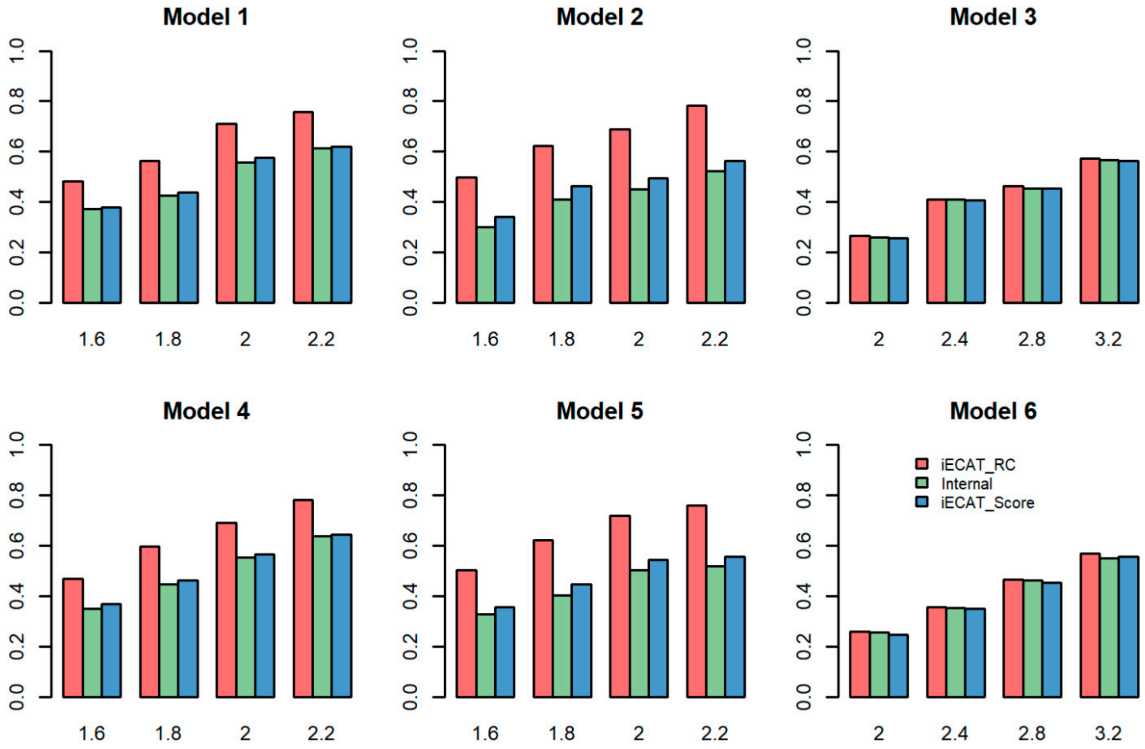

Figure 1. Power comparison of iECAT-RC, Internal, and iECAT-Score at the significance level 5×10^{-8} and $DVS = 0.03$. The horizontal axis represents the odds ratio, and the vertical axis represents power.

Application to the UK Biobank Data

The UK Biobank dataset, which contains approximately $500,000$ individuals with $784,256$ variants from across the United Kingdom, provides a prospective cohort study for discovering more genetic associations and the genetic bases of complex traits with deep genetic and phenotypic data [23–25]. In the UK Biobank dataset, genotypes are assayed using two genotyping calling procedures which are the Applied Biosystems UK BiLEVE Axiom Array (UKBL) and the UK Biobank Axiom Array (UKBB) [26,27]. However, the common practice of calling underlying genotypes and then treating the called values as known is prone to false positive findings, especially when genotyping errors are systematically different between cases and controls [28]. Therefore, we apply our proposed method to the real data from the UK Biobank based on two genotype calling procedures and consider genotype calling as the batch effect. The genotype quality control is performed by PLINK 1.9 https://www.cog-genomics.org/plink/1.9/ with missing rate 5%, Hardy-Weinberg equilibrium exact test threshold 10^{-6}, and MAF greater than 5% [29]. Then $288,647$ variants are obtained after quality control. We consider the M72 Fibroblastic disorders as phenotype, and choose individuals from UKBL as internal data with 229 cases and UKBB with controls as the external data. The overlap variants in these two samples are used in real analysis. And the covariates age, sex, and the first 10 principal components are adjusted in the model. The descriptive statistics of subjects from internal and external studies are shown in **Table 2**.

Table 2. Descriptive statistics of subjects from UK Biobank for real analysis.

Study	Samples size		
	Cases	Controls	Totals
UKBL (Internal)	229	22,472	22,701
UKBB (External)		297,068	297,068
Total	229	318,540	319,769

We apply iECAT-RC, Internal, and iECAT-Score to analyze M72 Fibroblastic disorders for two genotyping calling procedures in the UK Biobank. Four SNPs are detected to be associated with Fibroblastic disorders by all three methods at the significance level 5×10^{-8} (**Table 3** and **Figure 2**). iECAT-RC detect these three SNPs with smaller p-values. Among the four SNPs, SNP rs62228062 locates in gene WNT7B. A recent transcriptome study identified WNT7B as amongst the most enriched transcripts in anterior capsule tissue in patients undergoing arthroscopic capsulotomy surgery for frozen shoulder (tissue disorder) suggesting WNT7B as a potential causal gene at the locus [30]. SNP rs2290221 on chromosome 7 is identified for association with Fibroblastic disorders, and shows the strongest association signal with a p-value of 1.26×10^{-8} by iECAT-RC. This SNP is in the intronic of the genes secreted frizzled-related protein 4 (SFRP4) and ependymal related protein 1 (zebrafish) (EPDR1). And it is detected to be associated with Dupuytren's disease which has a large overlap with frozen shoulder-associated loci [31,32].

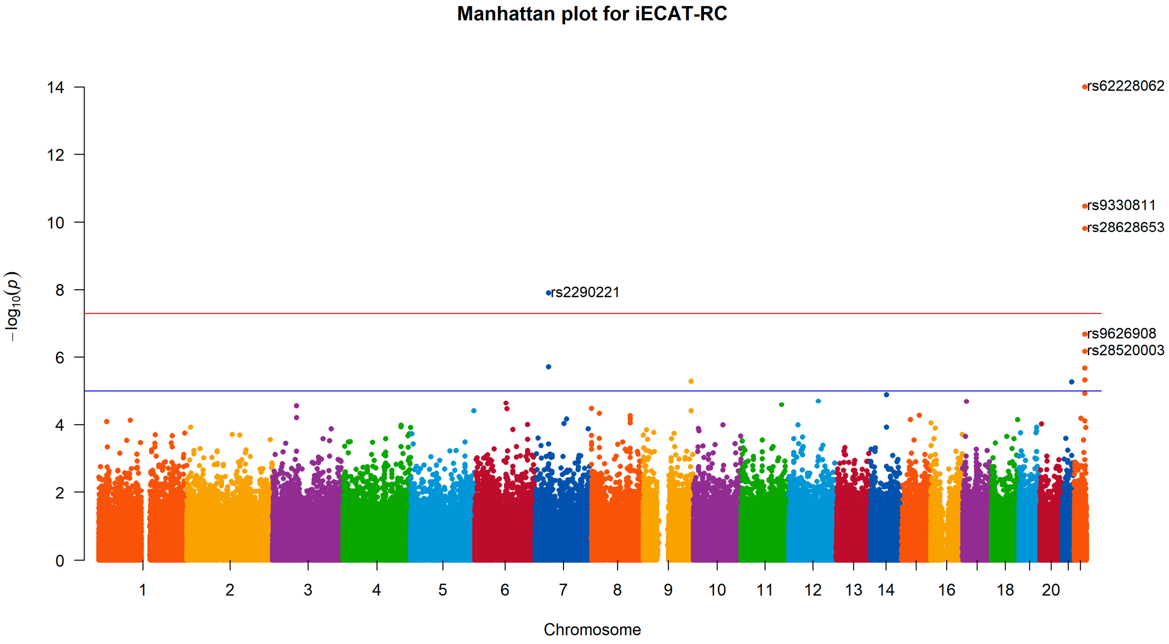

Figure 2. Manhattan plot for M72 Fibroblastic disorders based on iECAT-RC. The p-values are represented in genomic order by chromosome and position on the chromosome. The value on the y-axis represents the −log10 of the p-value. This plot is based on 22,701 individuals from UKBL and 297,068 individuals from UKBB. The genome-wide significance level is set at 5×10^{-8}. The most significant SNP in the experiment is rs62228062 in the WNT7B gene.

Table 3. Significant SNPs identified by iECAT-RC, iECAT-Score, and Internal at significance level of 5×10^{-8}.

Chromosome	SNP	Base Position	Genes	iECAT-RC	iECAT-Score	Internal
7	rs2290221	37987632	SFRP4, EPDR1	1.26E-08	2.91E-08	1.86E-08
22	rs9330811	46362396	WNT7B	1.65E-11	3.37E-11	3.00E-11
22	rs62228062	46381234	WNT7B	6.04E-18	8.82E-18	6.04E-18
22	rs28628653	46396925	LOC730668	1.54E-10	1.40E-10	1.54E-10

The Q-Q plot is used to assess the number and magnitude of observed associations between SNPs and the disease under study, compared to the association statistics expected under the null hypothesis of no association. The −log10 p-values calculated from each method are ranked in order from smallest to largest on the y-axis and plotted against the distribution that would be expected under the null hypothesis of no association on the x-axis. We test for association between the disease status of M72 Fibroblastic disorders and a SNP, adjusting for age, sex, and the first 10 principal components. The QQ plots from the tests integrating external control samples using the iECAT-RC method, Internal method, and iECAT-Score method are shown in **Figure 3**. We observe the similarity between the patterns of the three QQ plots, which are both close to the 45-degree line and show that all three methods can control Type I error rates well in this analysis.

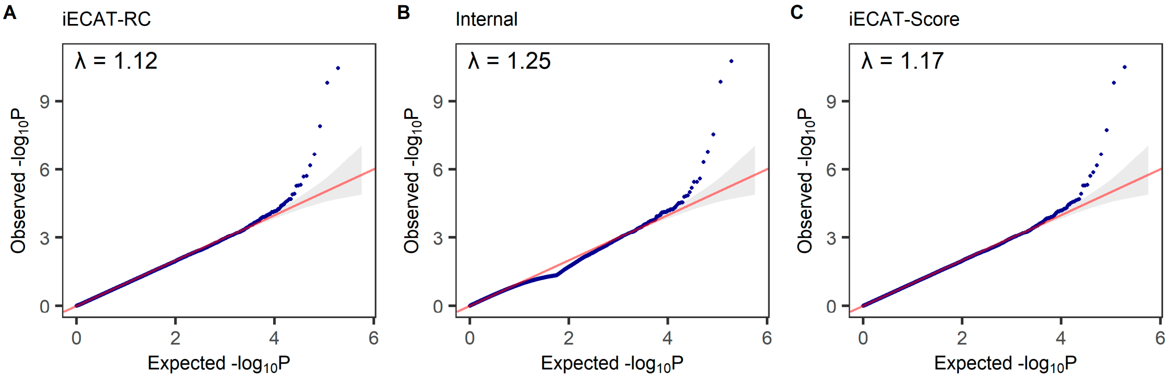

Figure 3. Quantile-quantile (QQ) plot of GWAS results based on iECAT-RC, Internal, and iECAT-Score. QQ plots show the distribution of expected p-value under the null model versus observed p-value in the -log10 scale. λ indicates the genomic inflation factor.

The case-control ratio of the combined samples has a significant impact on the performance of these three methods (iECAT-RC, Internal, and iECAT-Score), particularly in extremely unbalanced case-control studies, as observed in the simulation studies. Our method demonstrates increased statistical power when the case-control ratio is small. To assess the model's performance in real data analysis, we randomly select a subset from the real dataset while maintaining a value of $n_1^I : n_0^I : n_0^E$ is $1:1:2$. This allows us to compare the probabilities of detecting potentially significant SNPs using different methods. Specifically, we conduct $10,000$ random samples, with each sample comprising 229 internal cases, 229 internal controls, and 458 external controls. Then we implement different methods and the proportion of detected significant SNPs among the 10,000 samples is presented in **Table S2**. The proposed method, iECAT-RC, demonstrates a higher probability of detecting significant SNPs. For instance, the relative frequency of detecting SNP rs62228062 is 95.3%, surpassing that of the other two methods.

Discussion

In case-control studies, it is cost-effective to boost statistical power by increasing the sample size of case-control study. However, integrating external controls without considering systematic differences (batch effect) between studies, such as the differences in sequencing platforms, genotype calling procedures, population stratification, and so forth, may lead to inflated Type I error rates. In this paper, we propose an approach to integrate external control samples and allows for covariate adjustment. The proposed method, iECAT-RC, effectively addresses potential batch effects by calibrating bias using a regression model.

Simulation studies revealed that iECAT-RC can control for Type I error rates very well and boost power in the presence of batch effect. Specifically, we consider different simulation scenarios, including varying the batch effect level, DVS, and case-control ratios. Comparing iECAT-RC with three referenced methods, Internal, iECAT-Score, and iECAT-N, we demonstrate that all other methods could maintain Type I error rates except iECAT-N which naively combine internal and external samples without adjusting for the batch effects. Additionally, the simulation studies show that iECAT-RC has a higher power compared with other methods under different batch effect mechanisms.

In the real data analysis, we apply iECAT-RC, Internal, and iECAT-Score to genetic data from approximately 500,000 individuals with 784,256 SNPs across the United Kingdom. These individuals are used to identify the association between SNPs and M72 Fibroblastic disorders, while considering the genotype calling as the batch effect. Although all of the three methods, iECAT-RC, Internal, and iECAT-Score, identify four SNPs that are significantly associated with the disease, our proposed method has a higher probability of detecting these disease-associated SNPs compared to the other two methods when the case-control ratio is 1:3.

In conclusion, the proposed iECAT-RC method can integrate external control samples and at the same time, control type I error rate and boos statistical power. Through the linear regression calibration, we effectively reduce the batch effects arising from different platforms. Additionally, we employ SPA [18] and ER [19] methods to accurately calibrate p-values in scenarios of unbalanced case-control ratios and low MAFs. Our method provides a robust and effective improvement in score tests, ultimately contributing to a better understanding of the genetic architecture of complex diseases.

Author Contributions: Formal analysis and methodology: LZ, JS, XC, SZ, and QS; Data curation and visualization: LZ and XC; Writing original draft: LZ and SZ; writing review and editing: LZ, JS, XC, SZ, and QS.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Part of this research has been conducted using the UK Biobank Resource under application number 102999. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. Preparation of GAW 17 data was supported, in part, by NIH R01 MH059490 and used data from the 1000 Genomes Project (www.1000genomes.org).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Price AL, Spencer CC, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. Proceedings of the Royal Society B: Biological Sciences. 2015 Dec 22;282(1821):20151684.
2. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genetic epidemiology. 2012 Sep;36(6):561-71.
3. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics. 2017 Jul 6;101(1):5-22.
4. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature reviews genetics. 2005 Feb 1;6(2):95-108.
5. Fang S, Zhang S, Sha Q. Literature reviews on methods for rare variant association studies. Hum Genet Embryol. 2016;6:1-5.
6. Homann J, Osburg T, Ohlei O, Dobricic V, Deecke L, Bos I, Vandenberghe R, Gabel S, Scheltens P, Teunissen CE, Engelborghs S. Genome-wide association study of Alzheimer's disease brain imaging biomarkers and neuropsychological phenotypes in the European medical information framework for Alzheimer's disease multimodal biomarker discovery dataset. Frontiers in Aging Neuroscience. 2022 Mar 21;14:840651.
7. Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. The American Journal of Human Genetics. 2011 Sep 9;89(3):354-67.
8. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008 Oct;26(10):1135-45.
9. Skotte L, Korneliussen TS, Albrechtsen A. Association testing for next-generation sequencing data using score statistics. Genetic epidemiology. 2012 Jul;36(5):430-7.
10. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 2010 Jul;11(7):499-511.
11. Lee S, Kim S, Fuchsberger C. Improving power for rare-variant tests by integrating external controls. Genetic epidemiology. 2017 Nov;41(7):610-9.
12. Widmayer SJ, Evans KS, Zdraljevic S, Andersen EC. Evaluating the power and limitations of genome-wide association studies in Caenorhabditis elegans. G3. 2022 Jul 1;12(7):jkac114.

13. Liu DJ, Leal SM. SEQCHIP: a powerful method to integrate sequence and genotype data for the detection of rare variant associations. Bioinformatics. 2012 Jul 1;28(13):1745-51.
14. Derkach A, Chiang T, Gong J, Addis L, Dobbins S, Tomlinson I, Houlston R, Pal DK, Strug LJ. Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. Bioinformatics. 2014 Aug 1;30(15):2179-88.
15. Chen S, Lin X. Analysis in case–control sequencing association studies with different sequencing depths. Biostatistics. 2020 Jul;21(3):577-93.
16. Hendricks AE, Billups SC, Pike HN, Farooqi IS, Zeggini E, Santorico SA, Barroso I, Dupuis J. ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. PLoS genetics. 2018 Oct 16;14(10):e1007591.
17. Li Y, Lee S. Integrating external controls in case–control studies improves power for rare-variant tests. Genetic Epidemiology. 2022 Apr;46(3-4):145-58.
18. Dey R, Schmidt EM, Abecasis GR, Lee S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. The American Journal of Human Genetics. 2017 Jul 6;101(1):37-49.
19. Lee S, Fuchsberger C, Kim S, Scott L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. Biostatistics. 2016 Jan 1;17(1):1-5.
20. Li Y, Lee S. Novel score test to increase power in association test by integrating external controls. Genetic epidemiology. 2021 Apr;45(3):293-304.
21. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics. 2014 Jul 3;95(1):5-23.
22. Ma C, Blackwell T, Boehnke M, Scott LJ, GoT2D Investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genetic epidemiology. 2013 Sep;37(6):539-50.
23. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203-9.
24. McGuirl MR, Pattillo Smith S, Sandstede B, Ramachandran S. Hierarchical clustering of gene-level association statistics reveals shared and differential genetic architecture among traits in the UK Biobank. bioRxiv. 2019 Mar 4:565903.
25. Zhao Z, Bi W, Zhou W, VandeHaar P, Fritsche LG, Lee S. UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. The American Journal of Human Genetics. 2020 Jan 2;106(1):3-12.
26. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 2011 Jun;12(6):443-51.
27. Tängdén T, Gustafsson S, Rao AS, Ingelsson E. A genome-wide association study in a large community-based cohort identifies multiple loci associated with susceptibility to bacterial and viral infections. Scientific Reports. 2022 Feb 16;12(1):2582.
28. Hu YJ, Liao P, Johnston HR, Allen AS, Satten GA. Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. PLoS genetics. 2016 May 6;12(5):e1006040.
29. Liang X, Cao X, Sha Q, Zhang S. HCLC-FC: A novel statistical method for phenome-wide association studies. Plos one. 2022 Nov 9;17(11):e0276646.
30. Green HD, Jones A, Evans JP, Wood AR, Beaumont RN, Tyrrell J, Frayling TM, Smith C, Weedon MN. A genome wide association study of frozen shoulder identifies a common variant of WNT7B and diabetes as causal risk factors. medRxiv. 2020 Nov 16:2020-11.
31. Michou L, Lermusiaux JL, Teyssedou JP, Bardin T, Beaudreuil J, Petit-Teixeira E. Genetics of Dupuytren's disease. Joint Bone Spine. 2012 Jan 1;79(1):7-12.
32. Green HD, Jones A, Evans JP, Wood AR, Beaumont RN, Tyrrell J, Frayling TM, Smith C, Weedon MN. A genome-wide association study identifies 5 loci associated with frozen shoulder and implicates diabetes as a causal risk factor. PLoS genetics. 2021 Jun 10;17(6):e1009577.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.