RESEARCH ARTICLE

WILEY **Genetic Epidemiology**

OFFICIAL JOURNAL
INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Improving power for rare-variant tests by integrating external controls

Seunggeun Lee[1,2] (iD) | Sehee Kim[1] | Christian Fuchsberger[1,2,3]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA

[3]Center for Biomedicine, European Academy of Bolzano/Bozen, affiliated to the University of Lübeck, Bolzano/Bozen, Italy

Correspondence
Seunggeun Lee, M4148 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109-2029, USA.
Email: leeshawn@umich.edu

ABSTRACT

Due to the drop in sequencing cost, the number of sequenced genomes is increasing rapidly. To improve power of rare-variant tests, these sequenced samples could be used as external control samples in addition to control samples from the study itself. However, when using external controls, possible batch effects due to the use of different sequencing platforms or genotype calling pipelines can dramatically increase type I error rates. To address this, we propose novel summary statistics based single and gene- or region-based rare-variant tests that allow the integration of external controls while controlling for type I error. Our approach is based on the insight that batch effects on a given variant can be assessed by comparing odds ratio estimates using internal controls only vs. using combined control samples of internal and external controls. From simulation experiments and the analysis of data from age-related macular degeneration and type 2 diabetes studies, we demonstrate that our method can substantially improve power while controlling for type I error rate.

KEYWORDS

Rare-variant test, external controls, next-generation sequencing

1 | INTRODUCTION

Identification of genetic variants that predispose to complex diseases is an essential step toward understanding disease etiology, which can lead to breakthroughs in diagnosis, prevention, and treatment. The advance of sequencing technologies (Shendure & Ji, 2008) enables studying the full spectrum of genetic variants, including rare variants (minor allele frequency (MAF) <1%), in large studies. Although recent sequencing studies have started to identify disease-associated rare variants (Cruchaga et al., 2014; Steinthorsdottir et al., 2014), the number of discoveries is much smaller than initially predicted (Zuk et al., 2014). To facilitate further discoveries, more efficient and powerful statistical strategies and methods are needed.

Given decreasing sequencing costs, the number of sequenced whole exomes and even whole genomes are rapidly increasing. These sequenced samples provide a great opportunity to increase the power of rare-variant test. For a single study, if sequenced samples from other studies are used as control samples (external control samples), the power of rare-variant tests can be substantially improved without any additional sequencing cost. For example, using 1,529 external control samples from NHLBI GO Exome Sequencing Project (ESP) data in addition to their own 789 internal control samples, Zhan et al. (2013) identified a rare coding variant (MAF = 0.4%, P-value = 2.7×10^{-4}) in the C3 gene associated with age-related macular degeneration (AMD). When they exclusively used internal controls, the variant was substantially less significant.

Although the use of external control samples can greatly improve power, systematic differences between studies can have a negative impact on type I error control and power. Study heterogeneity can arise from differences in study populations, i.e., population stratification, or technical batch effects due to the use of different sequencing platforms or genotype calling pipelines (Quail et al., 2012). Recently, substantial progress has been made in identifying external control samples with similar genetic background (Bodea et al., 2016;

Wang et al., 2014); however, only limited success has been made in adjusting for technical batch effects. Derkach et al. (2014) developed a robust score test that uses genotype likelihoods. The method can control type I error rates when locations of variants are correctly known; however, inflated type I error rates are noted when variant locations should be inferred from genotype calls (Hu, Liao, Johnston, Allen, & Satten, 2016). More recently, Hu et al. (2016) developed a likelihood-based burden test that uses sequence reads, which can provide more accurate type I error rate controls. However, given that raw high depth sequence read data can be >1,000 times larger than processed genotype data, downloading, storing, and analyzing external control sequence read data can be a huge computational burden. In addition, the aforementioned methods need individual-level genotype and phenotype information, which is often difficult to obtain.

To address these challenges, we propose simple and robust rare-variant association tests that only require allele counts from external studies. Our approach, integrating External Controls into Association Test (iECAT), is based on the insight that the existence of batch effects on a given variant can be assessed by comparing two different odds ratio estimates: odds ratio estimate using internal controls only vs. odds ratio estimate using combined control samples of internal and external controls. If a variant is subject to batch effects, these two should be substantially different; even when ancestry-matched external controls were used (Mahajan & Robertson, 2015). In such a case, only internal control samples should be used to avoid false positives. Otherwise, external control samples can be added to increase the sample size. To approach this problem data adaptively, we propose using an empirical Bayesian-type method. We first construct a single variant test based on the shrinkage method and extend it to several gene- or region-based tests, such as burden tests (Li & Leal, 2008; Madsen & Browning, 2009; Morris & Zeggini, 2010), variance-component tests (Wu et al., 2011), and combinations of these two (e.g., SKAT-O; Lee, Wu, & Lin, 2012).

Given that only allele counts from external studies are needed for the proposed method, the method can be used with summary information publicly available in variant servers, such as the ESP data on the Exome Variant Server (ESP). Through extensive simulation studies and analysis of AMD and type 2 diabetes sequencing data, we demonstrate that the proposed method can improve power while controlling for type I error rates.

2 | MATERIAL AND METHODS

2.1 | Single-variant association test

For a single-variant test, we consider the following shrinkage estimation-based allelic test with an assumption

TABLE 1 Data setup for internal case-control and external-control samples for a single-variant association test

		$G_j = 1$	$G_j = 0$	Total
Internal	$Y = 1$	r^I_1	$2n^I_1 - r^I_1$	$2n^I_1$
	$Y = 0$	r^I_0	$2n^I_0 - r^I_0$	$2n^I_0$
External	$Y = 0$	r^E_0	$2n^E_0 - r^E_0$	$2n^E_0$

that the Hardy-Weinberg equilibrium (HWE) holds. Let $Y = 1(Y = 0)$ denote affected (unaffected) status, $G = 1(G = 0)$ denote the minor allele (major allele) in the variant. Let n^I_0 and n^I_1 (n^E_0 and n^E_1) denote the number of controls and cases of an internal study (external study), respectively. The data are represented by a 3×2 table in Table 1.

Considering the internal study only, let p^I_y denote the *unknown true* MAF (i.e., $p^I_y = Pr[G = 1|Y = y]$ for $y = 0, 1$), and r^I_Y denote the corresponding *observed* number of minor alleles. The observed counts can be viewed as a random sample from two independent binomial distributions, $r^I_0 \sim$ Binomial $(2n^I_0, p^I_0)$ and $r^I_1 \sim$ Binomial $(2n^I_1, p^I_1)$. Note that the number of trials is two times n^I_0 (or n^I_1) because each sample consists of two copies of chromosomes. Suppose the parameter of interest is a log odds ratio of the genetic effect, given by

$$\beta_0 = \log[p^I_1(1 - p^I_0)/\{p^I_0(1 - p^I_1)\}].$$

Our goal is to test whether $H_0 : \beta_0 = 0$ or not. By using the internal study only, β_0 can be consistently estimated by $\hat{\beta}_{Int} = \log\left[r^I_1\left(2n^I_0 - r^I_0\right)/\left\{r^I_0\left(2n^I_1 - r^I_1\right)\right\}\right]$.

Now, suppose n^E_0 external control samples are available, and assume the *observed* number of minor alleles, r^E_0, follows a binomial distribution with p^E_0, the proportion of the minor allele in the external controls, i.e., $r^E_0 \sim$ Binomial $(2n^E_0, p^E_0)$. Combining control samples from internal and external studies, the estimate of β_0 is

$$\hat{\beta}_{Com} = \log[r^I_1(2n^I_0 + 2n^E_0 - r^I_0 - r^E_0)/\{(r^I_0 + r^E_0)(2n^I_1 - r^I_1)\}]$$

When $p^I_0 = p^E_0$, indicating that no batch effect exists, $\hat{\beta}_{Com}$ converges to β_0 in probability. When $p^I_0 \neq p^E_0$, however, $\hat{\beta}_{Com}$ converges to a biased quantity, $\beta_{0,Com} = \log[p^I_1(1 - p^{Com}_0)/\{p^{Com}_0(1 - p^I_1)\}]$, where $p^{Com}_0 = n^I_0/(n^I_0 + n^E_0)p^I_0 + n^E_0/(n^I_0 + n^E_0)p^E_0$. Thus, the bias can be quantified as $\theta = \beta_{0,Com} - \beta_0 = \log[p^I_0(1 - p^{Com}_0)/\{p^{Com}_0(1 - p^I_0)\}]$. Taking it for granted that the bias would exist, we propose a bias-adjusted estimate of β_0 as follows.

We first estimate the bias θ by using an empirical Bayesian approach. It begins with an estimate $\hat{\theta} = \hat{\beta}_{Com} - \hat{\beta}_{Int}$ that asymptotically follows a normal distribution N$(\theta, \sigma^2_\theta)$, where σ^2_θ is the variance of $\hat{\theta}$. To account for the uncertainty in the bias, we assume that θ is a random variable following a normal distribution N$(0, \nu)$, where the unknown hyper-

parameter, ν, reflects the magnitude of uncertainty by the batch effects. Note that the consistent estimator of ν from the marginal distribution of $\hat{\theta}$ is $\hat{\nu}^+ = \max(0, \hat{\theta}^2 - \hat{\sigma}_{\theta}^2)$, where $\hat{\sigma}_{\theta}^2$ is an estimate of σ_{θ}^2 (Morris, 1983). Recently Mukherjee and Chatterjee (2008) have proposed to use $\hat{\nu}^* = \hat{\theta}^2$, a more conservative estimator of ν. The main advantage of using $\hat{\nu}^*$ instead of $\hat{\nu}^+$ is that $\hat{\nu}^*$ provides a closed form of variance estimator of the subsequent θ estimator. They also showed that there is essentially no loss of efficiency of using $\hat{\nu}^*$. Therefore, we propose to use $\hat{\nu}^*$, and the resulting estimate of the posterior mean of θ is $\tilde{\theta} = \frac{\hat{\theta}^2}{\hat{\theta}^2 + \hat{\sigma}_{\theta}^2} \hat{\theta}$. Finally, the proposed estimator of β_0 is

$$\hat{\beta} = \hat{\beta}_{Com} - \tilde{\theta} = (1 - \tau)\hat{\beta}_{Com} + \tau \hat{\beta}_{Int}; \ \tau = \frac{\hat{\theta}^2}{\hat{\theta}^2 + \hat{\sigma}_{\theta}^2}. \quad (1)$$

When $\hat{\theta}$ (i.e., the difference between $\hat{\beta}_{Com}$ and $\hat{\beta}_{Int}$) is large, τ is close to one, and hence $\hat{\beta}$ becomes $\hat{\beta}_{Int}$. In contrast, when τ is close to zero due to a small $\hat{\theta}$, $\hat{\beta}$ is close to $\hat{\beta}_{Com}$. Regardless of the size of the true bias θ, $\hat{\beta}$ is an asymptotically unbiased estimator of β_0 (supplementary Appendix A). The exact expression of the asymptotic variance of $\hat{\beta}$ can be found in the supplementary Appendix A. When τ is large and hence close to one, $\hat{\beta}$ can have a larger variance than $\hat{\beta}_{Int}$. In this case, we propose to use $\hat{\beta}_{Int}$ to achieve better power.

If there is a prior information on when θ could be zero under the null hypothesis (i.e., $p_1^I = p_0^I$), using $\tau = 0$ can increase power without inflating type I error rates. Suppose that observed MAFs of external controls (i.e., $r_0^E/2n_0^E$) is in between observed MAFs of cases and internal controls (i.e., $r_y^I/2n_y^I$), and hence $\hat{\beta}_{Com}$ is closer to zero than $\hat{\beta}_{Int}$ (i.e., $|\hat{\beta}_{Com}| < |\hat{\beta}_{Int}|$). Under the null hypothesis with the batch effects (i.e., $p_1^I = p_0^I \neq p_0^E$), because observed MAFs converge to true MAFs, the probability to observe this phenomenon will converge to zero. This indicates that using $\tau = 0$ in this situation will not increase type I error rates. In contrast, using $\tau = 0$ in this situation will increase power especially when the external controls include case samples, because the case contamination will cause either $p_1^I < p_0^E < p_0^I$ or $p_1^I > p_0^E > p_0^I$ under the alternative hypothesis. And hence when this phenomenon occurs, we used $\tau = 0$ as a default value in simulation studies and real data analysis.

2.2 | Gene- or region-based tests

For gene- or region-based tests, we extend the proposed single-variant test in the previous section to burden, SKAT- and SKAT-O type tests. The key idea is first to construct a single-variant score-type statistic for each variant using $\hat{\beta}$ and then to aggregate them using their weighted linear or quadratic sums. To aggregate association signals in each variant, the burden test uses a weighted linear sum of score statistics, whereas SKAT uses a weighted quadratic sum of score statistics. The combined test, SKAT-O, uses a linear combination of the burden and SKAT test statistics.

Suppose that a region being tested has p variant loci. For variant j, let $\hat{\beta}_j$ be the log odds ratio estimate in (1), $\hat{\sigma}_j$ the standard error estimate, and q_j the MAF estimate from the internal samples. In supplementary Appendix B, we show that a single-variant score statistic is approximately proportional to the product of log odds ratio estimate, sample size, and the genotype variance. Using this fact, we construct a score-type statistic for a single variant j as $S_j = n_{eff}q_j(1 - q_j)\hat{\beta}_j$, where n_{eff} is the effective sample size, and propose the following test statistics:

$$Q_{iECAT}(\rho) = (1 - \rho)\sum_{j=1}^{p}\left(w_j S_j\right)^2 + \rho\left(\sum_{j=1}^{p}w_j S_j\right)^2,$$

where w_j is a weight for variant j, and ρ is a parameter between 0 and 1. Clearly, $\rho = 0$ corresponds to the SKAT-type test, and $\rho = 1$ corresponds to the burden-type tests. Note that $Q_{iECAT}(\rho)$ has a similar form to the variance component test with the compound symmetric kernel structure proposed by Lee et al. (2012). More details on the test and P-value calculation are given in supplementary Appendix C.

Burden-type tests (i.e., $\rho = 1$) are powerful when large percentage of variants are causal and effects are in the same direction. SKAT-type tests (i.e., $\rho = 0$) are powerful when heterogeneity is noted regarding effect sizes and the direction of the effects (Basu & Pan, 2011; Lee, Abecasis, Boehnke, & Lin, 2014). SKAT-O type tests combine burden and SKAT tests using the minimum P-values from a grid of ρ. We propose the following SKAT-O type combined test as

$$Q_{iECAT-O} = \min P - value(\rho),$$

where P-value(ρ) is the P-value $Q_{iECAT}(\rho)$ with a given ρ. We used a grid of eight ρ values (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1) in simulation studies and real data analysis. This approach has been used in previous studies and shown to provide good performances in type I error control and power (Lee, Teslovich, Boehnke, & Lin, 2013). P-values of $Q_{iECAT-O}$ can be obtained using numerical integration, as described in Lee et al. (2012).

2.3 | Type I error and power simulations

We performed extensive simulation studies to evaluate the performance of the proposed iECAT method. To generate sequence information, we simulated 40,000 European-like and 40,000 African America like haplotypes for 200 kbps using the coalescent simulator COSI with the calibrated demographic model (Schaffner et al., 2005). The binary

phenotypes were generated from the following logistic regression model:

$$logit\ P(Y = 1) = \alpha_0 + G'_{causal}\beta_{causal}, \qquad (2)$$

where G_{causal} is a genotype matrix containing causal variants, and β_{causal} is a vector of the genetic effect coefficients. The intercept α_0 was chosen for the disease prevalence of 0.05.

We presumed that 3% of the variants had different MAFs between internal and external control samples due to batch effects, mimicking the level of the batch effects observed in the real data analysis (see Section 3 for details). For these variants, MAFs in external control samples were randomly generated from Uniform $(0.1 \times q, 4 \times q)$, where q is the MAFs in internal study samples. To mimic realistic scenarios of population stratification between internal and external control samples, we generated 0% (no population stratification), 5%, and 10% of external control samples from the African American like haplotypes. All other internal and remaining external control samples were simulated from the European-like haplotypes.

We compared the following methods for a region-based test: (1) the proposed iECAT-O; (2) SKAT-O using the internal control only; (3) iECAT-O without adjusting for batch effects (iECAT-O$_{Noadj}$)—i.e., $\hat{\beta} = \hat{\beta}_{Com}$ was always used regardless of the existence of the batch effects; (4) iECAT-O assuming that batch effect variants were known prior to the test (iECAT-O$_{Known}$)—i.e., $\hat{\beta} = \hat{\beta}_{Int}$ was used (i.e., τ is fixed with 1) in the presence of true batch effects. Otherwise, $\hat{\beta} = \hat{\beta}_{Com}$ was used (i.e., τ is fixed with 0). Because iECAT-O$_{Known}$ was constructed given that variants subject to batch effects were fully known, iECAT-O$_{Known}$ could reach the theoretical maximum power under the iECAT framework. Obviously, this batch effect information could not be known in real data. Therefore, iECAT-O$_{Known}$ will be used only for power comparison to quantify the efficiency loss in iECAT-O due to the uncertainty in τ. For a single-variant test, we used aforementioned approaches (iECAT-O, iECAT-O$_{Noadj}$, and iECAT-O$_{Known}$) and the Wald test (internal controls only) with testing one variant at a time.

For region-based tests, we randomly selected a 3 kbp region and tested for an association between variants in the region and the phenotypes. For very rare variants, the proposed methods cannot be used because $\hat{\theta}$ estimates can be unstable. We thus only used variants observed in both internal and external studies with internal study MAC > 3. For single-variant tests, we randomly selected a variant and tested for an association between the variant and the phenotypes.

In type I error simulations, phenotypes were generated from (2) with $\beta_{causal} = 0$. For each simulation, to evaluate type I error rates at genome-wide significance levels, we generated 10^7 datasets for region-based tests and 5×10^7 datasets for single-variant tests. Given that generating these large datasets was computationally intensive, we considered only two internal study sample sizes, moderate and large ($N_{internal} = 4,000$ and $10,000$). In each sample size, two different ratios of case-control (1:1 and 2:1) were considered. The external control sample size ($N_{external}$) was set to be the same as the internal study sample size.

For power simulations for gene-based tests, 5%, 10%, or 30% of rare variants (MAF < 1%) were assumed to be causal. In each setting, either all causal variants were risk-increasing or 80% of causal variants were risk increasing (the remaining 20% were risk decreasing). Given that it is possible that rarer variants have larger effect sizes, we modeled log OR as a decreasing function of MAF. Specifically, $\beta = c|\log_{10}(MAF)|$. When 30% of variants were causal, we used $c = \log(2)/2$, which led to OR = 2 for a causal variant with MAF = 1%. When 5% or 10% of rare variants were causal, we used $c = \log(4)/2$ and $c = \log(3)/2$ to compensate the decreased numbers of causal variants. For single-variant tests, we evaluated power for testing a variant with MAF = (0.01, 0.005) and OR = 2.

To evaluate the power when external control samples contain samples with the disease of interest (i.e., case contamination), we generated external control samples of which 0%, 5%, and 10% were the diseased samples. Because 5% disease prevalence was noted, 5% contamination was equivalent to using the general population as external controls. In all power simulation scenarios, the external control sample size was the same as the internal study sample size ($N_{external} = N_{internal}$). For each power simulation setting, we generated 5,000 data sets.

2.4 | Real data analysis

We applied our method to two sequence datasets while using 4,300 ESP European samples as external controls: targeted sequencing data from the AMD study (Zhan et al., 2013) and deep whole exome sequence data from the Genetics of Type-2 Diabetes (GoT2D) study (Fuchsberger et al., 2016). The AMD study sequenced 10 target loci spanning 56 genes (2.7 Mb in total). We downloaded 3,350 mapped SRA files from dbGaP (Tryka et al., 2014) (phs000684.v1.p1), applied variant calling and QC procedures as described in Zhan et al. (2013), and retained 2,317 cases and 791 controls. The GoT2D study whole exome sequenced 1,326 European T2D cases and 1,331 European T2D controls at high depth. For both datasets, we applied LASER software with the HGDP reference (Wang et al., 2014) to identify population structure. Given that Finish samples were separately clustered from other European samples (supplementary Fig. S8), and ESP has only a small number of samples clustered together with Finish samples (supplementary Fig. S8), we exclusively used the non-Finish cohorts (British, German, and Sweden) in the GoT2D data analysis, leading to a total of 650 cases and 646 controls.

For external controls, allele counts of ESP samples were downloaded from the ESP Exome Variant Sever. Variants observed in both internal and external studies were included in the analysis, after excluding variants that did not pass each study's own QC criteria. We further excluded variants with internal study MAC ≤ 3. For gene-based tests, we used genes with at least three rare variants. To investigate whether ancestry matching can improve type I error control and power for whole exome scale data analysis, we obtained individual-level genotype data of ESP for the GoT2D data analysis. We used LASER software with the HGDP reference for 1:2 matching between internal and external samples and identified 2,568 ancestry-matched external control samples (supplementary Fig. S8).

3 | RESULTS

3.1 | Type I error rate simulation results

The empirical type I error rates estimated for iECAT-O and the other methods are presented in Table 2 for $\alpha = 10^{-4}$ and 2.5×10^{-6}, corresponding to candidate gene studies of 500 genes and exome-wide studies of 20,000 genes. All internal and external samples were simulated using European-like haplotypes. In all the simulation scenarios we considered, iECAT-O had well controlled or slightly inflated type I error rates. As expected, when we used the combined control samples but ignored batch effects (iECAT-O$_{\text{Noadj}}$), type I error rates were significantly inflated. For example, empirical type I error rates were increased more than 10,000-fold compared to the nominal α level on average when $\alpha = 2.5 \times 10^{-6}$. When we exclusively used the internal control samples (SKAT-O), type I error rates were well controlled.

When only summary-level information for external control samples is available, samples with different genetic background cannot be excluded. We carried out additional type I error simulations with 5% and 10% of external control samples from the African American like haplotypes, and the results have shown that iECAT-O had robust type I error control even in the presence of population stratification (details See supplementary Appendix D).

3.2 | Power simulation results

Next, we compared power of the proposed and existing methods to identify genetic associations at $\alpha = 2.5 \times 10^{-6}$. Figure 1 presents power simulation results when all causal variants were deleterious variants. iECAT-O had substantially improved power compared to the approach using internal study samples only. For example, when the internal study case-control ratio was 1:1 and the internal study sample size was 10,000, iECAT-O was on average 38–60% more

powerful. When the internal study case-control ratio was 2:1, iECAT-O exhibited increased power. On average, when the internal study sample size was 10,000, the strategy to sequence more cases (2:1 case-control ratio compared to 1:1) resulted in 11–17% increased power for iECAT-O compared to sequencing the same number of cases and controls (Fig. 1, bottom panel). As expected, the power of iECAT-O$_{\text{Known}}$, where the optimal value of shrinkage weight τ was used, was higher than that of iECAT-O, but the difference was not substantial. This implies that estimating τ in iECAT-O method did not reduce the efficiency so much. Given that iECAT-O$_{\text{Noadj}}$ greatly increased type I error rates, we did not include it in this plot. The relative performance of the methods in the presence of both risk-increasing and risk-decreasing variants was quantitatively similar (supplementary Fig. S2).

Given that external studies were not performed for the target disease of interest in the internal study, it is possible that external control samples included cases (i.e., the diseased samples), and this case contamination can reduce the power. Figure 2 presents power evaluation results with 0%, 5%, and 10% of the external control samples being cases. Given that 5% disease prevalence was assumed, 5% contamination is equivalent to using the general population as external controls. When $N_{\text{internal}} = 10,000$, the power was decreased on average by 2–4% for 5% contamination and 3–8% for 10% contamination. When the sample size was small, the power reduction was slightly larger.

Next, we performed power simulations in the presence of population stratification between internal and external samples in which 5% and 10% of the external control samples were generated from African American like haplotypes and all other samples were generated from European-like haplotypes. Given that population stratification can inflate type I error rates, we used empirical α levels obtained in the type I error simulation studies for fair comparisons. Supplementary Figure S3 indicates that the power of iECAT-O slightly decreased in the presence of population stratification. These results suggest that the power of iECAT-O can be improved if the ancestry-based matching method is used to exclude external control samples with different genetic backgrounds.

As expected, the power simulation results with for iECAT with $\rho = 0$ (SKAT-type test) and $\rho = 1$ (burden type of tests) were quantitatively similar (supplementary Figs. S4 and S5). We obtained power with empirical α levels instead of nominal α levels for $N_{\text{internal}} = 4,000$ and 10,000, and the results were also quantitatively similar (supplementary Fig. S6). We then performed power simulations for single-variant tests for two different MAFs (1% and 0.5%) (supplementary Fig. S7). The power curves revealed that iECAT had greatly improved power compared with the method exclusively using internal controls (e.g., 43% vs. 12% when MAF = 0.5%, $N_{\text{internal}} = 10,000$ and case:control = 1:1). Interestingly, the strategy of

TABLE 2 Empirical type I error rates for iECAT-O, SKAT-O, and iECAT-O_Noadj

Internal sample size	Internal case:control	Level α	iECAT-O	SKAT-O	iECAT-O_Noadj
4,000	1:1	10^{-4}	7.40×10^{-5}	1.10×10^{-4}	7.00×10^{-2}
	1:1	2.5×10^{-6}	1.40×10^{-6}	2.10×10^{-6}	5.50×10^{-2}
	2:1	10^{-4}	8.00×10^{-5}	1.20×10^{-4}	8.40×10^{-2}
	2:1	2.5×10^{-6}	2.60×10^{-6}	2.10×10^{-6}	7.00×10^{-2}
10,000	1:1	10^{-4}	8.60×10^{-5}	1.10×10^{-4}	1.10×10^{-1}
	1:1	2.5×10^{-6}	3.50×10^{-6}	2.90×10^{-6}	9.30×10^{-2}
	2:1	10^{-4}	8.70×10^{-5}	1.10×10^{-4}	1.20×10^{-1}
	2:1	2.5×10^{-6}	2.40×10^{-6}	3.10×10^{-6}	1.10×10^{-1}

Each cell has an empirical type I error rate estimated from 10^7 simulated datasets. The external control sample sizes were the same as the internal sample sizes. All internal and external controls samples were simulated from European-like haplotypes.
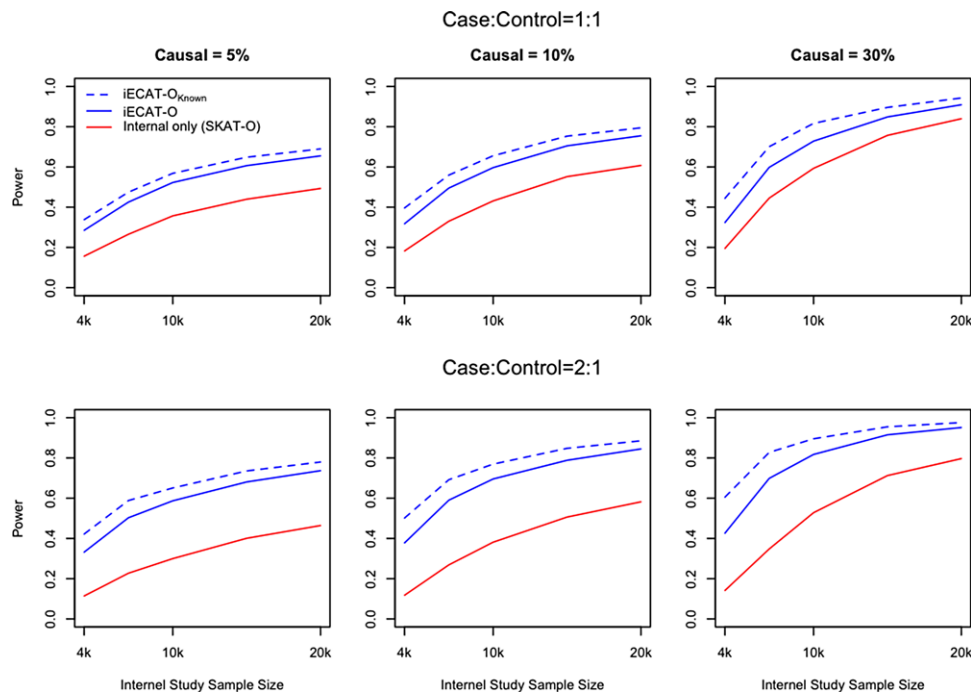
FIGURE 1 Power comparisons when all causal variants were risk increasing variants.
Each line represents empirical power at $\alpha = 2.5 \times 10^{-6}$. From left to right, the plots consider that 5%, 10%, and 30% of variants were causal variants, respectively. From top to bottom the plots consider that internal study case:control ratio was 1:1 and 2:1, respectively. The external control sample sizes were the same as the internal study sample sizes. For causal variants, we assumed that $\beta = c|\log_{10}(\text{MAF})|$ (see Section 2)

sequencing more cases (case:control = 2:1) did not always improve the power compared with the one case per one control strategy. For example, when MAF = 0.5% and N_{internal} = 15,000, case:control = 2:1 design had 72% power, whereas case:control = 1:1 design had 75% power. This occurred in part because having a smaller number of internal controls increased the variability of the bias estimate, which had a negative impact on power. We note that this increased variability had a smaller effect on gene-based tests given that the gene-based approach combines multiple single-variant odds ratios.

Overall, our simulation studies demonstrate that iECAT can substantially improve power compared with the existing

approaches that exclusively use internal control samples, while controlling for type I error in a wide range of scenarios. In contrast, if the external control samples are used without adjusting for batch effects, it can result in significantly increased type I error rates.

3.3 | Real data analysis

We applied our proposed methods to the analysis of AMD targeted sequencing and GoT2D whole exome sequencing datasets. For both datasets, we used allele count statistics of 4,300 ESP European samples downloaded from ESP as
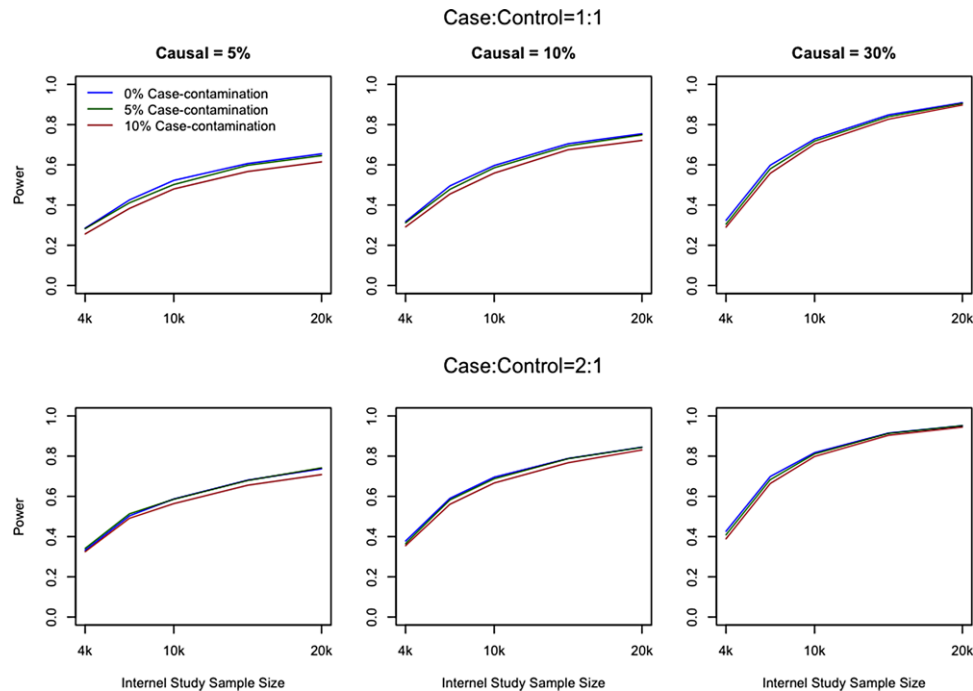
FIGURE 2 Empirical power of iECAT-O in the presence of case contamination in external controls.
Each line represents empirical power at $\alpha = 2.5 \times 10^{-6}$. From left to right, the plots consider that 5%, 10%, and 30% of variants were causal variants, respectively. From top to bottom the plots consider that internal study case:control ratio was 1:1 and 2:1, respectively. All causal variants were risk-increasing variants. The external control sample sizes were the same as the internal study sample size. For causal variants, we assumed that $\beta = c|\log_{10}(\text{MAF})|$ (see Section 2)

external controls. In addition, for the GoT2D data analysis we obtained the ancestry-matched external control samples using individual-level genotype data of ESP to investigate whether matching can improve type I error control and power in whole exome scale analysis.

AMD Study: We applied iECAT to 56 candidate genes in the AMD targeted sequencing dataset. After quality control, we retained 2,317 cases and 791 controls. Table 3 presents the top five genes by iECAT-O. For this analysis, we focused on rare variants (MAF < 1%). By combining the internal and external controls, iECAT-O revealed that two well-known AMD-related genes, *C3* and *CFH*, had the smallest *P*-values (*P*-value = 5.75×10^{-6} and *P*-value = 7.44×10^{-6}, respectively), and the *P*-values were still significant after the Bonferroni correction (corrected $\alpha = 0.05/56 = 8.9 \times 10^{-4}$). In contrast, when SKAT-O was performed with the AMD dataset alone, the *P*-values were greater than 0.01, indicating that the proposed test significantly improved power. We also performed single-variant tests for a total of 538 rare variants (supplementary Table S3). We found that SNV rs147859257 in *C3* (*P*-value = 1.23×10^{-5}) and rs121913059 in *CFH* (*P*-value = 1.24×10^{-5}) were the top two variants by iECAT. However, when the AMD dataset was used alone, these two SNVs resulted in *P*-values > 10^{-2}. The single-variant test results were largely consistent with Zhan et al. (2013), in which 1,529 ancestry-matched external controls were used.

GoT2D Study: We performed single-variant and gene-based tests with GoT2D deep exome sequencing data. Given that the ESP dataset contains few Finish samples (see Section 2), we focused on non-Finish GoT2D cohorts in this analysis, which included 650 cases and 646 controls. The gene-based rare-variant test results are summarized through QQ plots in Figure 3 (top panel); the iECAT-O QQ plot was fairly well calibrated. In contrast, iECAT-O$_{\text{Noadj}}$ resulted in a significantly inflated QQ plot. Even after the external control samples were ancestry matched (a total of 2,568 external control samples; supplementary Fig. S8), the QQ plot by iECAT-O$_{\text{Noadj}}$ remained significantly inflated, which suggests that technical batch effects were a major source of the type I error inflation (Fig. 3, bottom panel). In contrast, iECAT-O had a well-calibrated QQ plot. Supplementary Figure S9 compares *P*-values calculated with all 4,300 ESP external control samples and 2,568 ancestry-matched external control samples. These two *P*-values were largely consistent, indicating that iECAT-O analyses with and without ancestry-based matching were largely similar in this dataset.

We also performed single-variant tests (supplementary Fig. S10) for low frequency and rare variants with MAF < 5%. As expected, QQ plots of iECAT were well calibrated, whereas QQ plots of iECAT$_{\text{Noadj}}$ were greatly inflated. When we compared allele frequencies between the internal control samples and the matched external control samples, approximately 13% and 23% variants had *P*-values < 10% and 20%, respectively.

TABLE 3 Top five genes by iECAT-O *P*-values from the AMD data analysis

Gene	Chr	No. of variant[a]	iECAT-O *P*-values	SKAT-O *P*-values
CFH	1	14	5.75×10^{-6}	6.70×10^{-2}
C3	19	27	7.44×10^{-6}	1.04×10^{-2}
DOM3Z	6	7	1.16×10^{-4}	1.69×10^{-3}
CFHR4	1	6	8.03×10^{-4}	6.66×10^{-2}
SLC44A4	6	16	1.39×10^{-3}	4.47×10^{-2}

The 4,300 European ESP samples were used as external controls. Rare variants (MAF < 1%) were exclusively used for this analysis.

[a]Variants with internal study MAC ≤ 3 were excluded from the analysis.
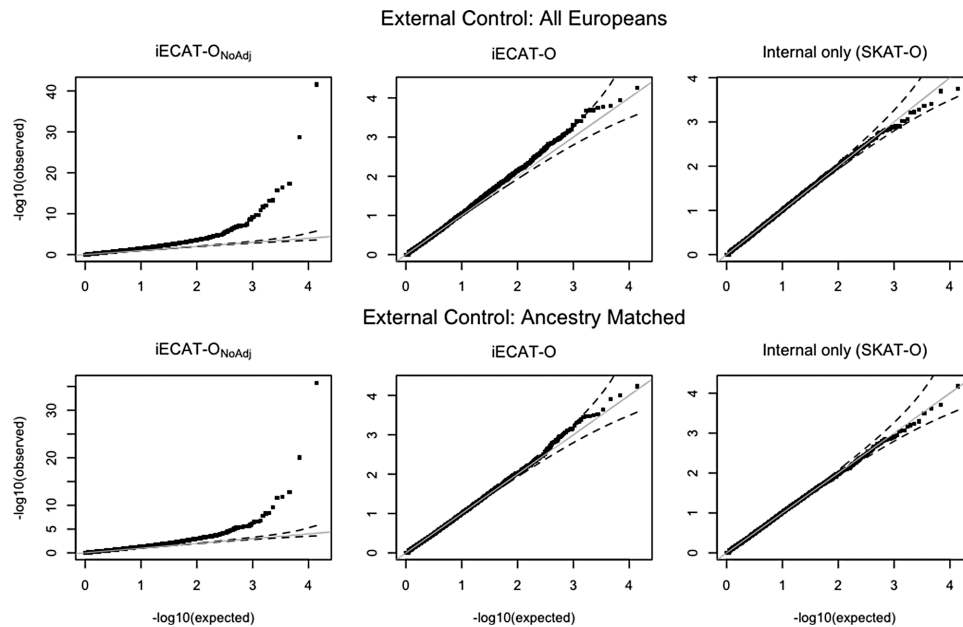
FIGURE 3 Analysis of the GoT2D exome data with ESP as external control samples.
QQ plots of $-\log_{10}$ *P*-values of gene-based tests for rare variants (MAF < 1%). A total of 11,283 genes with at least three rare variants were tested for associations with T2D status. The dashed line represents a 95% confidence band. Top panel: all 4,300 European ESP samples were used as external controls. Bottom panel: 2,568 ancestry-matched ESP samples were used as external controls

These 3% inflation may indicate that approximately 3% of variants were subject to technical batch effects. To investigate whether the batch effects vary by minor allele counts (MACs), we obtained a distribution of the shrinkage weight τ by MAC (supplementary Fig. S11). The box plots were similar across all four MAC bins, indicating that there was no clear pattern in batch effects by MAC.

Supplementary Table S4 presents the top five genes by iECAT-O. Although GoT2D analysis did not identify statistically significant T2D-associated genes, this analysis provided a suggestive association and demonstrated that the proposed approach can control for type I errors.

4 | DISCUSSION

In this paper, we proposed rare-variant tests that increase power by integrating external control samples. By estimating bias using an empirical Bayesian approach, the proposed

iECAT method provides an effective way to adjust for possible batch effects. The type I error simulation and GoT2D data analysis revealed that iECAT can control for type I errors in the presence of technical batch effects. The power simulation and AMD data analysis revealed that iECAT can improve power compared with the exclusive use of internal controls. The method is implemented in the iECAT R package (available on the author's website).

One of the important features of iECAT is that it only needs allele counts from external studies. By utilizing allele count information readily available in variant servers, such as ESP and ExAC browser (Lek et al., 2016), our method can greatly facilitate the use of external control samples. This is a desirable property considering the difficulties in obtaining individual-level genotype data. One possible issue of using allele count information in a variant server is that we cannot filter out samples with different genetic background. In such case, iECAT could yield slightly inflated type I error rates. We believe that it will become less problematic in the future

as variant servers are starting to provide allele counts for fine-scale ancestry.

There are several limitations in the present proposal. First, the proposed method cannot adjust for covariates such as age, gender, and principal components. In supplementary Appendix E, we present a simple approach that can adjust for covariates by estimating covariate-adjusted MACs. The approach is similar to the principal component adjustment method used in EigenStrat (Price et al., 2006), and a small-scale simulation study shows that the method can adjust for population stratification when the population index was used as a covariate. The presented approach, however, requires individual-level genotype and phenotype data from external controls. Second, the proposed method mainly assumed that one external study is used for external controls. If multiple study data are used as external controls, it is possible that there exist multiple batch effects. One possible way to addressing it is to apply the covariate adjustment in supplementary Appendix E to adjust for batch effects in the external control genotypes and to use adjusted MAC for the association tests. Third, the method assumes that HWE holds (Guedj, Nuel, & Prum, 2008). Because it is uncommon to observe rare allele homozygotes, the violation of HWE condition would have limited effects on the proposed tests. Fourth, we do not recommend using the method for singletons or doubletons because log odds ratio estimates can be unstable for these variants. In future work, we will extend the method to incorporate singletons and doubletons.

With the advances in sequencing technologies, the number of sequenced genomes is increasing rapidly. Our method provides a robust and effective way to utilize these sequenced genomes in rare-variant tests and will contribute to the understanding of the genetic architecture of complex diseases.

ACKNOWLEDGMENTS

REFERENCES

Basu, S., & Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7), 606–619.

Bodea, C. A., Neale, B. M., Ripke, S., Daly, M. J., Devlin, B., Roeder, K., & Consortium, International IBD Genetics. (2016). A method to exploit the structure of genetic ancestry space to enhance case-control studies. *American Journal of Human Genetics*, 98(5), 857–868.

Cruchaga, C., Karch, C. M., Jin, S. C., Benitez, B. A., Cai, Y., Guerreiro, R., … Bertelsen, S. (2014). Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*, 505(7484), 550–554.

Derkach, A., Chiang, T., Gong, J., Addis, L., Dobbins, S., Tomlinson, I., … Strug, L. J. (2014). Association analysis using next generation sequence data from publicly available control groups: The robust variance score statistic. *Bioinformatics*, 30(15), 2179–2188.

ESP, NHLBI GO. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA. (2016). Retrieved from http://evs.gs.washington.edu/EVS/.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., … McCarthy, D. J. (2016). The genetic architecture of type 2 diabetes. *Nature*, 536(7614), 41–47.

Guedj, M., Nuel, G., & Prum, B. (2008). A note on allelic tests in case-control association studies. *Annals of Human Genetics*, 72(3), 407–409.

Hu, Y.-J., Liao, P., Johnston, H. R., Allen, A., & Satten, G. (2016). Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. *PLoS Genetics*, 12(5), e1006040.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, 95(1), 5–23.

Lee, S., Teslovich, T. M., Boehnke, M., & Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *American Journal of Human Genetics*, 93(1), 42–53.

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4), 762–775.

Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., … Cummings, B. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291.

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, 83(3), 311–321.

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2), e1000384.

Mahajan, A., & Robertson, N. (2015). Rare variant quality control. In Eleftheria Zeggini & Andrew Morris (Eds.), *Assessing rare variation in complex traits* (pp. 33–43). New York: Springer.

Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2), 188–193.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381), 47–55.

Mukherjee, B., & Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case–control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3), 685–694.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for

stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., & Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, *15*(11), 1576–1583.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145.

Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., … Gudjonsson, S. A. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature Genetics*, *46*, 294–298.

Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., … Kimura, M. (2014). NCBI's database of genotypes and phenotypes: DbGaP. *Nucleic Acids Research*, *42*(D1), D975–D979.

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., … Fulton, R. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. *Nature Genetics*, *46*(4), 409–415.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. C., & Lin, X. (2011). Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *American Journal of Human Genetics*, *89*(1), 82–93.

Zhan, X., Larson, D. E., Wang, C., Koboldt, D. C., Sergeev, Y. V., Fulton, R. S., … Bragg-Gresham, J. (2013). Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nature Genetics*, *45*, 1375–1379.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., … Lander, E. S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, *111*(4), E455–E464.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.