Novel Score Tests to Increase Power in Association Test by Integrating External Controls

by

Yatong Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2021

Doctoral Committee:

Professor Michael Boehnke, Co-Chair
Associate Professor Seunggeun Shawn Lee, Co-Chair
Research Professor Laura Scott
Professor Elizabeth Speliotes

Yatong Li

yatongli@umich.edu

ORCID iD: 0000-0003-2021-4076

Dedication

This dissertation is dedicated to my family, for a lifetime of love and inspiration.

Acknowledgements

Looking back on my education at the University of Michigan, I feel extremely fortunate to have the opportunity to work and learn from the brightest people I know. First and foremost, I am immensely grateful for my advisor, Dr. Seunggeun Shawn Lee. Shawn has always given me his precious time and provide me guidance on my research. He has patiently offered his insights on method development and skills of problem solving. I also feel extremely lucky to have the opportunity to work with Dr. Michael Boenhke. I have always been inspired by his insightful questions and advice on my research projects. His dedication and enthusiasm to science and our community always inspires me, making him a role model of the type of scientist I hope to be a pale shadow of one day. I am very grateful for Dr. Laura Scott and Dr. Elizabeth Speliotes, who gave me great advice on my presentation and provided feedback on my dissertation.

During my time and the University of Michigan, I am grateful to have received function from the Lee's lab including the NIH grant R01-HG008773 and Brain Pool Plus (BP+, Brain Pool+) Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020H1D3A2A03100666). I am similarly grateful for two years of Pre-Doctoral Traineeship in Genome Science funded by the National Human Genome Research Institute of the National Institutes of Health.

Table of Contents

List of Tables

List of Figures

Abstract

Recent advances in genotyping and sequencing technologies have enabled genetic association studies to leverage high-quality genotype or sequence data to identify high-impact variants accounting for a substantial portion of disease risk. The usage of external controls, whose genomes have already been genotyped and are publicly available, could be a cost-effective approach to increase the power of association testing. Various challenges in practice, however, hinder the use of external sources of controls, among which include differences in sequencing platforms, genotype calling procedures, population stratification, etc. Differences in these aspects could lead to a systematic batch effect between genetic data in different studies.

There has been recent effort to integrate external controls while adjusting for possible batch effects, such as the integrating External Controls into Association Test (iECAT). The original iECAT test, however, cannot adjust for covariates such as age, gender, etc. Hence, based on the insight of iECAT, we propose a novel score-based test, iECAT-Score, that allows for covariate adjustment and constructs a shrinkage score statistic that is a weighted sum of the score statistics using exclusively internal samples and uses both internal and external control samples. We show by simulation studies that our method has increased power over the original iECAT while controlling for type I error rates. We present the application of our method to the association studies of age-related macular degeneration (AMD) utilizing data from the International AMD Genomics Consortium (IAMDGC) and Michigan Genomics Initiative (MGI).

The iECAT-Score test has improved power for testing association between a single variant and the disease status, and yet single-variant tests could be underpowered to identify causal rare variants. Hence, in the second project, we extend the single-variant iECAT-Score test to a region-based test, which assesses the combined genetic effect of rare variants within a gene or region. The iECAT-Score region-based test aggregates the single-variant test statistics using a weighted linear or quadratic sum, or a linear combination of both. Through simulation studies

and the application of our method to the rare-variant association studies of AMD from the IAMDGC and UK Biobank data, we show that our proposed iECAT-Score region-based test efficiently identifies disease-associated genes while controlling for type I error rates.

When sequenced data are used in association studies, quality of the genotype calls could influence the performance of the testing methods. The quality of genotype calls is subject to many factors such as read depth, genotype-calling error rates, quality control (QC) pipelines, etc., all of which could result in bias in the estimation of minor allele frequencies (MAFs), leading to more profound batch effect between internal and external control samples. As whole genome/exome sequencing become the design of choice, to address the associated problems using genotyped data, we propose in the third project to integrate the above-mentioned QC parameters utilizing sequencing data. Through the incorporation of these factors, we develop a framework of integrating external controls that is applicable to both genotyped and sequencing data, further honing the statistical methods needed to identify disease-causing variants within the human genome.

Chapter 1 Introduction

1.1 Demand for Improved Power in Genetic Association Studies

Pinpointing genetic variants involved in human disease pathogenesis and trait development has long been of interest. A genome-wide association study (GWAS) using case and controls is a powerful tool to identify loci that are associated with disease through the comparison between the allele frequencies of single-nucleotide polymorphisms (SNPs) between one healthy control group and the disease-carrying case group. GWAS have successfully identified tens of thousands of SNPs associated with diseases and have prompted subsequent functional analyses regarding disease etiology and intervention (Freedman et al., 2011; Gallagher & Chen-Plotkin, 2018).

Early GWAS discoveries primarily consist of common variants with small effect sizes. To uncover more disease associated low-frequency or rare variants explaining the missing heritability, one trend of case-control study has been increasing the sample size (Ioannidis, Thomas, & Daly, 2009). Genotyping or sequencing a large number of samples for individual studies could be expensive. Hence, to improve power in variant discovery via an increased number of controls, investigators have gained interest in either using shared common set of controls for multiple case phenotypes (The Wellcome Trust Case Control Consortium, 2007), or augment control sample size using previously existed study samples (Zhan et al., 2013).

In one seminal study, investigators performed a joint GWA study to study the association between SNPs and seven major diseases using a shared set of controls of ~3,000 samples (Wellcome Trust Case Control Consortium, 2007). This pioneering study demonstrated that using shared or preexisting controls is an effective approach to have adequate sample size for GWAS and allow researchers to focus resources on collecting disease cases. In another association study for risks of age-related macular degeneration (AMD) (Zhan et al., 2013), an initial analysis on the exome sequence data of 2,335 cases and 789 controls revealed no rare coding variants with frequency < 1% that passed experiment-wide significance level. After augmenting the controls to 2,268 through ancestry matching within the Exome Sequencing

Project (Tennessen et al., 2012), however, the investigators were able to identify two large-effect within genes *CFH* and *C3*, before verifying their mechanistic impact through functional studies.

In the above-mentioned studies, investigators underwent careful selection of the (augmented) control samples: (1) cases and shared controls were genotyped on the same platforms and/or had the same filters applied for read depths and call rates (in the case of sequence data); (2) carefully matched the genetic ancestry between external controls and internal samples and excluded samples that showed population stratification with the study population. These procedures tried to use a *bottom-up* strategy to include control samples that have shown little heterogeneity due to technical batch effect or population stratification, thus reducing the likelihood of false discoveries secondary to the heterogeneity.

For most investigators, however, it is extremely challenging to be able to genotype study samples and potential external control samples using the same platforms and pipelines. The original sequence data are often not available for researchers to apply the exact same quality control filters to select additional controls. Even if the sequence data are available, their large file size brings huge computational burden; joint genotyping calling and applying quality control filters may not fully adjust for the technical batch effect if samples are sequenced on different platforms. In addition, selecting control samples that show little to no population stratification with the internal samples might considerably limit the employable controls, diminishing the opportunity of achieving improved power from increased sample size.

Recent advances in genotyping and sequencing technologies have enabled large scale genotype data to be publicly available through various consortia and biobanks. Under this trend, a more realistic scenario for researchers would be to take advantage of the readily available data as external controls to increase the sample size and to assist variant discovery in the study of interest. Thus, it would be preferable to adopt a *top-down* approach to integrate the additional controls without having to jointly re-genotype or manually select which samples to include based on quality control filters and genetic ancestry.

Various challenges in practice, however, hinder the use of external sources of controls, among which include differences in sequencing platforms, genotype calling procedures, population stratification, etc. Differences in these aspects could lead to a systematic batch effect between genetic data in different studies. If external controls are integrated without accounting for the batch effect, false discoveries are likely to happen because of inflated type I error rates,

and thus new analytical methods are required to address the possible technical batch effect and population stratification.

Therefore, in this dissertation, we aim to tackle the following research questions related to integrating external controls in case-control studies using genotype or sequence data:

1. How can we assess the batch effect between samples from different studies using genotype data and integrate the external controls to test for association between single genetic variants and the phenotype? How do we adjust for covariates such as age, gender, and population stratification?
2. How can we achieve improved power for rare variant association test while integrating external control samples?
3. What are the technical factors that lead to the batch effect between studies? Is it possible to use the same testing framework on both genotype and sequence data? Would using genotype dosages improve the performance of the association tests in sequence data?

1.2 Dissertation Chapter Outlines

We address the above-mentioned open research questions in Chapters 2 to 4. There has been recent effort to integrate external controls while adjusting for possible batch effects, such as the integrating External Controls into Association Test (iECAT) (Lee, Kim, & Fuchsberger, 2017). The original iECAT test, however, cannot adjust for covariates such as age, gender, etc. In Chapter 2, we discuss the single-variant iECAT-Score test that tests for association between a single genetic variant and the disease status while integrating external controls with covariate adjustment. In Chapter 3, we develop the iECAT-Score region-based test, which allows the iECAT-Score method to test for association between rare variants within a gene or region. In Chapter 4, we investigated the technical factors that affect the minor allele frequency estimation using hard called genotypes, and propose a strategy to exploit genotype dosages to complete the iECAT-Score framework to be applicable to both genotype and sequence data. We give concluding remarks in Chapter 5. In the sections below, we provide brief outlines for each Chapter 2 to 4 in this dissertation.

1.2.1 Chapter 2: Novel score test to increase power in association test by integrating external controls

Based on the insight of iECAT, we propose a novel score-based test, iECAT-Score, that allows for covariate adjustment and constructs a shrinkage score statistic that is a weighted sum of the score statistics using exclusively internal samples and uses both internal and external control samples. Similar to the original iECAT test, we assess the existence of batch effect at a variant by comparing control samples of internal and external sources. If little batch effect exists, we incorporate data from external control samples to calculate the score statistic; if substantial batch effect exists between internal and external control samples, we exclusively use internal control samples to prevent spurious discoveries. We further extend our method by utilizing the joint distribution of internal score statistic and iECAT score statistic to calculate an omnibus minimum p value. We show by simulation studies that our method has increased power over the original iECAT while controlling for type I error rates. We present the application of our method to the association studies of age-related macular degeneration (AMD) utilizing data from the International AMD Genomics Consortium (IAMDGC) and Michigan Genomics Initiative (MGI).

1.2.2 Chapter 3: Novel score test to increase power in association test by integrating external controls

The iECAT-Score test has superior performance in testing association between a single variant and the disease status, and yet single-variant tests could be underpowered to identify causal rare variants. Hence, in the second project, we extend the single-variant iECAT-Score test to region-based tests, which assess the combined genetic effect of rare variants within a gene or region. This method assesses the systematic batch effect between internal and external samples at each variant and constructs compound shrinkage score statistics to test for the joint genetic effect within a gene or a region, while adjusting for covariates and population stratification. Through simulation studies, we demonstrate that the proposed method controls for type I error rates and improves power in rare-variant tests. The application of the proposed method to the association studies of age-related macular degeneration (AMD) from the International AMD Genomics Consortium (IAMDGC) and UK Biobank revealed novel rare-variant associations in gene *DXO*. Through incorporation of external controls, the iECAT methods offer a powerful suite to identify

disease-associated genetic variants, further shedding light on future directions to investigate roles of rare variants in human diseases.

1.2.3 Chapter 4: Integrating External Controls into Association Analysis Using Sequencing Data

When sequenced data are used in association studies, quality of the genotype calls could influence the performance of the testing methods. The quality of genotype calls is subject to many factors such as read depth, genotype-calling error rates, quality control (QC) pipelines, etc., all of which could result in bias in the estimation of minor allele frequencies (MAFs), leading to more profound batch effect between internal and external control samples. As the whole genome/exome sequencing become the design of choice and to address the associated problems using genotyped data, we propose in the third project to integrate the above-mentioned QC parameters utilizing sequencing data through genotype dosages. Compared to the hard called genotype, which is selected as the most likely genotype given read data, the genotype dosage, is calculated to be the weighted average of all possible genotypes given their respective posterior genotype likelihood and accounts for the uncertainty about the true genotype. Using genotype dosages offer consistent minor allele frequency estimation, controls type I error rate, and improves power for association discovery especially in rare variants; opting to use genotype dosages also preserves more variants available for association test. Through the incorporation of these factors, we develop a complete framework of integrating external controls that is applicable to both genotyped and sequencing data, further honing the statistical methods needed to identify disease-causing variants within the human genome.

Chapter 2 Novel Score Test to Increase Power in Association Test by Integrating External Controls

2.1 Introduction

In case-control studies, large numbers of control samples are often necessary to achieve adequate power to identify disease-susceptible genetic variants. Genotyping large control samples, however, are expensive. A cost-effective strategy could be utilizing publicly available genotyped data and incorporating such external control samples. This approach allows for efficient use of existing resources and boosts the power of association testing by increasing the control sample size.

Various challenges in practice, however, hinder the use of external sources of controls, among which include differences in sequencing platforms, genotype calling procedures, population stratification, etc. Differences in these aspects could lead to a systematic batch effect between genotyped data in different studies. If the batch effect is unaccounted for, incorporating external sources of data could result in undesired type I error inflation and erroneous discoveries. Several developments have been made recently to address the systematic differences between genotyped data of internal and external sources. Derkach et al. (Derkach et al., 2014) developed a score test that replaces called genotype with expected genotype using probabilities of genotypes given sequencing reads, accounting for differential read depths between studies. Building on Derkach et al., Chen and Lin (Chen & Lin, 2018) proposed regression calibration (RC)-based and maximum-likelihood (ML)-based methods to account for differential sequencing errors between cases and controls. The RC method extended Derkach's method to allow for non-confounding covariate adjustment; the ML method enables the parameters of interest to be identifiable. Hu et al. (Hu, Liao, Johnston, Allen, & Satten, 2016) proposed a likelihood-based method that models sequencing reads, which involves first estimating single nucleotide variant (SNV) locations using read data and then using a burden test with the bootstrap procedure to

assess the significance of the association between an SNV and a trait. Although the aforementioned methods control type I error rates in most settings, they all require genotype probabilities or sequence reads data to be available. Computing genotype probabilities and storing sequence reads data, however, could be challenging and expensive for large scale studies, and thus are often unavailable. Hendricks et al. (Hendricks et al., 2018) proposed ProxECAT which focuses on integrating external controls to estimate enrichment of rare variants using allele counts. This method does not include internal control samples in the analyses, which potentially limits the power of the association test.

Lee et al. (Lee et al., 2017) proposed a method, integrating External Controls into Association Test (iECAT), that uses allele counts from external studies to assess the batch effect between internal and external studies. The method examines batch effect by comparing odds ratio estimates of alleles using internal control samples and using combined control samples from internal and external studies. The degree to which the odds ratios differ indicates the amount of batch effect that exists between the two studies. If little batch effect exists from this comparison, control samples from the external study are included to increase sample size; otherwise, external control samples are not used to avoid type I error inflation. The single variant test of iECAT uses an empirical Bayesian-type shrinkage estimator, which is a weighted sum of test statistics using exclusively internal samples and using both internal and external control samples. This iECAT method can control the type I error rates while improving power for association testing.

The original iECAT test, however, cannot adjust for covariates such as age, gender, etc. Hence, based on the insight of iECAT, we propose a novel score-based test that allows for covariate adjustment. Score tests are not only computationally efficient but are more stable than the tests using odds ratio estimated from allele counts. Recent improvements of score tests by applying the Saddlepoint approximation (SPA) (Dey, Schmidt, Abecasis, & Lee, 2017) and Efficient resampling (ER) (Lee, Fuchsberger, Kim, & Scott, 2015) methods also allow for controlling type I error in the scenario of case-control imbalance and low minor allele count (MAC). We construct a shrinkage score statistic, which is a weighted sum of the score statistics using exclusively internal samples and using both internal and external control samples. Similar to the original iECAT test, we assess the existence of batch effect at a variant by comparing control samples of internal and external sources. If little batch effect exists, we incorporate data

from external control samples to calculate the score statistic; if substantial batch effect exists between internal and external control samples, we exclusively use internal control samples to prevent spurious discoveries. We further extend our method by utilizing the joint distribution of internal score statistic and iECAT score statistic to calculate an omnibus minimum *p* value.

Our method controls type I error rates while increasing samples from external studies to improve power for association tests. In Section 2.2.1, we present the model for a single variant test of genetic effect in case-control studies and propose the iECAT-Score methods that allow for covariate adjustment when integrating external control samples in case-control association testing. In Section 2.2.2, we present the simulation studies to evaluate the type I error rates and power of our proposed methods. In Section 2.2.3, we present the application of our methods to the association studies of age-related macular degeneration (AMD) combining data from the International AMD Genomics Consortium (IAMDGC) and Michigan Genomics Initiative (MGI). In Section 2.3, we present the results from simulation studies and data analyses of the proposed methods and compare their performance with the method that does not integrate external control samples.

2.2 Materials and Methods

2.2.1 Single-variant association test

The iECAT score test is a shrinkage estimation-based test for the variant effect on a phenotype of case or control, combining external control samples. For each variant, the iECAT score test analytically assesses the batch effect between internal and external controls and determines the weight of external control samples to be combined, which is inversely correlated to the batch effect. The test then calculates a *p* value for association using the weighted sum of internal and external samples while adjusting for covariates.

Single Variant Score Test of Genetic Effect

We first consider the internal study only with sample size *n*. For the *i*th subject, let $y_i = 0/1$ be the dichotomous phenotype for control/case, $\boldsymbol{X}_i = \left(X_{i1}, X_{i2}, .., X_{ip}\right)^T$ the covariates, $\boldsymbol{G} = (G_1, G_2, .., G_n)^T$ the genotypes at a variant for *n* subjects ($G_i = 0, 1, 2$ represent 0, 1, 2 copies of

the minor allele). We relate the phenotype Y_i to the covariate $\boldsymbol{X}_i$, and the genotype G_i using the logistic model

$$logit[\Pr(Y_i = 1 \mid \boldsymbol{X}_i, G_i)] = \boldsymbol{X}_i^T \boldsymbol{\alpha} + G_i \beta \qquad (2.1)$$

where the phenotype Y_i follow a Bernoulli distribution. In this equation, $\boldsymbol{\alpha}$ is an $p \times 1$ vector of coefficients for p covariates including an intercept, and β is the genotype effect at the variant of interest. Assessing whether the association exists between the phenotype Y_i and the genotype at a variant is equivalent to testing $H_0: \beta = 0$ in Equation (1).

Let $\boldsymbol{\mu} = \{\mu_i\} = \{\Pr(Y_i = 1 \mid X_i)\}$ under H_0, and $\hat{\mu}_i$ is the maximum likelihood estimate of μ_i. The score test statistic is

$$S = \boldsymbol{G}^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$$

Under the null hypothesis of no genetic effect, $E(S) = 0$ and $Var(S) = \sum_{i=1}^{n} \tilde{G}_i^2 \hat{\mu}_i(1 - \hat{\mu}_i))$, where $\widetilde{\boldsymbol{G}} = \{\tilde{G}_i\} = \boldsymbol{G} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{G}$ is the covariate-adjusted genotype vector, and $\boldsymbol{V} = diag\{\hat{\mu}_i(1 - \hat{\mu}_i)\}$. Then $\frac{S^2}{Var(S)}$ asymptotically follows χ_1^2, and a p value can be obtained as $p = P(\chi_1^2 > \frac{S^2}{Var(S)})$.

Integrating External Control Samples in Score Test

In addition to the control samples from the study, which we refer to as internal control samples, we introduce external control samples to increase the sample size. Let n_1^I, n_0^I, n_0^E denote the sample sizes of internal cases, internal controls, and external controls, respectively. Let $Y_i = 0/1(i = 1,2, \ldots, n_1^I + n_0^I + n_0^E)$ be the dichotomous phenotype for control/case. When only the internal samples are used to test the association between the variant and the phenotype, the score statistic is given by $S_{int} = \boldsymbol{G}_{int}^T(\boldsymbol{Y}_{int} - \widehat{\boldsymbol{\mu}}_{int})$. In this equation, $\boldsymbol{G}_{int} = \left(G_{int,1}, G_{int,2}, .., G_{int,(n_1^I + n_0^I)}\right)^T$ is the vector of genotypes of internal samples; similarly, $\boldsymbol{Y}_{int} = \left(Y_{int,1}, Y_{int,2}, .., Y_{int,(n_1^I + n_0^I)}\right)^T$ is the vector of phenotypes of internal samples, and $\widehat{\boldsymbol{\mu}}_{int} = \left(\hat{\mu}_{int,1}, \hat{\mu}_{int,2}, .., \hat{\mu}_{int,(n_1^I + n_0^I)}\right)^T$ is the vector of maximum likelihood estimate of $\boldsymbol{\mu}_{int}$ under the null logistic regression model of no genetic effect built using internal samples only. When external control samples are included assuming no systematic differences between internal and

external studies, a similar score statistic can be constructed by $S_{all} = \boldsymbol{G}_{all}^T(\boldsymbol{Y}_{all} - \widehat{\boldsymbol{\mu}}_{all})$. In this equation, $\boldsymbol{G}_{all}, \boldsymbol{Y}_{all}$ and $\widehat{\boldsymbol{\mu}}_{all}$ are vectors of length $n_1^I + n_0^I + n_0^E$, denoting genotypes, phenotypes, and expected mean outcome under a null model built of combined internal and external samples.

The systematic differences between internal and external studies, however, are likely to exist, in which case distributions of genotypes are different between internal and external control samples after adjusting for covariates. To quantify the extent to which the batch effect exists between the two studies, we test for a relationship between the genetic variant and whether a control sample belongs to the internal or external study while adjusting for covariates. Considering the internal controls and external controls with sample sizes n_0^I and n_0^E, respectively, let $\widetilde{\boldsymbol{Y}}_{IvE} = (\tilde{Y}_j) = 0/1\ (j = 1, 2, \dots, n_0^I + n_0^E)$ represent a control sample belonging to the external/internal study, $\boldsymbol{X}_j = \left(X_{j1}, X_{j2}, .., X_{jp}\right)^T$ the covariates for the *j*th subject, and $\boldsymbol{G}_{IvE} = \left(G_1, G_2, .., G_{n_0^I + n_0^E}\right)^T$. The genotypes at a variant for $n_0^I + n_0^E$ control samples. To relate phenotype $\tilde{Y}_j$ to the covariate $\boldsymbol{X}_j$, and the genotype G_j, we consider the logistic model

$$logit[\Pr(\tilde{Y}_j = 1 \mid X_j, G_j)] = \boldsymbol{X}_j^T\, \tilde{\boldsymbol{\alpha}} + G_j^T\, \tilde{\beta}$$

If we test the null hypothesis of no batch effect between the internal and external control samples, a score test statistic can be constructed as $S_{IvE} = \boldsymbol{G}_{IvE}^T(\widetilde{\boldsymbol{Y}}_{IvE}^T - \widetilde{\boldsymbol{\mu}}_{IvE})$, where $\boldsymbol{\mu}_{IvE} = (\mu_{IvE,j}) = \left(\Pr(\tilde{Y}_j = 1 \mid X_j)\right)$ and $\widetilde{\mu}_{IvE,j}$ is the maximum likelihood estimate of $\mu_{IvE,j}$.

When there is no genetic effect on the phenotypes and no batch effect exists, $E(S_{int}) = E(S_{all}) = E(S_{IvE}) = 0$. Under such condition, we assume that $(S_{int}, S_{all}, S_{IvE})^T \sim N_3(\boldsymbol{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$ (Supplementary Materials **2.5.1.1**). To test the hypothesis of no genetic effect using both internal and external control samples, we propose a compound score statistic

$$S_w = a\tau S_{int} + (1 - \tau)S_{all} \qquad (2.2)$$

where $a = \frac{(n_1^I+n_0^I)(n_1^I n_0^I+n_1^I n_0^E)}{n_1^I n_0^I(n_1^I+n_0^I+n_0^E)}$ (Supplementary Materials **2.5.1.2**) and $\tau = \frac{\tau_1}{1+\tau_1}$ with $\tau_1 = \frac{S_{IvE}^2}{Var(S_{IvE})}$. Under the null hypothesis of no genetic effect, $E(S_w) = 0$ and $Var(S_w)$ can be calculated using the delta method. The asymptotic distribution of $\frac{S_w^2}{Var(S_w)}$ approximately follows a χ_1^2. Thus, a p value can be approximated by $P(\chi_1^2 > \frac{S_w^2}{Var(S_w)})$.

When minor allele frequencies (MAFs) of external controls are in between those of internal cases and internal controls, it is implied that we could make $\tau = 0$, i.e., include all of the external control samples without inflating type I error rates, following Lee and others (Lee et al., 2017).

Calibrating Single-Variant Test Using Saddlepoint Approximation and Efficient Resampling Methods

The single-variant score test approximates the null distribution by a normal distribution. The variance estimates based on such asymptotic tests behave well for common variants and balanced case-control studies. When allele frequency is extremely low resulting from low minor allele count (MAC), or when the case-control ratio is unbalanced, the underlying distribution of test statistics could be discrete or highly skewed. In such cases, the traditional asymptotic-based score test performs poorly with conservative or anti-conservative results.

To account for scenarios of low MAC and unbalanced case-control ratio, we apply a recently developed robust approach to iECAT score by properly calibrating the variance estimates of score statistics (Zhao et al., 2020). Specifically, we update the variance estimates of $Var(S_{int})$, $Var(S_{all})$ and $Var(S_{IvE})$ in $\boldsymbol{\Sigma}$ corresponding to the scores S_{int}, S_{all} and S_{IvE} by applying the Saddlepoint approximation (SPA) method (Dey et al., 2017) or Efficient resampling (ER) method (Lee et al., 2015). When the score estimates lie far from mean (zero), we apply the SPA method to obtain the $\tilde{p}$ values; when the MAC is low (MAC<10) either in internal samples, combined samples, or control samples, we apply the ER method to obtain the $\tilde{p}$ values. The updated variance $\widetilde{Var}(S_{(.)})$ is then derived from $\frac{S_{(.)}}{\widetilde{Var}(S_{(.)})} \sim \chi^2(\tilde{p})$, where $S(.)$ represents S_{int}, S_{all} or S_{IvE}. A calibrated variance estimate $\widetilde{Var}(S_w)$ of S_w is thus obtained by applying the delta

method to the updated covariance matrix $\widetilde{\boldsymbol{\Sigma}}$. The p value of the robust iECAT score single-variant test is approximated by $P(\chi_1^2 > \frac{S_w^2}{\widehat{Var}(S_w)})$.

Minimum P-value Based on Combination of S_w and S_{int}

When the variance estimate of S_w is large, the resulting p value can be larger than that from the test of using internal samples only. Hence, we calculate a minimum p value (iECAT-Score minP) following Conneely and Boehnke (Conneely & Boehnke, 2007). Specifically, (S_w, S_{int}) jointly follow a multivariate normal distribution. The minimum p value is calculated as the probability of observing one or both the p values as small as the smaller one of the two under the null hypothesis of no association (Supplementary Materials **2.5.1.3**).

2.2.2 Type I error and power simulations

We carried out simulation studies under a range of scenarios to evaluate the performance of the proposed iECAT score test with regard to type I error rates and power. For both type I error and power simulations, we generated binary phenotypes of case/control from the logistic regression model:

$$logit[\Pr(Y = 1 \mid \boldsymbol{X}, G)] = \alpha_0 + 0.5X_1 + 0.5X_2 + \beta G$$

where X_1 was a continuous covariate generated from a standard normal distribution, X_2 was a dichotomous covariate with the probability of 0.5 being 0 or 1, α_0 was chosen such that the disease prevalence was 0.05, G is the genotype at the variant of interest generated from a binomial (2, MAF) distribution, and β is the effect size of the variant. MAF was sampled from the MAF distribution in the MGI data.

To mimic the batch effect between internal and external genetic samples, we assumed that 3% of the variants were subject to different MAFs in internal and external control samples. For such variants, we set the MAFs of the external controls to be randomly generated from $Uniform(0.1 \times q, 4 \times q)$, where q is the MAF of corresponding variants in the internal samples.

For both type I error and power simulations, we considered two combinations of case-control ratios ($n_1^I : n_0^I : n_0^E$): (1) 5,000: 5,000: 10,000; (2) 6,667: 3,333: 10,000. In type one error simulations, $\beta = 0$. We generated 10^9 datasets to evaluate type I error rates at 5×10^{-5} and 5×10^{-8} level. To save computation time, we generated 10^7 sets of genotypes and phenotypes,

and resampled the disease phenotypes of internal samples 100 times for each set, while keeping other data fixed. In power simulations, β was set to values from a grid of $\log(1.1), \log(1.15), \ldots, \log(1.5)$, representing the odds ratio (OR) of 1.1, 1.15, …, 1.5 for the causal variant. We generated 100,000 data sets in each setting of effect size and case-control ratio to evaluate empirical power at the significance level of 5×10^{-8}.

2.2.3 Real data analysis

We applied our proposed method to genotype data from the International AMD Genomics Consortium (IAMDGC) (Fritsche et al., 2015) downloaded from dbGaP (phs001039.v1.p1). The IAMDGC dataset consists of 17,286 cases and 14,373 controls. As external controls, we used 40,971 samples from Michigan Genomics Initiative (MGI). The MGI samples consist of individuals who received surgical procedures at the University of Michigan Health System. A broad range of clinical phenotypes was collected from the MGI samples and the individuals were genotyped on the Illumina HumanCoreExome v1.12.1 array on >500,000 variants. To compare the performance of our proposed iECAT-Score methods with the method of the sole usage of internal samples, we performed analyses on the genotype data of 316,822 overlapping variants between the AMD and MGI studies. For both studies, the samples used in our analyses are of European ancestry. We applied the Fruposa software (Zhang, Dey, & Lee, 2020) with the 1000 Genomes reference (The 1000 Genomes Project Consortium, 2015) to obtain population structure. We assessed the relationship between the origin of control samples and genetic variants to examine the existence of batch effect between internal and external controls. We tested for association between the disease status of age-related macular degeneration (AMD) and single genetic variants that are shared by IAMDGC and MGI data sets, adjusting for age, sex, and the first 10 principal components.

2.3 Results

2.3.1 Type I error and power simulations

Empirical type I error rates of the proposed methods are given in **Table 2.1** at the significance level of 5×10^{-5} and 5×10^{-8}. The results show that various versions of the iECAT-Score

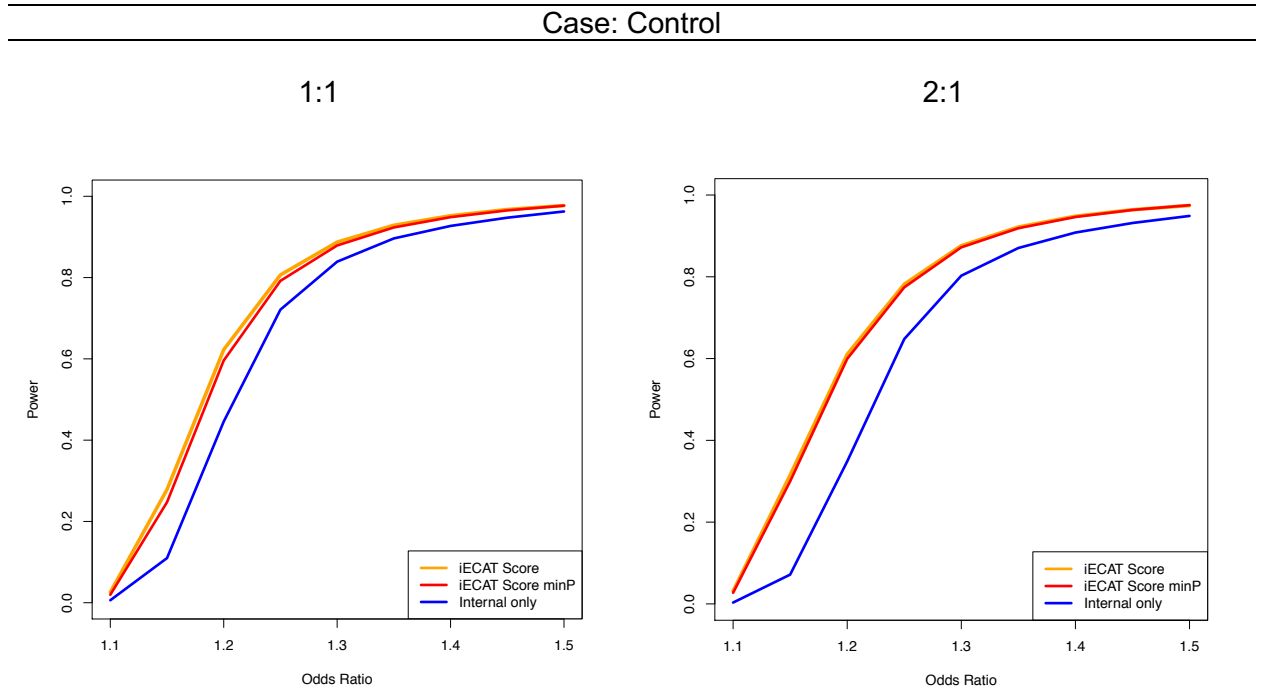
method controlled type I error at both nominal levels. Type I error rates were well controlled when the internal control samples were exclusively used, but they were erroneously inflated when the external control samples were combined without considering the batch effects.

Table 2.1: Empirical type I error rates of iECAT-Score tests.
Empirical type I error rates of score tests using internal samples exclusively, using combined samples without adjusting for batch effect, and various versions of the iECAT-Score method, under different scenarios. Type I error rates were estimated based on 10^9 simulations. 3% variants are simulated to have different minor allele frequencies between internal and external controls.

Sample size (Internal: external)	Internal case:control	Level alpha	Internal only	iECAT-Score	iECAT-Score minP	Combined internal and external samples
10000:10000	1:1	$5x10^{-5}$	5.00e-05	3.36e-05	3.91e-05	2.57e-02
		$1x10^{-6}$	1.00e-06	7.10e-07	8.00e-07	2.48e-02
		$5x10^{-8}$	6.00e-08	3.20e-08	3.80e-08	2.42e-02
10000:10000	2:1	$5x10^{-5}$	5.00e-05	3.43e-05	4.15e-05	2.68e-02
		$1x10^{-6}$	1.01e-06	8.68e-07	9.62e-07	2.60e-02
		$5x10^{-8}$	4.00e-08	8.40e-08	5.80e-08	2.56e-02

We compared the power of iECAT-Score methods and method using internal controls exclusively to assess genetic association at empirical alpha levels which provided type I error rates 5×10^{-8}. The empirical alpha levels were estimated from type I error simulation studies. Since the method using all external control samples had severely inflated type I error rates, we did not include this method in the power comparison. **Figure 2.1** compares powers of different methods at changing effect size, represented by the OR of 1.1, 1.15, …, 1.5 on the x-axis. The two panels of the figure show such comparison of different combinations of case-control ratios $(n_1^I : n_0^I : n_0^E)$: (1) 5,000: 5,000: 10,000; (2) 6,667: 3,333: 10,000. The results show that all versions of our proposed iECAT-Score method had improved power over the method that used exclusively internal control samples.

Figure 2.1: Power plot of the iECAT-Score methods.
Empirical power comparisons using 10,000 internal samples and 10,000 external control samples, with internal case and control ratios 1:1 and 2:1. Lines represent empirical powers at $\alpha = 5 \times 10^{-8}$. The effect size of the causal variant was $\beta = \log(Odds\ Ratio)$.



2.3.2 Application to Age-Related Macular Degeneration (AMD) Data

We applied our iECAT-Score methods to the analysis of AMD from IAMDGC, using samples from MGI as external controls. The female samples consist of 41.29%, 43.99%, and 45.90% in samples of internal cases, internal controls, and external controls, respectively (**Table 2.2**). The percentages of females did not vary significantly in different case groups. The mean age of cases and controls in the IAMDGC dataset were 75.86 years and 70.08 years, respectively, showing that samples with AMD tended to be older. Samples of the external controls were on average younger than those from IAMDGC, with a mean age of 53.31 years.

Table 2.2: Descriptive statistics of study subjects from internal (IAMDGC) and external (MGI) studies.
Shown in the table are the sample sizes, the number (percentage) of female samples, and mean (standard deviation of sample age in years in IAMDGC and MGI data.

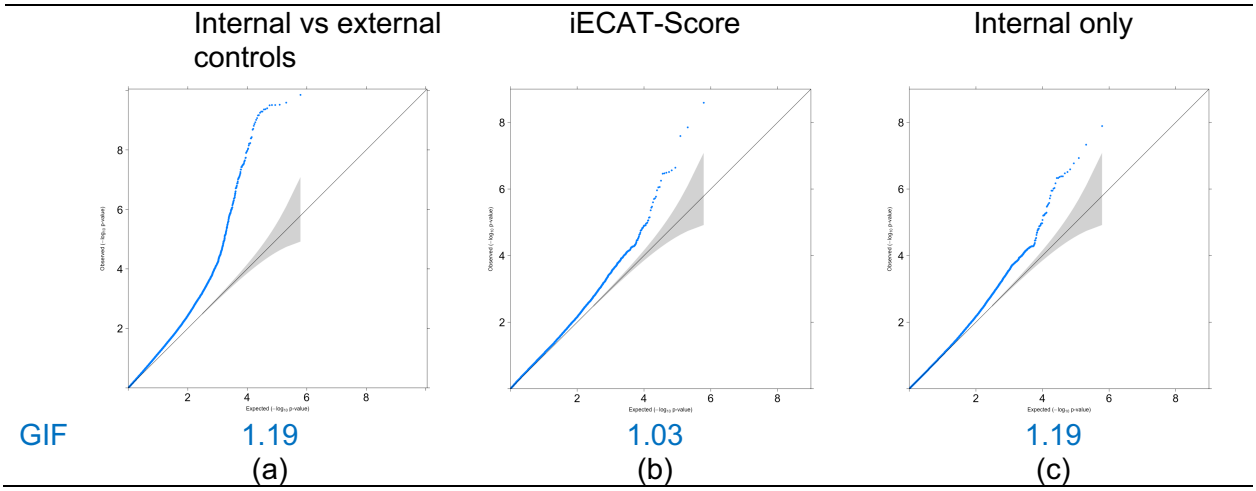
Study	Sample Size N		Female N (%)			Age (yrs) Mean (SD)		
	Cases	**Controls**	**Cases**	**Controls**	**Total**	**Cases**	**Controls**	**Total**
IAMDGC (internal)	17286	14373	7137 (41.29)	6322 (43.99)	13459 (42.51)	75.86 (8.10)	70.08 (9.71)	73.24 (9.32)
MGI (external)		26598		12479 (46.92)	12479 (46.92)		53.31 (15.88)	53.31 (15.88)
Total	40971	40971	7137 (41.29)	18801 (45.89)	25938 (44.52)	75.86 (8.10)	59.19 (16.15)	64.14 (16.15)

The QQ plot from testing the relationship between the origin of control samples and genetic variants is presented in **Figure 2.2(a)**. A significant deviation from the 45-degree line indicates that there exist systematic differences between internal and external control data. We tested for association between the disease status of age-related macular degeneration (AMD) and single genetic variants, adjusting for age, sex, and the first 10 principal components. The QQ plots from the tests integrating external control samples using the iECAT-Score method and using internal samples exclusively are shown in **Figure 2.2(b)** and **2.2(c)**. We observe the similarity between the patterns of the two QQ plots, which are both close to the 45-degree line and show that our method of iECAT-Score controlled for type I error rates in this analysis.

Table 2.3 presents the top variants showing genome-wide significance. Of the 13 presented variants, 11 variants had smaller *p* values resulted from the iECAT-Score method than using internal samples only, implying that the iECAT-Score method can have a higher power of detecting associations. The AMD-associated genes revealed by the iECAT-Score were consistent with some well-known associations such as *CFH* (Maller et al., 2006), *C2* (Gold et al., 2006), *CFI* (Fagerness et al., 2008; Helgason et al., 2013; Seddon et al., 2013; Zhan et al., 2013), *RAD51B* (Fritsche et al., 2013; Seddon, Reynolds, Yu, & Rosner, 2014) and *C3* (Maller et al., 2007; Yates, 2007), indicating the validity of our method. In particular, the association locus rs3784099 of gene *RAD51B* was revealed by applying the iECAT-Score method (*p* value: 4.63e-08), but did not reach the significance level of 5e-08 (*p* value: 1.10e-06) with the sole usage of internal samples from the IAMDGC dataset used by the GWAS study (Fritsche et al., 2015).

Although the AMD GWAS reported association of gene *RAD51B* at a different locus rs61985136, the reported locus is an imputed variant among ~11.8 million imputed variants analyzed in the study, and thus not included in our analyses. Most of our index variants found by the iECAT-Score method were either identical or in close LD with those found in the GWAS study.

Figure 2.2: QQ plots for analysis of age-related macular degeneration (AMD).
QQ plots for analysis of AMD from the internal study of International AMD Genomics Consortium (IAMDGC) and external control study of Michigan Genomics Initiative (MGI). For better visualization, the maximum of x and y axes in the plots are set to be 9, corresponding to *p* values of 1e-09. The GIF shows the genomic inflation factor calculated at the median of test statistics.



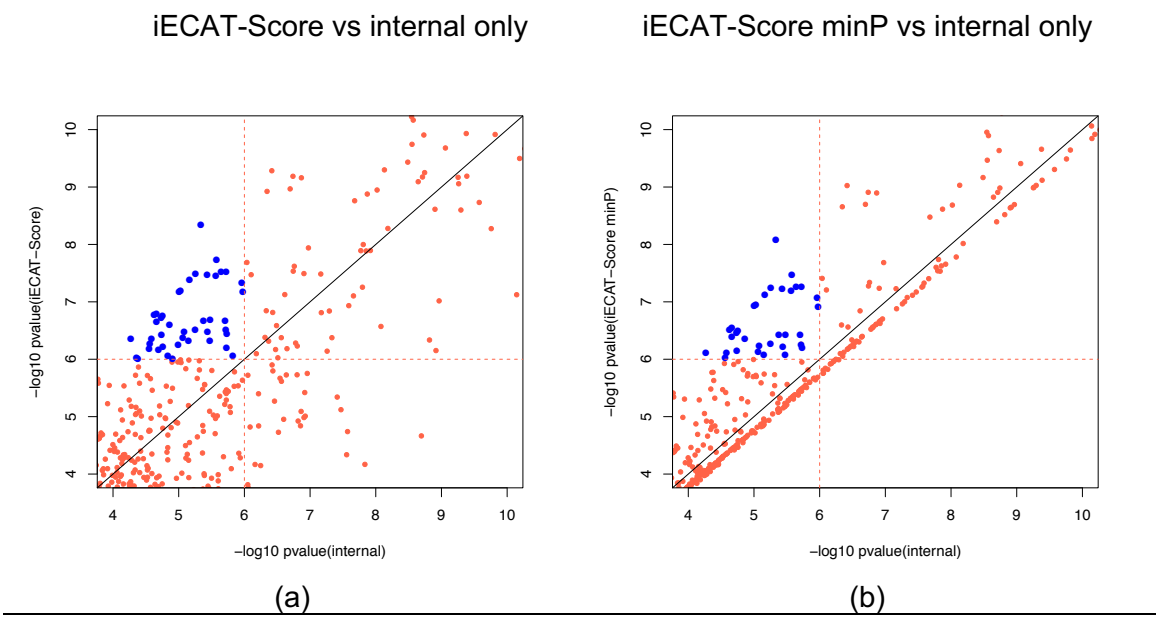
We compared *p* values (in log 10 scale) from analyses using iECAT-Score, iECAT-Score minP, and the method using internal samples only (**Figure 2.3**). Shown in blue color are variants that were not significant at the 1e-06 level by solely using internal samples, but showed signals of association when the iECAT-Score method was used. Hence, the iECAT-Score method could improve power in detecting associations for variants of borderline significance. Interestingly there were substantial numbers of variants in which the internal sample-only approach produced smaller *p* values than iECAT-Score. iECAT-Score minP test addressed this issue by leveraging the minimum *p* value between internal sample-only and iECAT-Score *p* values, so producing slightly less significant *p* value than the internal sample only when it has the smallest *p* value.

Table 2.3: Identification of variants showing significance (5e-8 level) based on iECAT-Score.
Shown are allele frequencies, effect sizes, p values of analyses from exclusive usage of internal samples and using the iECAT-Score method.

Signal Number	Index variant			Major / minor allele	Minor allele frequency					Pval.IvE	Odds Ratio	*p* values		
	Name	dbSNP ID	Chr:Pos		Internal		External	Combined				Internal only	iECAT-Score	iECAT-Score minP
					Case	Control	Control	Case	Control					
1	*CFH*	rs800292	1:196642233	G/A	.138	.238	.232	.138	.234	.253	.525	1.47e-212	5.17e-240	1.03e-239
2	*CFI*	rs10033900	4:110659067	C/T	.509	.476	.486	.509	.483	.079	1.11	1.89e-15	1.69e-16	3.29e-16
3	*C9*	rs34882957	5:39331894	G/A	.015	.0089	.0096	.015	.0093	.848	1.59	2.31e-11	5.26e-15	9.89e-15
4	*C2/CFB/CKIV2L*	rs429608	6:31930462	G/A	.098	.149	.144	.098	.146	.115	.638	3.60e-77	4.22e-83	9.44e-83
5	*VEGFA*	rs4711741	6:43828582	T/C	.472	.501	.498	.472	.499	.386	.897	2.03e-12	3.84e-14	7.41e-14
6	*ARMS2/HTRA1*	rs714816	10:124256345	G/A	.419	.336	.340	.419	.339	.795	1.41	9.15e-95	1.33e-121	2.65e-121
7	*B3GALTL*	rs11147458	13:31823239	A/G	.285	.305	.304	.285	.304	.265	0.91	2.59e-09	3.06e-12	5.81e-12
8	*RAD51B*	rs3784099	14:68749927	G/A	.275	.293	.287	.275	.289	.928	.934	1.10e-06	4.63e-08	8.35e-08
9	*LIPC*	rs415799	15:58690754	G/A	.492	.468	.477	.492	.474	.318	1.08	6.07e-11	5.30e-12	1.01e-11
10	*CETP*	rs247616	16:56989590	C/T	.351	.319	.322	.351	.321	.968	1.15	7.79e-17	1.40e-20	2.75e-20
11	*C3*	19:6718146	19:6718146	T/G	.012	.0040	.0042	.012	.0041	.326	2.89	4.22e-25	5.43e-19	8.44e-25
12	*APOE*	rs769449	19:45410002	G/A	.084	.111	.112	.084	.111	.032	.735	1.68e-20	5.20e-14	3.35e-20
13	*SLC16A8*	rs8135665	22:38476276	C/T	.215	.195	.202	.215	.200	.073	1.10	1.40e-08	1.30e-08	2.37e-08

Pval.IvE is the *p* value from comparing internal and external control samples, using the indicator of the source of control sample as outcome. The odds ratio was calculated as the odds of the disease of the minor allele as compared to the major allele, assuming an additive model.

Figure 2.3: Comparison of *p* values (in -log 10 scale) from analyses of age-related macular degeneration data using the iECAT-Score methods.
Panel (a): -log 10 scaled *p* values using the iECAT-Score method vs. using internal samples only; panel (b): -log 10 scaled *p* values using the iECAT-Score minP method vs. using internal samples only.



2.4 Discussion

Utilizing publicly available sequenced or genotyped data as external controls is a cost-effective approach to increase statistical power in case-control studies. In this paper, we proposed the score-based test, iECAT-Score, which allows for the integration of external sources of genotyped data into association testing while adjusting for systematic batch effects. Compared to the original iECAT, our method is not only computationally efficient, but is able to adjust for covariates such as age, sex, and population stratification.

The simulation studies showed that iECAT-Score methods control for type I error rates and have improved power for association testing compared to the sole usage of internal control samples. Analysis of the AMD from IAMDGC and MGI revealed that iECAT-Score reaches a resembling level of type I error control as the method that uses internal samples exclusively. With the integration of external controls and adjusting for batch effects, iECAT-Score can improve power for association discovery.

In our simulation studies, we mimicked the batch effect by setting different MAFs between internal and external control samples. An alternative to such a batch effect mechanism is assuming different genotyping error rates. In supplementary materials **Table S2.1** and **Figure S2.1**, we present additional simulation results by setting certain genotyping error rate in external control samples, where each allele could be mis-genotyped as the alternative allele, provided the variant is prone to batch effect. Regardless of the mechanisms through which batch effects occurred, our method of iECAT-Score controlled for type I error rates. In the simulation studies with either batch effect mechanism, we assumed that three percent of variants shared between internal and external studies were subject to batch effect. We show in additional simulation results in supplementary materials **Table S2.2** and **Figure S2.2**, that when variants subjecting to batch effect were increased to 12 percent, the presented iECAT-Score method still efficiently controlled type I error rates and increased power. We compared the level of batch effect between the above-mentioned simulation settings and real data analyses of the IAMDGC and MGI. We calculated the genomic inflation factors from the comparisons between internal controls and external controls, in real data, and simulated data when batch effect existed in three and twelve percent of variants. The results in supplementary materials **Table S2.3** show that the level of batch effect observed in real data between IAMDGC and MGI was less than that in the simulation studies with twelve percentage of batch-effect-prone variants, and yet iECAT-Score still had superior performance in such an extreme scenario.

Our data analysis results revealed that it is possible for iECAT-Score to yield a larger *p* value for association at a variant than when only the internal controls are used, such as variant 19:6718146 near gene *C3* and variant 19:45410002 near gene *APOE* (**Table 2.3**). This was due to the large variance estimate of the iECAT-Score statistic, resulting in an overall weaker signal than the test using internal samples. To address such a scenario, one possible strategy is to calculate a two-sided minimum *p* value following Conneely and Boenhke (Conneely & Boehnke, 2007). This method calculates the probability of observing at least one *p* value as small as the observed smaller *p* value, while considering a correlation between the two tests. Similar to iECAT-Score, the minimum *p* value method combining iECAT-Score and internal control methods controlled for type I error rates and improved power compared with the exclusive use of internal samples.

In most cases, integrating external controls increases power by increasing the sample size. When MAF of external controls is closer to MAF of internal cases than to MAF of internal controls, the improvement in power could be weakened from the increase in noise. We performed both analytical calculations (Supplementary Materials **2.5.1.4**) and simulations to compare the power of the different methods when MAF of external control samples varied relative to MAFs of internal cases and internal controls (Supplementary **Figure S2.3**). Both versions of the iECAT-Score method achieved the greatest power when MAF of external controls was close to MAF of internal controls. When MAF of external controls was close to that of internal cases, iECAT-Score showed no improvement in power as a result of added noise and lowered power using the combined control samples. The iECAT-Score minP method, on the other hand, had comparable or improved power across all values of MAF in external controls, as compared to exclusively using internal samples.

Our method adjusts for population structure by including genetic principal component eigenvectors as covariates. Results from additional simulation studies show that when mild to moderate level of population stratification existed, our methods, the iECAT-Score minP method especially, controlled for type I error rates (Supplementary **Figure S2.4a**). When samples of more extreme population stratification were used, there could be mild inflation in type I errors (Supplementary **Figure S2.4b**). In such scenarios, we could apply ancestry matching (Guan, Liang, Boehnke, & Abecasis, 2009; Wang et al., 2014; Zhang et al., 2020) to select external control samples. Thus, despite the advantage of our methods, we recommend a careful examination of genetic backgrounds in the target populations, and possible ancestry matching before making the decision to integrate subjects as external controls, so as to avoid false positive discoveries as a result of extreme population stratification.

In addition to the possible population stratification, it would not be unusual that the available external study samples are different from the internal study samples in distributions of age, gender, and/or other variables that are not confounded by genetic background. In the analyses of the AMD from IAMDGC and MGI, for instance, external control samples in MGI tended to be younger than samples in the internal samples. We carried out additional simulation studies to assess the performance of iECAT-Score methods, where we assumed internal and external samples to have different distributions of covariates, which contributed to the disease risk. Our results (Supplementary **Table S2.4**, **Figure S2.5**) show that iECAT-Score had similarly

controlled type I error rates and increased statistical power, allowing for different distributions of between internal and external samples.

Since we applied the methods of saddlepoint approximation and efficient resampling, our method could handle small allele counts (single digit), as long as the following conditions are satisfied: (1) total minor allele counts in internal controls and external controls combined are greater than zero, i.e. there is no monomorphism in the combined control samples; (2) total minor allele counts in internal samples are greater than zero, i.e. there is no monomorphism in the internal samples. In simulation studies and real data analysis, we used all the variants that satisfied the conditions, and type I error rates were well controlled. When there exists monomorphism in internal samples or combined control samples, iECAT-Score methods cannot be applied, as it is impossible either to assess association exclusively using internal samples or to assess the batch effect between internal controls and external controls.

There are some scenarios when using external studies could be challenging, even with the assistance of iECAT-Score methods. (1) When the case-to-control ratio of internal study is small (less than 1:2), the increase of power in association testing by integrating external controls would be limited, despite that iECAT-Score controls type I error rates (Supplementary **Table S2.5**, **Figure S2.6**). (2) When the sample size of the internal controls is small compared to that of external study, iECAT-Score methods could result in type I error inflation at a low nominal level (Supplementary **Table S2.6**). Such inflation is due to a lack of confidence while comparing internal and external control samples, on which our method is dependent.

One additional challenge when using external study samples as controls is the potential misclassification of disease status. The phenotype of interest may not be recorded in the external study, leaving the possibility that some samples used as controls may be in fact cases. We expect that misclassification of cases as controls alleviates the overall signal and should not result in false positive discoveries. The results of our simulation studies (Supplementary **Table S2.7**, Supplementary **Figure S2.7**) were in line with this reasoning: with the presence of misclassification in external control samples, the iECAT-Score tests controlled type I error rates, although the power improvement was slightly reduced compared to when no misclassification existed. Nonetheless, we recommend attentive examination of the study samples and the prevalence of phenotype of interest before making the decision to integrate subjects as external controls.

The current version of iECAT-Score methods is constructed to analyze one set of external controls. However, there is no technical difficulty to apply our method to more than one set of controls. Multiple sources of controls could be treated as a whole and our method could be applied. When there exists drastic differentiation among different sources of external controls, it is possible that combining them could increase the batch effect between the combined set and internal samples. Our method is applicable and performs well with the existence of batch effect. However, by introducing additional controls, the added noise could decrease the power of association testing, offsetting some improved power from an increased sample size. Hence, although there is no technical problem to apply our method to more than one set of external controls, we suggest that such a decision can only be made after careful consideration of the possible additional batch effect and noise introduced.

In this article, our iECAT-Score method tests association at a single variant in its currently presented format. We will extend the approach to region-based rare variant association tests, such as burden (Li & Leal, 2008), SKAT (Wu et al., 2011), and SKAT-O (Lee, Wu, & Lin, 2012) type tests. The proposed method uses genotype data to assess the existence of batch effect between internal and external studies. As pointed out by Derkach (Derkach et al., 2014), differential misclassification rates during the genotyping procedure resulted from different read depths could lead to batch effect. We will extend our method to utilizing genotype probabilities given sequencing reads to further adjust for systematic batch effect when such information is available.

2.5 Supplementary Materials

2.5.1 Validation of Theoretical Results

2.5.1.1 Covariance of $(\boldsymbol{S_{int}, S_{all}, S_{IvE}})$

We assume that $(S_{int}, S_{all}, S_{IvE})^T \sim N_3(\mathbf{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$, under the null hypothesis of no genetic effect on the phenotype of case or control and no batch effect between internal and external controls. We omit the subscript to show the three statistics and their variances. Marginally, each statistic takes the form $S = \boldsymbol{G^T(Y - \hat{\mu})}$, where $\boldsymbol{\mu} = \{\mu_i\} = \{\Pr(Y_i = 1 \mid X_i)\}$ under H_0, and $\hat{\mu}_i$ is the maximum likelihood estimate of μ_i. Under the null hypothesis, $E(S) = 0$

and $Var(S) = \sum_{i=1}^{n} \tilde{G}_i^2 \hat{\mu}_i(1-\hat{\mu}_i))$, where $\widetilde{\boldsymbol{G}} = \{\tilde{G}_i\} = \boldsymbol{G} - \boldsymbol{X}(\boldsymbol{X^T V X})^{-1}\boldsymbol{X^T V G}$ is the covariate-adjusted genotype vector, and $\boldsymbol{V} = diag\{\hat{\mu}_i(1-\hat{\mu}_i)\}$.

We show the covariance terms of $\boldsymbol{\Sigma}$ by using general subscripts. Let S_a and S_b be two score statistics taking forms $S_a = \Sigma_{i\in A} G_i(Y_i - \hat{\mu}_i) = \Sigma_{i\in A}\tilde{G}_i(Y_i - \hat{\mu}_i)$ and $S_b = \Sigma_{j\in B} G_j\big(Y_j - \hat{\mu}_j\big) = \Sigma_{i\in A}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)$, where A and B are sets containing samples i and j, respectively, and $\widetilde{\boldsymbol{G}}$ is the covariate-adjusted genotype vector. Then $cov(S_a, S_b) = cov\left(\Sigma_{i\in A}\tilde{G}_i(Y_i - \hat{\mu}_i), \Sigma_{j\in B}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right) = E\left[\left(\Sigma_{i\in A}\tilde{G}_i(Y_i - \hat{\mu}_i)\right)\left(\Sigma_{j\in B}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right)\right] - E\left(\Sigma_{i\in A}\tilde{G}_i(Y_i - \hat{\mu}_i)\right) \times E\left(\Sigma_{j\in B}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right) = E\left[\left(\Sigma_{i\in A}\tilde{G}_i(Y_i - \hat{\mu}_i)\right)\left(\Sigma_{j\in B}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right)\right] = E\left\{\left[\left(\Sigma_{i\in(A\cap B)}\tilde{G}_i(Y_i - \hat{\mu}_i)\right)\left(\Sigma_{i\in(A\setminus B)}\tilde{G}_i(Y_i - \hat{\mu}_i)\right)\right] \times \left[\left(\Sigma_{j\in(A\cap B)}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right)\left(\Sigma_{j\in(B\setminus A)}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right)\right]\right\} = E\left[\left(\Sigma_{i\in(A\cap B)}\tilde{G}_i(Y_i - \hat{\mu}_i)\right)\left(\Sigma_{j\in(A\cap B)}\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right)\right] = E\left[\left(\Sigma_{i=j}\tilde{G}_i(Y_i - \hat{\mu}_i)\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right) + \left(\Sigma_{i\neq j}\tilde{G}_i(Y_i - \hat{\mu}_i)\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right)\right] = E\left(\Sigma_{i=j}\tilde{G}_i(Y_i - \hat{\mu}_i)\tilde{G}_j\big(Y_j - \hat{\mu}_j\big)\right)$. Hence, $\widehat{cov(S_a, S_b)} = \Sigma_{i=j}\tilde{G}_i(Y_i - \hat{\mu}_i)\tilde{G}_j\big(Y_j - \hat{\mu}_j\big), i \in B, j \in B$.

2.5.1.2 Coefficient $\boldsymbol{a}$ in iECAT-Score statistic $\boldsymbol{S_w}$

Without loss of generality, we derive the coefficient a in iECAT-Score statistic without covariates X. Let n_1^I, n_0^I, n_0^E denote the sample sizes of internal cases, internal controls, and external controls, respectively. Let p_1^I, p_0^I, p_0^E denote the minor allele frequencies (MAFs) of internal cases, internal controls, and external controls. The score statistic using internal samples only is $S_{int} = \boldsymbol{G_{int}^T}(\boldsymbol{Y}_{int} - \widehat{\boldsymbol{\mu}}_{int}) = \boldsymbol{G_{int}^T Y}_{int} - \boldsymbol{G_{int}^T}\widehat{\boldsymbol{\mu}}_{int} = 2n_1^I p_1^I - \left(2n_1^I p_1^I \frac{n_1^I}{n_1^I + n_0^I} + 2\, n_0^I p_0^I \frac{n_1^I}{n_1^I + n_0^I}\right) = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I}(p_1^I - p_0^I)$. Similarly, S_{all} can be represented by $S_{all} = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I + n_0^E}(p_1^I - p_0^I) + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_1^I - p_0^E)$. Hence, we can rewrite S_{all} as a function of S_{int} as $S_{all} = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I + n_0^E}(p_1^I - p_0^I) + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_1^I - p_0^I) + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_0^I - p_0^E) = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I}(p_1^I - p_0^I) \times \frac{n_1^I n_0^I}{n_1^I + n_0^I + n_0^E} + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_1^I - p_0^I) + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_0^I - p_0^E) = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I}(p_1^I - p_0^I) \times \frac{(n_1^I + n_0^I)(n_1^I n_0^I + n_1^I n_0^E)}{n_1^I n_0^I (n_1^I + n_0^I + n_0^E)} + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_0^I - p_0^E) = a \times S_{int} + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_0^I - p_0^E)$, where $a = \frac{(n_1^I + n_0^I)(n_1^I n_0^I + n_1^I n_0^E)}{n_1^I n_0^I (n_1^I + n_0^I + n_0^E)}$. Under the null hypothesis of no batch

effect between internal and external control samples, $p_0^I = p_0^E$, i.e., $S_{all} = a \times S_{int}$. Thus, a compound statistic as a weighted sum of S_{int} and S_{all} with weight τ requires the formation $S_w = a\tau S_{int} + (1-\tau)S_{all}$.

2.5.1.3 Calculation of iECAT-Score minimum p value

Under the null hypothesis, (S_w, S_{int}) jointly follow a bivariate normal distribution. The covariance is given by $cov(S_w, S_{int}) = cov(a\tau S_{int} + (1-\tau)S_{all}, S_{int}) = a\tau Var(S_{int}) + (1-\tau)cov(S_{int}, S_{all}) = (a\tau + 1 - \tau)Var(S_{int})$, where $\tau = \frac{\tau_1}{1+\tau_1}$ with $\tau_1 = \frac{S_{IvE}^2}{Var(S_{IvE})}$. We estimate the covariance here by treating τ as a constant instead of a random variable to simplify the calculation.

The minimum *p* value (iECAT-Score minP) is calculated as the probability of observing one or both the *p* values as small as the smaller one of the two under the null hypothesis of no association. Specifically, let $p(S_{int})$ and $p(S_w)$ denote the *p* values calculated from using internal samples only and using the iECAT-Score method, respectively. Let Z_{int} and Z_w denote the standardized z-scores: $Z_{int} = \Phi^{-1}(1 - \frac{p(S_{int})}{2})$ and $Z_w = \Phi^{-1}(1 - \frac{p(S_w)}{2})$. The minimum *p* value is given by $1 - Prob\{\max(|Z_{int}|, |Z_w|) < \Phi^{-1}(1 - \frac{\min(p(S_{int}), p(S_w))}{2}\}$, which can be calculated using numerical integration in R.

2.5.1.4 Power of tests using internal controls only ($\boldsymbol{S_{int}}$) and naively combining controls ($\boldsymbol{S_{all}}$)

We provide below some numerical guidelines for the test of internal only (S_{int}) and the naïve strategy of combined control samples (S_{all}) to show the conditions where having the external controls would not provide an increase in power, without considering covariates.

Let n_1^I, n_0^I, n_0^E denote the sample sizes of internal cases, internal controls, and external controls, respectively. Let p_1^I, p_0^I, p_0^E denote the minor allele frequencies (MAFs) of internal cases, internal controls, and external controls. Without considering covariates, we showed in Supplementary materials Appendix B that score statistics S_{int} and S_{all} can be represented as functions of minor allele frequencies of internal cases, internal controls, and external controls: $S_{int} = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I}(p_1^I - p_0^I)$, and $S_{all} = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I + n_0^E}(p_1^I - p_0^I) + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_1^I - p_0^I) + 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}(p_0^I - p_0^E) = 2\frac{n_1^I n_0^I + n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}p_1^I - 2\frac{n_1^I n_0^I}{n_1^I + n_0^I + n_0^E}p_0^I - 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}p_0^E$.

Hence, $S_{int} \sim N(\hat{\mu}_{int}, \hat{\sigma}^2_{int})$, where $\hat{\mu}_{int} = 2\frac{n_1^I n_0^I}{n_1^I + n_0^I}(\hat{p}_1^I - \hat{p}_0^I)$ and $\hat{\sigma}^2_{int} = \left(2\frac{n_1^I n_0^I}{n_1^I + n_0^I}\right)^2 \left\{\frac{\hat{p}_1^I(1-\hat{p}_1^I)}{2n_1^I} + \frac{\hat{p}_0^I(1-\hat{p}_0^I)}{2n_0^I}\right\}$. Then $S^2_{int}/\hat{\sigma}^2_{int}$ approximately follows a non-central chi-squared distribution with non-centrality parameter $\frac{\hat{\mu}^2_{int}}{\hat{\sigma}^2_{int}}$. Similarly, $S_{all} \sim N(\hat{\mu}_{all}, \hat{\sigma}^2_{all})$, where $\hat{\mu}_{all} = 2\frac{n_1^I n_0^I + n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}\hat{p}_1^I - 2\frac{n_1^I n_0^I}{n_1^I + n_0^I + n_0^E}\hat{p}_0^I - 2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}\hat{p}_0^E$ and $\hat{\sigma}^2_{all} = \left(2\frac{n_1^I n_0^I + n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}\right)^2 \times \frac{\hat{p}_1^I(1-\hat{p}_1^I)}{2n_1^I} + \left(2\frac{n_1^I n_0^I}{n_1^I + n_0^I + n_0^E}\right)^2 \times \frac{\hat{p}_0^I(1-\hat{p}_0^I)}{2n_0^I} + \left(2\frac{n_1^I n_0^E}{n_1^I + n_0^I + n_0^E}\right)^2 \times \frac{\hat{p}_0^E(1-\hat{p}_0^E)}{2n_0^E}$. $S^2_{all}/\hat{\sigma}^2_{all}$ approximately follows a non-central chi-squared distribution with non-centrality parameter $\frac{\hat{\mu}^2_{all}}{\hat{\sigma}^2_{all}}$.

Under the null hypothesis of no associations, both standardized squared score statistics, $S^2_{int}/\hat{\sigma}^2_{int}$ and $S^2_{all}/\hat{\sigma}^2_{all}$, follow the (central) chi-squared distribution with one degree of freedom, and hence the critical value at $\alpha = 5 \times 10^8$ is 29.7168. Using a built-in non-central chi-squared CDF function in R, we can calculate power as the probability of observing a standardized squared score statistic larger than 29.7168. Since the power monotonically increases as the non-centrality parameter increases, the power of test that naively combines controls is *greater* than that of the test using exclusively internal samples when $\frac{\hat{\mu}^2_{all}}{\hat{\sigma}^2_{all}} > \frac{\hat{\mu}^2_{int}}{\hat{\sigma}^2_{int}}$.

2.5.2 Supplementary Tables and Figures

Table S2.1: Empirical type I error rates under alternative batch effect mechanism.
We assumed that 3% of the variants were subject to batch effect in internal and external control samples. For such variants, each allele had a 2% chance of being mis-genotyped as the alternative allele. Shown in the table are empirical type I error rates of score tests using internal samples exclusively, using combined samples without adjusting for batch effect, and various versions of the iECAT-Score method, under different scenarios. Type I error rates were estimated based on 10^9 simulations.

Sample size (Internal: external)	Internal case:control	Level alpha	Internal only	iECAT-Score	iECAT-Score minP	Combined internal and external samples
10000:10000	1:1	$5x10^{-5}$	5.06e-05	3.58e-05	4.06e-05	2.81e-02
		$1x10^{-6}$	1.10e-06	7.74e-07	8.74e-07	1.52e-03
		$5x10^{-8}$	7.20e-08	4.40e-08	5.60e-08	9.41e-04

Figure S2.1: Empirical power comparisons under alternative batch effect mechanism.
We assumed that 3% of the variants were subject to different MAFs in internal and external control samples. For such variants, each allele had a 2% chance of being mis-genotyped as the alternative allele. Shown are empirical powers at $\alpha = 5 \times 10^{-8}$ from simulations using 10,000 internal samples and 10,000 external control samples, with internal case and control ratio 1:1. The effect size of the causal variant was $\beta = \log(Odds\ Ratio)$.

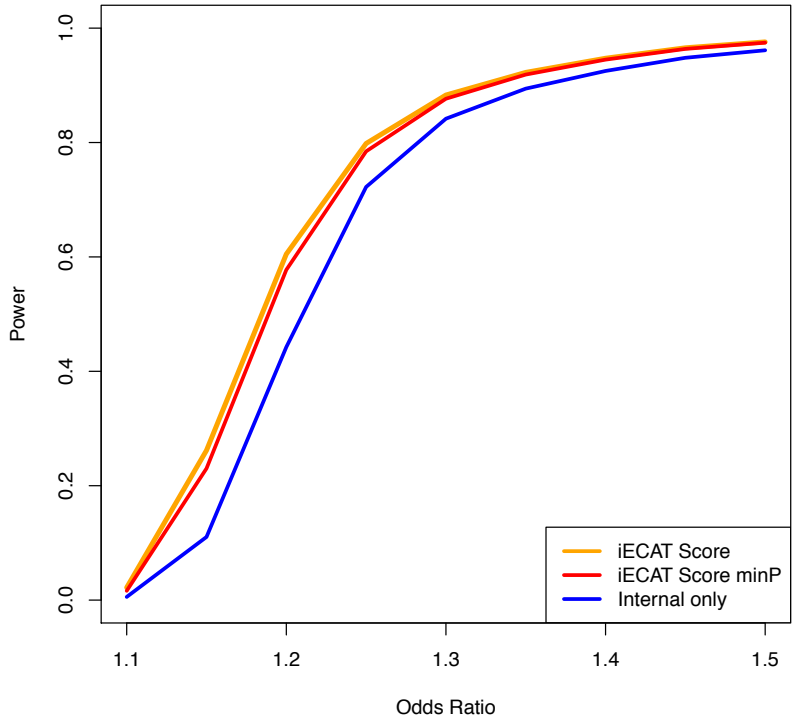


Table S2.2: Empirical type I error rates under increased batch effect between internal and external samples.

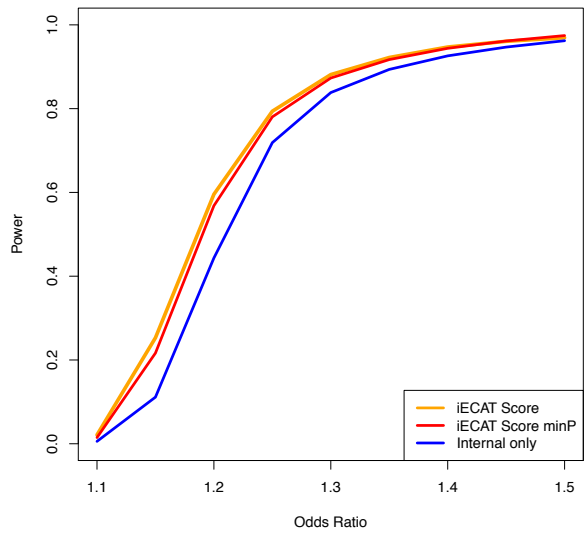
We assumed that 12% of the variants were subject to different MAFs in internal and external control samples. Shown in the table are empirical type I error rates of score tests using internal samples exclusively, using combined samples without adjusting for batch effect, and various versions of the iECAT-Score method, under different scenarios. Type I error rates were estimated based on 5×10^8 simulations.

Sample size (Internal: external)	Internal case:control	Level alpha	Internal only	iECAT-Score	iECAT-Score minP	Combined internal and external samples
10000:10000	1:1	$5x10^{-5}$	4.96e-05	3.59e-05	3.57e-05	1.04e-01
		$1x10^{-6}$	8.58e-07	7.44e-07	6.33e-07	1.00e-01
		$5x10^{-8}$	3.44e-08	3.70e-08	1.85e-08	9.80e-02
10000:10000	2:1	$5x10^{-5}$	5.04e-05	4.07e-05	3.67e-05	1.07e-01
		$1x10^{-6}$	9.82e-07	1.45e-06	1.12e-06	1.04e-01
		$5x10^{-8}$	5.33e-08	1.60e-07	1.21e-07	1.02e-01

Figure S2.2: Empirical power comparisons under increased batch effect between internal and external samples.
We assumed that 12% of the variants were subject to different MAFs in internal and external control samples. Shown are empirical powers at $\alpha = 5 \times 10^{-8}$ from simulations using 10,000 internal samples and 10,000 external control samples, with internal case and control ratio 1:1 and 2:1. The effect size of the causal variant was $\beta = \log(Odds\ Ratio)$.

Case: Control

1:1



2:1

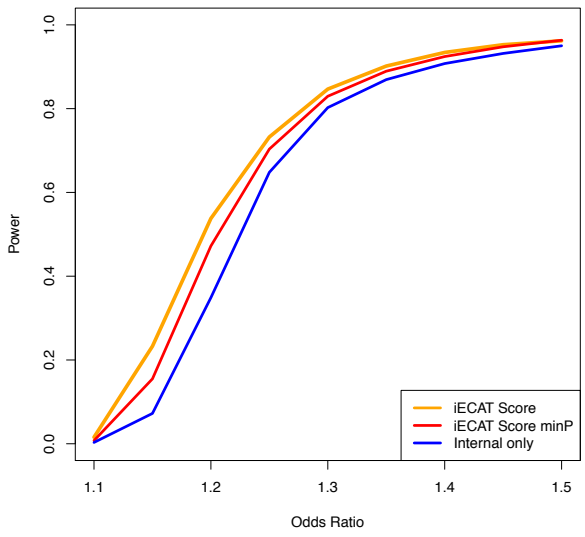


Table S2.3: Genomic inflation factors (GIF) from the comparisons between internal controls and external controls.

The GIF shows the genomic inflation factor calculated at the median of test statistics, in real data, and simulated data when batch effect existed in three and twelve percent of variants.

	Real Data	Simulation: 3% batch effect	Simulation: 12% batch effect
GIF	1.19	1.12	1.31

Figure S2.3: Power comparisons with varying MAFs.
Powers of the iECAT-Score method, iECAT-Score minP method, the method using exclusively internal samples, and method that naively combines internal and external control samples, when MAF of external controls varied relative to MAF of internal cases and controls. MAF of internal cases was 0.2; the odds ratio of causal allele was 1.2. The sample sizes of internal cases, internal controls and external controls were 5000, 5000, 10000, respectively. Nominal level $\alpha = 5 \times 10^{-8}$. Powers of the internal only and naïve approaches were based on analytical calculations assuming non-central chi-squared distributions. Powers of the iECAT-Score methods were based on 5×10^5 simulations at each value of MAF for external controls.

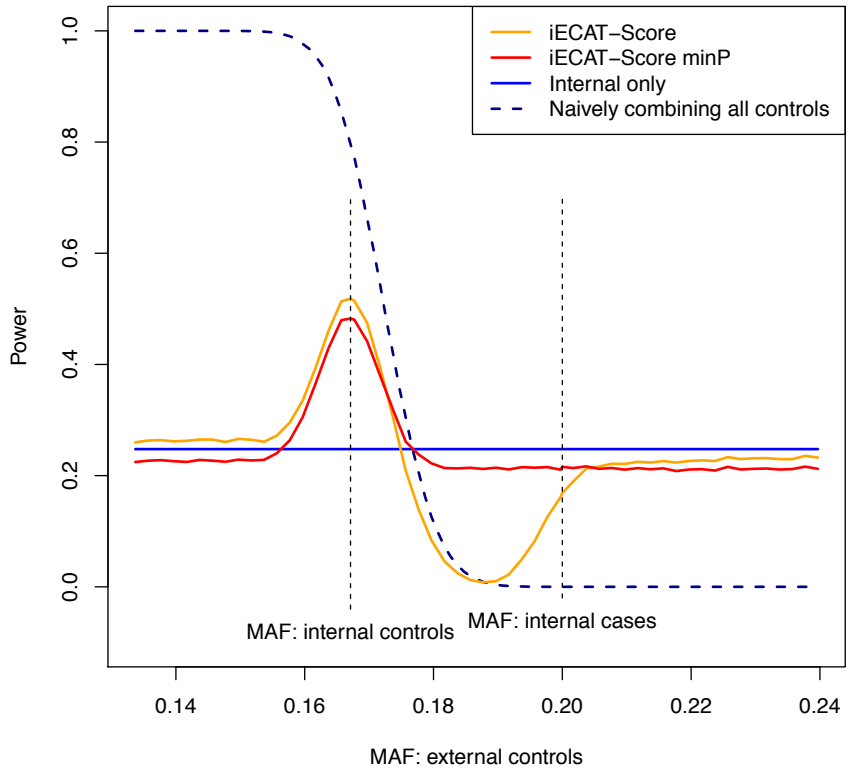


Figure S2.4: Quantile-quantile plots of association *p* values for 500,000 variants from simulation studies with population stratification.
Sample sizes for internal cases, internal controls, and external controls were 5000, 5000, 10000, respectively. Samples were assumed to come from Italian and Finnish populations. MAFs for 500,000 variants were generated following the MAF spectrum of those of Toscani in Italia (TSI) and Finnish in Finland (FIN) from The 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). Genetic principal components were calculated using 100,000 simulated genotypes, using the Fruposa software (Zhang, Dey, & Lee, 2020). Disease prevalence in TSI and FIN populations was assumed to be 0.05 and 0.07, respectively. Internal samples were assumed to consist of 50/50 subjects from TSI and FIN populations; external control samples consisted of (a) 60/40 and (b) 90/10 TSI and FIN samples, respectively.

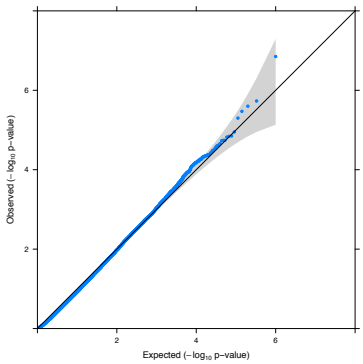
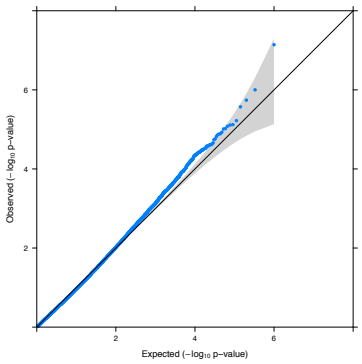
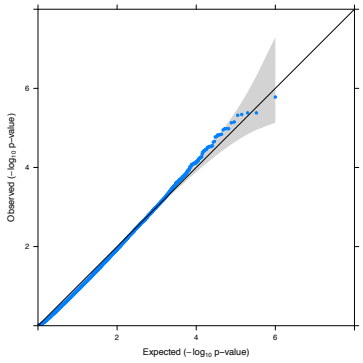
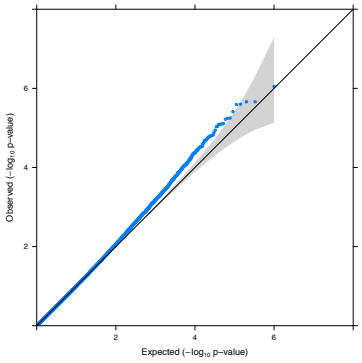
External samples constitution	iECAT-Score minP	iECAT-Score
(a) 60% TSI, 40% FIN		
(b) 90% TSI, 40% FIN		

Table S2.4: Empirical type I error rates when internal and external samples have different distributions of covariates that contribute to disease risk.

We generated binary phenotypes of case/control from the logistic regression model: $logit[\Pr(Y = 1 \mid X, G)] = \alpha_0 + 0.5X_1 + 0.5X_2$. For internal samples, X_1 was a continuous covariate generated from normal distribution $N(0, 1)$ and X_2 was a dichotomous covariate with a probability of 0.5 being 1; for external samples, X_1 was a continuous covariate generated from normal distribution $N(2, 1)$ and X_2 was a dichotomous covariate with a probability of 0.8 being 1. The intercept α_0 was chosen such that the disease prevalence in internal and external samples was 0.05 and 0.1, respectively. Shown in the table are empirical type I error rates of score tests using internal samples exclusively, using combined samples without adjusting for batch effect, and various versions of the iECAT-Score method, under different scenarios. Type I error rates were estimated based on 5×10^8 simulations.

Sample size (Internal: external)	Internal case:control	Level alpha	Internal only	iECAT-Score	iECAT-Score minP	Combined internal and external samples
10000:10000	1:1	$5x10^{-5}$	4.99e-05	4.88e-05	3.79e-05	2.44e-02
		$1x10^{-6}$	9.73e-07	1.03e-06	7.76e-07	2.32e-02
		$5x10^{-8}$	3.60e-08	6.80e-08	2.28e-08	2.25e-02

Figure S2.5: Empirical power comparisons when internal and external samples have different distributions of covariates.
We generated binary phenotypes of case/control from the logistic regression model: $logit[\Pr(Y = 1 \mid X, G)] = \alpha_0 + 0.5X_1 + 0.5X_2 + \beta G$. For internal samples, X_1 was a continuous covariate generated from normal distribution $N(0, 1)$ and X_2 was a dichotomous covariate with a probability of 0.5 being 1; for external samples, X_1 was a continuous covariate generated from normal distribution $N(2, 1)$ and X_2 was a dichotomous covariate with a probability of 0.8 being 1. G is the genotype at the variant of interest generated from a binomial (2, MAF) distribution, and β is the effect size of the variant. The intercept α_0 was chosen such that the disease prevalence in internal and external samples was 0.05 and 0.1, respectively. We assumed that 3% of the variants were subject to different MAFs in internal and external control samples. Shown are empirical powers at $\alpha = 5 \times 10^{-8}$ from simulations using 10,000 internal samples and 10,000 external control samples, with internal case and control ratio 1:1. The effect size of the causal variant was $\beta = \log(Odds\ Ratio)$.

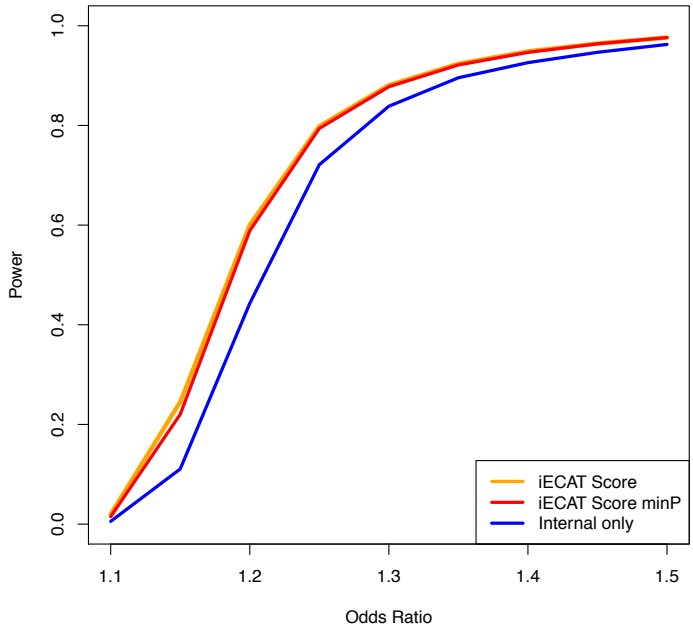


Table S2.5: Empirical type I error rates with small internal case sample size compared to internal controls.

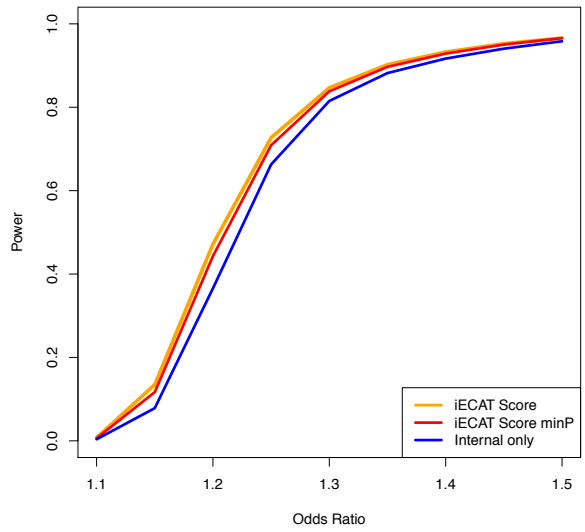
Shown in the table are empirical type I error rates of score tests using internal samples exclusively, using combined control samples without adjusting for batch effect, and various versions of the iECAT-Score method, under different scenarios. Type I error rates were estimated based on 10^9 simulations.

Sample size (Internal: external)	Internal case:control	Level alpha	Internal only	iECAT-Score	iECAT-Score minP	Combined internal and external samples
10000:10000	1:2	4.99e-05	3.42e-05	3.87e-05	2.50e-02	4.99e-05
		1.01e-06	6.41e-07	7.38e-07	2.39e-02	1.01e-06
		4.51e-08	4.67e-08	4.34e-08	2.32e-02	4.51e-08
10000:10000	1:5	4.97e-05	3.67e-05	4.08e-05	2.26e-02	4.97e-05
		9.87e-07	6.70e-07	7.47e-07	2.10e-02	9.87e-07
		4.17e-08	3.50e-08	3.34e-08	2.00e-02	4.17e-08

Figure S2.6: Empirical power comparisons with small internal case sample size compared to internal controls.
We assumed that 3% of the variants were subject to different MAFs in internal and external control samples. Shown are empirical powers at $\alpha = 5 \times 10^{-8}$ from simulations using 10,000 internal samples and 10,000 external control samples, with internal case and control ratio 1:2 and 1:5. The effect size of the causal variant was $\beta = \log(Odds\ Ratio)$.

Case: Control

1:2



1:5

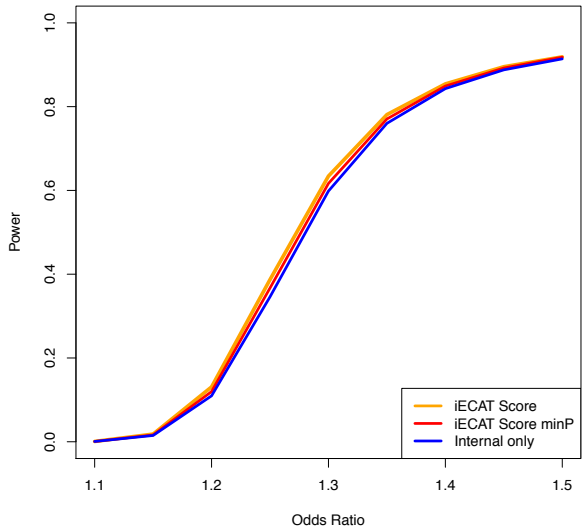


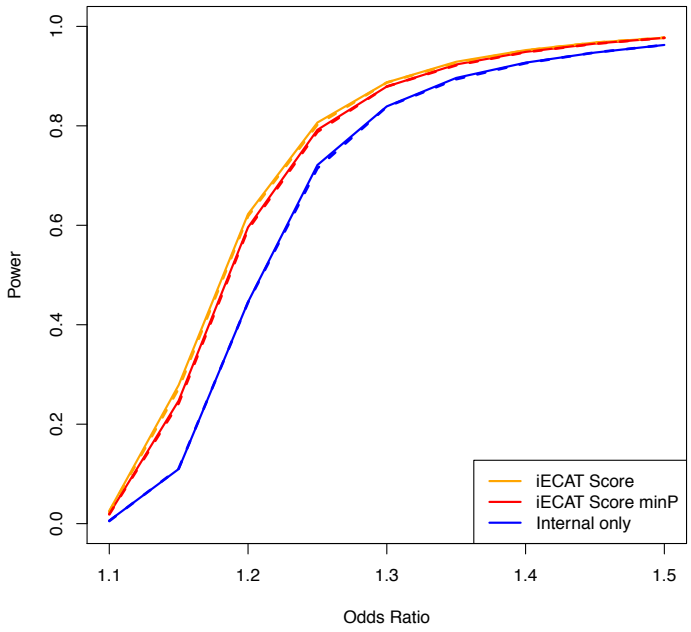
Table S2.6: Empirical type I error rates with small internal control sample size.
Shown in the table are empirical type I error rates of score tests using internal samples exclusively, using combined control samples without adjusting for batch effect, and various versions of the iECAT-Score method, under different scenarios. Type I error rates were estimated based on 10^9 simulations.

Sample size (Internal: external)	Internal case:control	Level alpha	Internal only	iECAT-Score	iECAT-Score minP	Combined internal and external samples
1500:10000	2:1	$5x10^{-5}$	5.03e-05	4.29e-05	4.88e-05	2.46e-02
		$1x10^{-6}$	1.04e-06	3.28e-06	2.83e-06	2.34e-02
		$5x10^{-8}$	6.00e-08	8.24e-07	6.40e-07	2.27e-02

Table S2.7: Empirical type I error rates with misclassification in external control samples.
Of the external samples, there were 1% of samples that were misclassified as controls. Shown in the table are empirical type I error rates of score tests using internal samples exclusively, using combined control samples without adjusting for batch effect, and various versions of the iECAT-Score method, under different scenarios. Type I error rates were estimated based on 10^9 simulations.

Sample size (Internal: external)	Internal case:control	Level alpha	Internal only	iECAT-Score	iECAT-Score minP	Combined internal and external samples
10000:10000	1:1	$5x10^{-5}$	5.01e-05	3.34e-05	4.04e-05	2.58e-02
		$1x10^{-6}$	9.96e-07	6.80e-07	8.12e-07	2.50e-02
		$5x10^{-8}$	4.20e-08	3.60e-08	4.20e-08	2.44e-02

Figure S2.7: Empirical power comparisons when 1% of external control samples were case samples.
The power was estimated with 10,000 internal samples and 10,000 external control samples, with internal case and control ratio 1:1. Lines represent empirical powers at $\alpha = 5 \times 10^{-8}$, where dashed lines show powers when misclassification of disease status exists in external control samples. The effect size of the causal variant was $\beta = \log(Odds\ Ratio)$.



Chapter 3 Integrating External Controls in Case-Control Studies Improves Power for Rare-Variant Tests

3.1 Introduction

Recent advances in genotyping and sequencing technologies have enabled progressively larger scale sequencing and genotyping projects to identify disease-associated rare (Cruchaga et al., 2014). For instance, the Michigan Genomics Initiative has collected genotype data from over 66,000 unrelated individuals within Michigan Medicine; the UK Biobank has produced genome-wide genotype data on approximately 500,000 individuals from the United Kingdom and exome sequence data on 200,200. This rapid increase in the number of genotyped and sequenced individuals provides a unique opportunity to develop methods that can leverage the large-scale sequencing and genotyping projects, whose data are publicly available, as additional control samples to increase the power of rare-variant association testing in case-control studies.

When combining controls from external studies, systematic batch effect between genotyped data from different studies are likely to exist due to differences in sequencing platforms, genotype calling procedures and population stratification. Undesired type I error inflation could result from the systematic batch effect if they are left unaddressed. Several recent methodologic developments have attempted to address the systematic differences between genotyped data of internal and external sources, most of which directly or indirectly use sequence read data. Derkach et al. (Derkach et al., 2014) developed a score test that replaces the called genotype with an expected genotype. The calculation of expected genotype requires known read depths, base-calling error rates of the sequencing platform and prior knowledge on allele frequencies. Using expected genotype accounts for several factors that contribute to the systematic batch effect between genotyped data from different studies and thus reduces inflation in type I error rates, but the calculation of which could be challenging if the posterior genotype likelihoods are not provided in the genotype vcf files. In addition, by considering the

retrospective setting, this method does not allow for covariate adjustment. Extending Derkach's method, Chen and Lin (Chen & Lin, 2018) proposed regression calibration (RC)-based and maximum-likelihood (ML)-based methods to account for differential sequencing errors between cases and controls; these methods allow for parameter's effect size estimation, with the assumption that weak confounding from population stratification are the only potential confounders. Hu et al. (Hu, Liao, Johnston, Allen, & Satten, 2016) proposed a likelihood-based method that directly models sequencing reads using sequence data without calling the genotypes. This method first estimates the single nucleotide variant (SNV) locations and then applies a burden-type test to assess the significance of the association between an SNV and a trait. Hendricks et al. (Hendricks et al., 2018) proposed ProxECAT which uses allele counts from genotyped data to estimate enrichment of rare variants in external controls. Although this method does not require genotype probabilities or sequence read data to be available, it does not include internal control samples in the analyses as a baseline reference and results in consistent inflation in Type I error rates. Thus, the author suggests using more conservative significance level, which potentially limits the power of the association test.

To address the shortcomings of the above methods, we recently proposed a novel score-based test, the iECAT-Score test, that uses genotype data to integrate external control samples into association test. We built upon the original iECAT test developed by Lee et al (Lee et al., 2017), which assesses the batch effect and includes external control samples using allele counts from genotype data, and developed a score test that further allows for covariate adjustment. Compared to the iECAT test in its originally presented format, the score tests are more stable, computationally efficient, and allow to adjust for covariates and population stratification. Through applying recent improvements of score tests including the Saddlepoint approximation (SPA) (Dey et al., 2017) and efficient resampling (ER) (Lee et al., 2015) methods, the iECAT-Score test protects the type I error in the scenario of case-control imbalance and low minor allele count (MAC).

The iECAT-Score test we previously proposed tests association between a single variant and the disease status. It controls type I error rates while increasing samples from external studies to improve power for association tests. However, in the case of rare-variant association test, single-variant tests are often underpowered to identify causal rare variants. Hence, in this work, we extend the single-variant iECAT-Score test to burden (B. Li & Leal, 2008), SKAT (Wu

et al., 2011) and SKAT-O (Lee, Wu, & Lin, 2012) type tests to test for the combined genetic effects in a gene or region. Similar to the burden, SKAT and SKAT-O tests that are used in the common case-control studies setting, the iECAT-Score region test aggregates the single-variant test statistics using a weighted linear (burden) or quadratic (SKAT) sum, or a linear combination of both (SKAT-O). Association between the rare variants in the region and the phenotype is then assessed by comparing the compound statistic to a specified distribution under the null of no genetic effect.

We organize this article as follows. We first introduce the model for the rare variant association test in case-control studies using burden, SKAT, and SKATO tests, and propose the iECAT-Score region tests that allow for integration of external control samples in case-control association tests. We then describe the simulation studies to assess the type I error rates and power of our proposed methods, as well as their applications to the association studies of age-related macular degeneration (AMD) combining data from the International AMD Genomics Consortium (IAMDGC) (Fritsche et al., 2015) and the UK Biobank (Bycroft et al., 2018). Finally, we present the results from simulation studies and data analyses of the proposed methods, discuss our findings, and provide guidelines for integrating external control samples in case-control studies.

3.2. Materials and Methods

3.2.1 iECAT-Score Region-based association test

The iECAT-Score region-based test is a shrinkage-estimation-based test for aggregated genetic effects within a genomic region. At each variant within the region, the iECAT-Score test assesses the batch effect between internal and external control samples and constructs a shrinkage estimator to access the single-variant genetic effect. The iECAT-Score region-based test then groups the single variant test statistics to test for association between the joint effect of variants in a region and outcome of disease status, using burden, SKAT, and SKAT-O type methods.

Region-Based Test for Genetic Effect

To present the model for a region-based test that tests for the aggregated genetic effect within a gene or a region, we first consider a scenario of no external controls. Consider the internal study

of sample size *n.* For subject *i*, let $y_i = 0/1$ be the dichotomous phenotype for control/case; $\boldsymbol{X}_i = \left(X_{i1}, X_{i2}, .., X_{ip}\right)$ is the covariate vector of length *p*; $\boldsymbol{G}_i = (G_{i1}, G_{i2}, .., G_{im})$ is the vector of genotypes consisting genotypes at *m* variants within a region. Hence, $\boldsymbol{G} = (\boldsymbol{G}_1, \boldsymbol{G}_2, .., \boldsymbol{G}_n)^T$ is the $n \times m$ genotype matrix for the *n* subjects at *m* variants. To relate the phenotype Y, the covariates $\boldsymbol{X}$, and the genotype $\boldsymbol{G}$, we consider the following logistic regression model

$$logit[\Pr(Y_i = 1 \mid X_i, G_i)] = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{\beta} \qquad (3.1)$$

where the phenotype Y_i follows a Bernoulli distribution, $\boldsymbol{\alpha}$ is an $p \times 1$ vector of coefficients for the covariates, and $\boldsymbol{\beta} = (\beta_1, \beta_2, .., \beta_m)^T$ are the regression coefficients for the *m* variants in a region. We assume that $\boldsymbol{\beta}$ is a random vector with $E\left(\beta_j\right) = 0$, $Var\left(\beta_j\right) = w_j^2\tau$ where w_j is the weight assigned to variant $j11$, and $corr\left(\beta_j, \beta_k\right) = \rho, j, k \in \{1,2, .., m\}$. To test for the association between the phenotypes Y_i and the genotypes within the region, we want to test for the null hypothesis H_0: $\boldsymbol{\beta} = \boldsymbol{0}$ in Equation (1).

Within the region of *m* variants, the score test statistic at variant j is given by $S_j = \sum_{i=1}^{n} G_{ij}(Y_i - \hat{\mu}_i)$ where $\hat{\mu}_i$ is the maximum likelihood estimate of μ_i with $\boldsymbol{\mu} = \{\mu_i\} = \{\Pr(Y_i = 1 \mid X_i)\}$ under H_0. To test the null hypothesis under the assumption that $\tau = 0$, the burden- and SKAT-type score test statistics can be constructed as

$$Q_B = \left(\sum_{j=1}^{m} w_j S_j\right)^2, \text{ and } Q_S = \sum_{j=1}^{m}\left(w_j S_j\right)^2,$$

where w_j is the weight assigned to variant j (Wu et al., 2011). The omnibus SKAT-O type test takes the form of weighted sum of the burden and SKAT test statistics and can be constructed

$$Q_\rho = (1 - \rho)Q_B + \rho Q_S,$$

where ρ is a parameter between 0 and 1 (Lee et al., 2012).

Under the null hypothesis of no genetic effect, $\boldsymbol{S}$ approximately follows a multivariate Gaussian distribution with mean zero and variance $\boldsymbol{\Phi} = \boldsymbol{W\Sigma W}$, where $\boldsymbol{W} = \{w_j\}$ is the diagonal weight matrix and $\boldsymbol{\Sigma}$ is the covariance matrix of $(S_1, S_2, \ldots, S_m)'$. The covariance matrix of $(S_1, S_2, \ldots, S_m)'$ is given by $\boldsymbol{\Sigma} = \boldsymbol{A}^{1/2}\boldsymbol{C}\boldsymbol{A}^{1/2}$, with $\boldsymbol{A} = diag\{Var(S_1), Var(S_2), \ldots, Var(S_m)\}$ and $\boldsymbol{C}$ a $m \times m$ correlation matrix of m variants. The statistic Q_ρ approximately follows mixture of chi-square distributions under the null hypothesis, and Davies method (Davies, 1980) can be applied to obtain the *p* value. As the parameter ρ is unknown, the SKAT-O test adaptively applies the minimum *p* values over a grid of ρ to search for a ρ that maximizes power.

iECAT-Score Region-Based Test

We introduce external control samples to the internal study samples that consist of cases and controls. We apply the single-variant iECAT-Score method (Y. Li & Lee, 2021) to each variant that integrates external controls to improve the power. Let n_1^I, n_0^I, n_0^E denote the sample sizes of internal cases, internal controls, and external controls, respectively. Similar to the notations in model (1), we let $Y_i = 0/1(i = 1,2, \dots, n_1^I + n_0^I + n_0^E)$ be the dichotomous phenotype for control/case.

At variant j within a region, the score statistic that tests for the association between the variant j and the phenotype using exclusively internal samples is given by $S_{int,j} = \boldsymbol{G}_{int,j}^T(\boldsymbol{Y}_{int} - \widehat{\boldsymbol{\mu}}_{int})$. In this equation, $\boldsymbol{G}_{int,j} = \left(G_{int,1,j}, G_{int,2,j}, .., G_{int,(n_1^I + n_0^I),j}\right)^T$ is the vector of genotypes at variant j of internal samples; similarly, $\boldsymbol{Y}_{int} = \left(Y_{int,1}, Y_{int,2}, .., Y_{int,(n_1^I + n_0^I)}\right)^T$ is the vector of phenotypes of internal samples, and $\widehat{\boldsymbol{\mu}}_{int} = \left(\hat{\mu}_{int,1}, \hat{\mu}_{int,2}, .., \hat{\mu}_{int,(n_1^I + n_0^I)}\right)^T$ is the vector of maximum likelihood estimate of $\boldsymbol{\mu}_{int}$ under the null logistic regression model of no genetic effect built using internal samples only as in model (1). When external control samples are included assuming no systematic differences between internal and external studies, we construct a score statistic at variant j as $S_{all,j} = \boldsymbol{G}_{all,j}^T(\boldsymbol{Y}_{all} - \widehat{\boldsymbol{\mu}}_{all})$. In this equation, $\boldsymbol{G}_{all,j}, \boldsymbol{Y}_{all}$ and $\widehat{\boldsymbol{\mu}}_{all}$ are vectors of length $n_1^I + n_0^I + n_0^E$, denoting genotypes at variant j, phenotypes, and expected mean outcome under a null model built of combined internal and external samples.

Using a similar approach to the single variant iECAT-Score method, we quantify the level of batch effect between internal and external control samples at each variant within a region. Specifically, we test for an association between each genetic variant and whether a control sample belongs to the internal or external study, while adjusting for covariates. We define a new outcome variable $\widetilde{\boldsymbol{Y}}_{IvE} = (\tilde{Y}_i) = 0/1\ (j = 1, 2, \dots, n_0^I + n_0^E)$ to represent a control sample belonging to the external/internal study, $\boldsymbol{X}_j = \left(X_{j1}, X_{j2}, .., X_{jp}\right)^T$ the covariates for jth subject, and $\boldsymbol{G}_{IvE,j} = \left(G_{1,j}, G_{2,j}, .., G_{n_0^I + n_0^E,j}\right)^T$ be the genotypes at variant j for the $n_0^I + n_0^E$ controls samples. To test for the relationship between the genetic variant and source of control samples, we consider the logistic model

$$logit[\Pr(\tilde{Y}_i = 1 \mid X_i, G_{i,j})] = \boldsymbol{X}_i^T\, \tilde{\boldsymbol{\alpha}} + G_{i,j}^T\, \tilde{\beta}$$

A score test statistic can be constructed to test the null hypothesis of no batch effect between the internal and external controls samples as $S_{IvE} = \boldsymbol{G}_{IvE,j}^T(\tilde{\boldsymbol{Y}}_{IvE}^T - \tilde{\boldsymbol{\mu}}_{IvE})$, where $\boldsymbol{\mu}_{IvE} = (\mu_{IvE,j}) = \left(\Pr(\tilde{Y}_i = 1 \mid \boldsymbol{X}_i)\right)$ and $\tilde{\mu}_{IvE,i}$ is the maximum likelihood estimate of $\mu_{IvE,i}$.

Following the single-variant iECAT-Score method, a compound score statistic that tests the hypothesis no genetic effect at variant j is given by

$$S_{w,j} = a\tau_j S_{int,j} + (1 - \tau_j) S_{all,j} \qquad (2)$$

where $a = \frac{(n_1^I + n_0^I)(n_1^I n_0^I + n_1^I n_0^E)}{n_1^I n_0^I (n_1^I + n_0^I + n_0^E)}$ adjusts for the different sample sizes used to calculate $S_{int,j}$ and $S_{all,j}$, and $\tau_j = \frac{\tau_{1_j}}{1+\tau_{1_j}}$ with $\tau_1 = \frac{S_{IvE_j}^2}{Var(S_{IvE_j})}$ is a variant-specific weight that reflects the level of batch effect existed between the internal and external control samples at the variant j. When minor allele frequencies (MAFs) of external controls are in between those of internal cases and internal controls, and the MAFs are such that $\frac{\hat{\mu}_{all}^2}{\hat{\sigma}_{all}^2} > \frac{\hat{\mu}_{int}^2}{\hat{\sigma}_{int}^2}$, we let $\tau_j = 0$ follwing Li and Lee (Y. Li & Lee, 2021). Under the null hypothesis of no genetic effect, $E(S_{w,j}) = 0$ and $Var(S_{w_j})$ can be calculated using the delta method. Additionally, we update the $Var(S_{w_j})$ to its robust estimate by applying the Saddlepoint approximation (SPA) or Efficient resampling (ER) method, allowing for scenarios of unbalanced case-control ratio and low MAC.

After obtaining the iECAT-Score statistic at each variant within a region, we test the joint genetic effect of variants by performing Burden-, SKAT-, and SKAT-O-type tests to the region. Consider m variants in a region. Let $\boldsymbol{S_w} = (w_1 S_{w,1}, w_2 S_{w,2}, \dots, w_m S_{w,m})'$, where S_j is the single variant iECAT-Score statistic at variant j, $j = 1, 2, \dots, m$. Under the null hypothesis of no genetic effect, $\boldsymbol{S}$ approximately follows a multivariate Gaussian distribution with mean zero and variance $\boldsymbol{\Phi} = \boldsymbol{W\Sigma W}$, where $\boldsymbol{W} = \{w_j\}$ is the diagonal weight matrix and $\boldsymbol{\Sigma}$ is the covariance matrix of $(S_1, S_2, \dots, S_m)'$. We use $w_j = Beta(MAF_j, a_1, a_2)$ where $(a_1, a_2) = (1, 25)$ with MAF_j estimated based on the combined samples. Such choice of (a_1, a_2) upweights rare variants (MAF less than 1%) while giving adequate nonzero weights to less common variants (MAF 1%-5%) (Wu et al., 2011). The covariance matrix of $(S_1, S_2, \dots, S_m)'$ is given by $\boldsymbol{\Sigma} = \boldsymbol{A}^{1/2}\boldsymbol{C}\boldsymbol{A}^{1/2}$, where $\boldsymbol{A} = diag\{Var(S_1), Var(S_2), \dots, Var(S_m)\}$ and $\boldsymbol{C}$ is a $m \times m$ correlation matrix of m variants. As we are interested in maintaining the correlation structure between the variants reflected

through the internal sample population, we estimate the correlation matrix $\boldsymbol{C}$ by the empirical correlation between genetic variants within the region using exclusively the internal case and control samples.

The SKAT statistic is $Q_{SKAT} = \sum_{j=1}^{m} w_j^2 S_j^2$ and the burden test statistic is $Q_{Burden} = \left(\sum_{j=1}^{m} w_j S_j\right)^2$. The weighted average of the SKAT and burden test statistics is $Q(\rho) = (1-\rho)Q_{SKAT} + \rho Q_{Burden}$. Let $\boldsymbol{R}_\rho = (1-\rho)\boldsymbol{I} + \rho \boldsymbol{1}\boldsymbol{1}'$ and $\boldsymbol{L}_\rho$ be its Choleskey decomposition matrix such that $\boldsymbol{L}_\rho \boldsymbol{L}_\rho' = \boldsymbol{R}_\rho$. Then $Q(\rho)$ asymptotically follows a mixture of chi-square distributions, $\Sigma_{j=1}^{m} \lambda_j \chi_j^2$, where $(\lambda_1, \ldots, \lambda_m)$ are the eigenvalues of $\boldsymbol{L}_\rho \boldsymbol{\Phi} \boldsymbol{L}_\rho'$. We apply the Davies method (Davies, 1980) to obtain the *p* value for association between the genetic region and phenotype.

Minimum P-value Based on Combination of S_w and S_{int}

Similar to the single-variant iECAT-Score test, the iECAT-Score method may yield a larger *p* value than exclusively using internal samples only as a result of large variance estimate (Y. Li & Lee, 2021). Thus, to improve power, we use the minimum *p* value calculation procedure (Conneely & Boehnke, 2007; Y. Li & Lee, 2021). At variant j, $(S_{w,j}, S_{int,j})$ jointly follow a multivariate normal distribution. The *p* value of the minimum *p* value of $S_{w,j}$ and $S_{int,j}$, i.e. minP *p* value, are calculated as the probability of observing one or both the *p* values as small as the smaller one of the two under the null hypothesis of no association (Conneely & Boehnke, 2007).

For region-based test, we first obtain the minP *p* value of $(S_{w,j}, S_{int,j})$ $(j = 1, 2, \ldots, m)$ in each variant to re-construct the single variant score statistics, i.e., $S_{minP,j}$, and then combine them for association analysis. One issue of this approach is that to estimate $S_{minP,j}$ from minP *p* value, the variance of $S_{minP,j}$ should be specified. To address this, we use the geometric mean of the score statistics using internal samples only and using the iECAT-Score method, i.e., $Var(S_{minP,j}) = \sqrt{Var(S_{w,j}) \times Var(S_{int,j})}$. This choice of the geometric mean does not only reflect that $Var(S_{minP,j})$ is on the same scale as $Var(S_{w,j})$ and $Var(S_{int,j})$, but also takes into consideration the correlation among $S_{w,j}$, $S_{int,j}$, and $S_{minP,j}$. The minP score statistic $S_{minP,j}$ is then derived by $S_{minP,j} = Var\left(S_{minP,j}\right) \times \chi^2_{quantile}(1 - p_{minP})$. The score statistics and their

variances at each variant are then used to calculate the SKAT-, Burden-, and SKATO-type statistics for the region.

3.2.2 Type I error and power simulations

We conducted simulation studies of various scenarios to evaluate the performance of the proposed iECAT score region-based test regarding type I error rates and power. We used the coalescent simulator COSI (Schaffner et al., 2005) to generate genotyped data of 3000 bps of European ancestry on samples sizes of two combinations of case-control ratios ($n_1^I : n_0^I : n_0^E$): (1) 5,000: 5,000: 10,000; (2) 6,667: 3,333: 10,000.

For both type I error and power simulations, we generated binary phenotypes of case/control from the logistic regression model:

$$logit[\Pr(Y = 1 \mid \boldsymbol{X}, \boldsymbol{G})] = \alpha_0 + 0.5X_1 + 0.5X_2 + \boldsymbol{\beta G}$$

where X_1 was a binary covariate following a Bernoulli distribution with probability of 0.5 being 1, X_2 was a continuous covariate following the standard normal distribution, and α_0 was chosen such that the disease prevalence was 0.01. $\boldsymbol{G}$ consist of variants of 3kb regions randomly selected from the 3kb regions generated by the coalescent simulator.

We assumed that 3% of the variants were subject to different MAFs in internal and external control samples to mimic the batch effects between internal and external samples. When batch effect existed at a variant, the MAFs of the external controls were randomly generated from $Uniform(0.1 \times q, 4 \times q)$, where q was the MAF of corresponding variants in the internal samples.

In type one error simulations, the genetic effect size $\beta = 0$. We generated 5×10^6 datasets to evaluate type I error rates at 1.0×10^{-4} and 2.5×10^{-6} levels. In power simulations, we randomly selected 5%, 10%, 20%, and 50% of variants with MAF < 1% in the 3kb region as causal variants. The effect size of causal variants $\beta = c|\log_{10} MAF|$ where $c = 0.6$, 0.46, 0.35, 0.27 when 5%, 10%, 20%, 50% of the rare variants were causal. We assumed that either all causal SNPs had positive effect (homogeneous effect), or 80% had positive effect and 20% negative (heterogeneous effect). We generated 100,000 data sets in each simulation setting and case-control ratio to evaluate power at the significance level of 2.5×10^{-6}.

3.2.3 Real data analysis

We applied our proposed method to genotype data from the International AMD Genomics Consortium (IAMDGC) (Fritsche et al., 2015) downloaded from dbGaP (phs001039.v1.p1). The IAMDGC dataset consists of 17,286 cases and 14,377 controls. We used 348,465 unrelated samples from the UK Biobank as external controls. We used ICD-9 code to select samples from UK Biobank who are free from macular degeneration of retina and posterior pole of retina. For both studies, the samples used in our analysis are of European ancestry.

We performed analyses on the genotype data of overlapping variants between the AMD and UK Biobank studies to compare the performance of our proposed iECAT-Score method with the method that solely uses internal samples. We applied the ANNOVAR software (K. Wang, Li, & Hakonarson, 2010) for gene-based annotation, using the hg19 build downloaded from the UCSC Genome Browser Annotation Database (Haeussler et al., 2018). We included exonic, intronic, splicing, and UTR variants for analyses.

We applied the Fruposa software (Zhang et al., 2020) with the 1000 Genomes reference (The 1000 Genomes Project Consortium, 2015) to obtain population principal component scores. We used a logistic regression model to test for the association between the disease status of age-related macular degeneration (AMD) and single common genetic variants that are shared by IAMDGC and UK Biobank data sets, adjusting for age, sex, and first four principal components. Then we tested for association between rare variants within genes and the phenotype, conditioned on significant (p value <1e-06) common variant within 3kb region of the gene, based on single-variant association results using the iECAT-Score minP method. In the region-based test, we adjusted for age, sex, and first four principal components. We compared the performance of iECAT-Score, iECAT-Score minP, and methods that exclusively use internal samples and that naively combines control samples without adjusting for batch effect.

3.3. Results

3.3.1 Type I error and power simulations

We present in **Table 3.1** the type I error rates of the proposed methods at the significance level of 1e-04 and 2.5e-06, for two settings of case-control ratios ($n_1^I: n_0^I: n_0^E$): (1) 5,000: 5,000:

10,000; (2) 6,667: 3,333: 10,000. For each setting, we present type I error rates of SKAT, burden, and SKAT-O type tests using test statistics that are constructed using the following methods: (1) exclusively using internal samples; (2) iECAT-Score method; (3) iECAT-Score minP method; (4) naively combining external controls without adjusting for batch effect. The results show that for all of SKAT, burden, and SKAT-O type tests, both versions of iECAT-Score methods controlled type I error rates at both significance levels, although iECAT-Score methods tended to be more conservative than the method that exclusively uses internal samples. If external control samples are naively integrated without adjusting for batch effect, however, substantial inflation of type I error rates is observed.

Table 3.1: Type I error rates of iECAT-Score region-based tests.
Comparison of type I error rates of score tests using internal samples exclusively, various versions of iECAT-Score method, and using combined samples without adjusting for batch effect, under different scenarios. Type I error rates were estimated based on 5×10^6 simulations.

Internal cases: internal controls: external controls	α level		Internal only	iECAT-Score	iECAT-Score MinP	All
5000:5000:10000	1e-04	SKAT	9.73e-05	5.27e-05	5.27e-05	9.13e-02
		Burden	9.86e-05	5.85e-05	2.98e-05	2.08e-02
		SKAT-O	1.14e-04	6.80e-05	5.75e-05	8.85e-02
	2.5e-06	SKAT	3.10e-06	8.61e-07	1.89e-06	7.15e-02
		Burden	2.75e-06	1.55e-06	8.61e-07	1.02e-02
		SKAT-O	2.58e-06	1.89e-06	1.38e-06	6.98e-02
6667:3333:10000	1e-04	SKAT	1.03e-04	4.12e-05	5.75e-05	1.10e-01
		Burden	1.04e-04	4.50e-05	3.78e-05	3.37e-02
		SKAT-O	1.22e-04	4.67e-05	5.57e-05	1.08e-01
	2.5e-06	SKAT	3.24e-06	1.24e-06	2.62e-06	9.30e-02
		Burden	3.08e-06	9.25e-07	9.25e-07	1.95e-02
		SKAT-O	4.32e-06	1.08e-06	2.31e-06	9.10e-02

We compared the power of iECAT-Score, iECAT-Score minP methods and method using internal controls exclusively to assess genetic association at alpha level of 2.5e-06, using SKAT, burden, and SKAT-O type tests. As the method that naively combines external controls does not control for type I error rates, we did not include this method in the power comparison. **Figure 3.1** compares powers of different methods for varying percentage of causal rare variants, when all causal SNPs had positive effect (homogeneous effect). Top three panels show powers of such comparison when case-control ratios are $(n_1^I : n_0^I : n_0^E)$ =5,000: 5,000: 10,000; bottom three

panels show a similar power comparison for case-control ratio of 6,667: 3,333: 10,000. In **Figure 3.2**, we show power comparison in a similar manner when 80% of causal variants had positive effect and 20% had negative effect (heterogeneous effect). The results show that in all settings, both iECAT-Score and iECAT-Score minP methods had improved power over the method that used exclusively internal control samples. Under homogeneous causal effect, iECAT-Score method had higher power than iECAT-Score minP method when a small percentage (5%) of rare variants were causal; iECAT-Score minP method had higher power than iECAT-Score method when a large percentage (50%) of rare variants were causal. Under heterogenous causal effect, however, such relationship was reversed.

Figure 3.1: Empirical power comparisons of iECAT-Score region-based tests under homogeneous genetic effect.
Empirical power comparisons of iECAT-Score minP, and Internal only methods for varying percentage of causal rare variants, when all causal SNPs had positive effect (homogeneous effect). Shown are power from simulations using 10,000 internal samples and 10,000 external control samples, with internal case and control ratios 1:1 (top panels) and 2:1 (bottom panels).

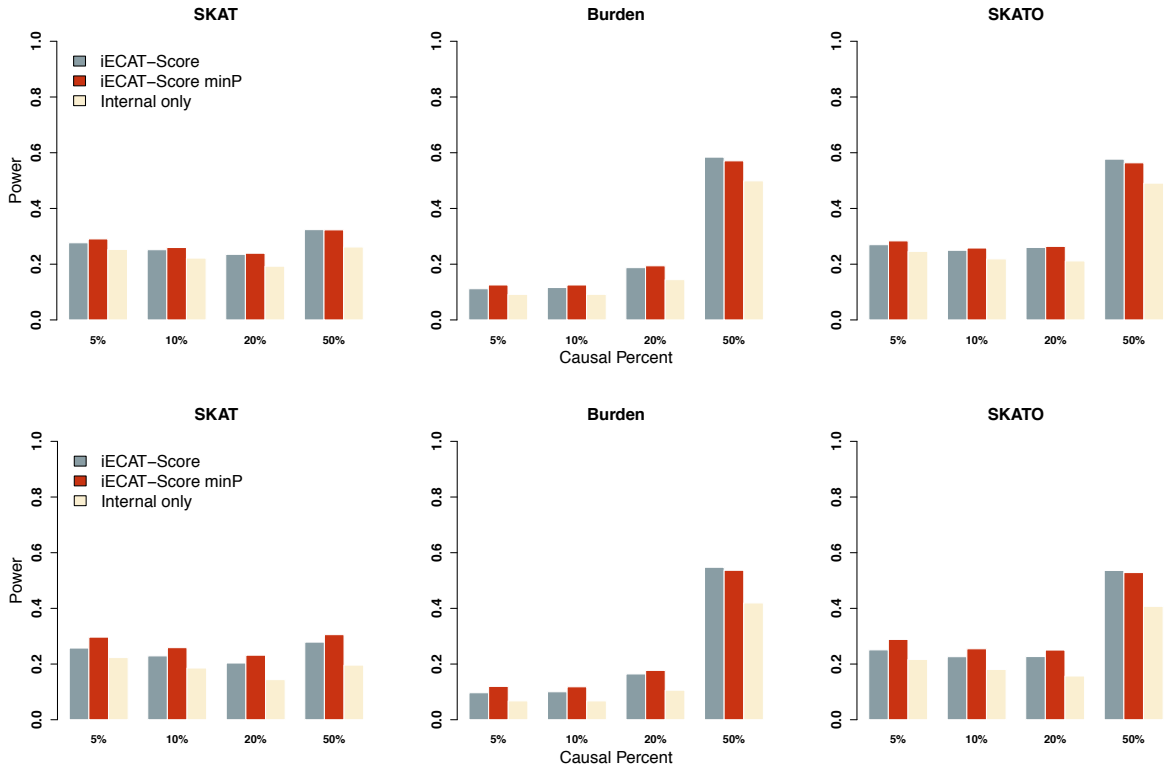
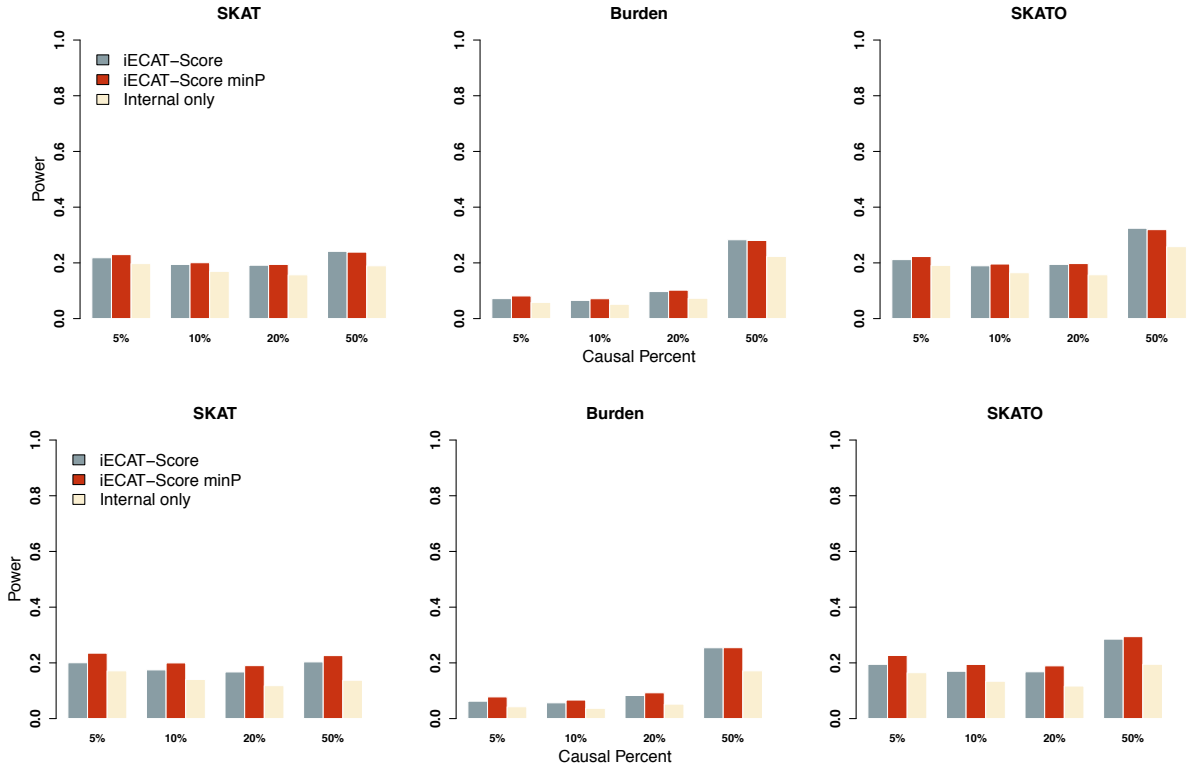


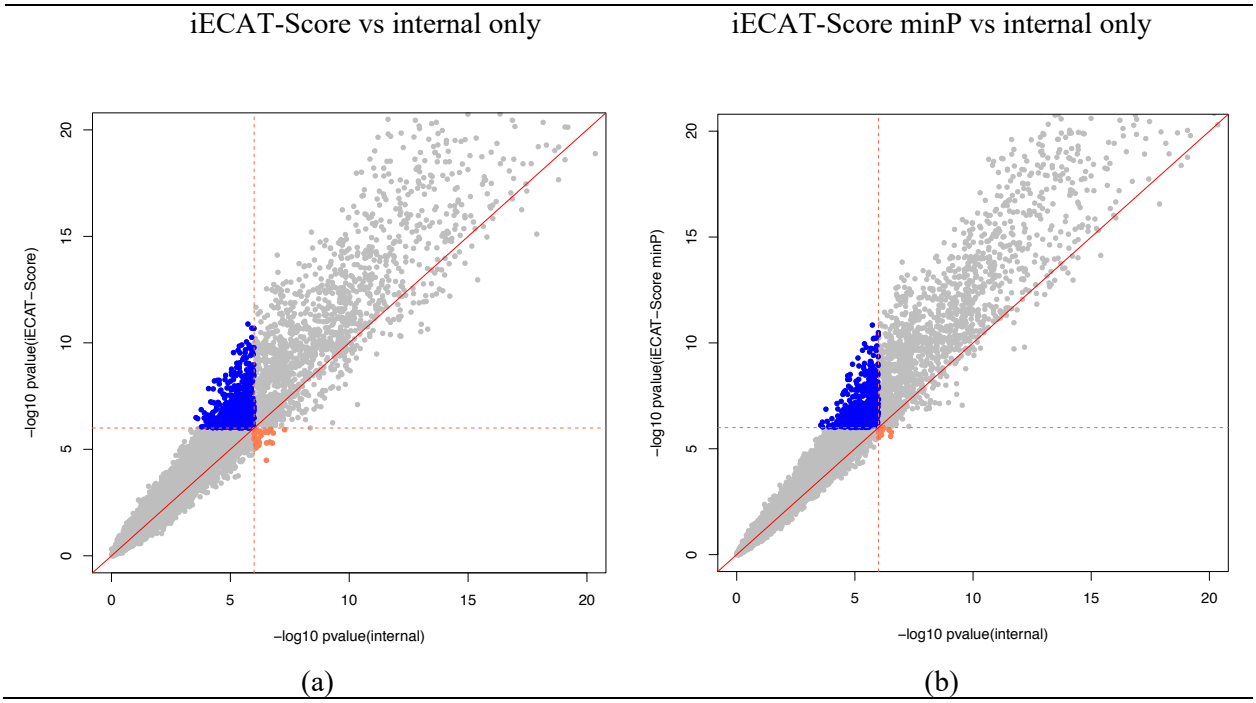
Figure 3.2: Empirical power comparisons of iECAT-Score region-based tests under heterogeneous genetic effect.
Empirical power comparisons of iECAT-Score, iECAT-Score minP, and Internal only methods for varying percentage of causal rare variants, 80% of causal variants had positive effect and 20% had negative effect (heterogeneous effect). Shown are power from simulations using 10,000 internal samples and 10,000 external control samples, with internal case and control ratios 1:1 (top panels) and 2:1 (bottom panels).



We compared *p* values from the power simulations between iECAT-Score method, iECAT-Score minP method, and the method using internal samples only. **Figure 3.3** shows the log-10 scaled *p* values of SKAT-O test under homogeneous causal effect of iECAT-Score method (panel a) and iECAT-Score minP method (panel b) vs. using internal samples only, when case-control ratios are $(n_1^I : n_0^I : n_0^E)$ =5,000: 5,000: 10,000. Shown in blue color are variants that were not significant at 1e-06 level by solely using internal samples but showed signals of association when iECAT-Score method was used; shown in coral color are variants that showed significance at 1e-06 level by exclusively using internal samples, but did not get picked up by iECAT-Score methods. There were some cases where the internal sample-only approach produced smaller *p* values than iECAT-Score, which is a result of large variance estimates using the iECAT-Score method. However, the iECAT-Score minP approach substantially rescued this

disadvantage by leveraging the minimum *p* value between internal sample-only and iECAT-Score *p* values. In both comparisons, there were significant number of variants that showed significance only through the usage of iECAT-Score methods, indicating higher power of iECAT-Score methods in detecting associations.

Figure 3.3: Comparison of *p* values (in -log 10 scale) from analyses of age-related macular degeneration data using the iECAT-Score region-based methods.
Panel (a): -log 10 scaled *p* values using the iECAT-Score method vs. using internal samples only; panel (b): -log 10 scaled *p* values using the iECAT-Score minP method vs. using internal samples only.



3.3.2 Application to Age-Related Macular Degeneration (AMD) Data

We analyzed the association between genes and age-related macular degeneration from IAMDGC, using samples from UK Biobank as external controls. The UK Biobank data include markers directly genotyped or imputed by the TOPMed reference panel (Taliun et al., 2021). We restricted our analysis to the markers shared by the IAMDGC and UK Biobank samples. All samples used in the analyses are of European ancestry. The female samples consist of 41.29%, 43.99%, and 42.51% in samples of internal cases, internal controls, and external controls, respectively (**Table 3.2**). We observed that samples with AMD tended to be older than the

controls, with mean ages 85.86 years and 70.08 years, respectively. Samples of the external controls had a mean age of 56.75 years.

Table 3.2: Descriptive statistics of study subjects from internal (IAMDGC) and external (UK Biobank) studies.
Shown in the table are the sample sizes, the number (percentage) of female samples, and mean (standard deviation of sample age in years in IAMDGC and MGI data.

Study	Sample Size N		Female N (%)			Age Years (SD; min-max)			Total
	Cases	**Controls**	**Cases**	**Controls**	**Total**	**Cases**	**Controls**	**Total**	
IAMDGC (internal)	17,286	14,377	7,137 (41.29)	6,322 (43.97)	13,459 (42.51)	75.86 (8.10; 50-90)	70.08 (9.71; 35-90)	73.24 (9.32; 35-90)	31,663
UKB (external)		334,088		179,984 (53.87)	179,984 (53.87)		56.75 (8.00; 39-73)	56.75 (8.00; 39-73)	334,088
Total	17,286	348,465	7137 (41.29)	188,039 (53.96)	198,188 (54.19)	75.86 (8.10; 50-90)	57.30 (8.50; 35-90)	58.17 (9.35: 35-90)	365,751

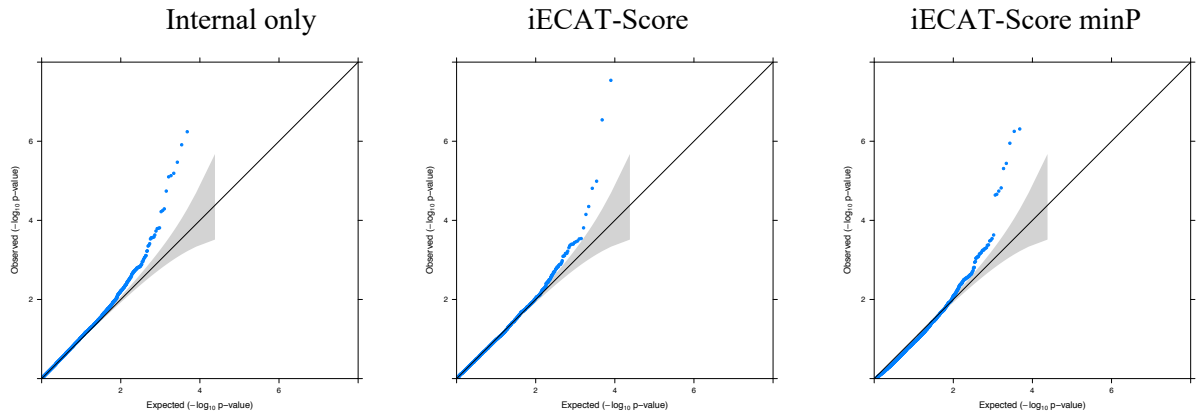
A total of 74,676 variants of exonic, intronic, splicing, and UTR regions are present in both IAMDGC and UK Biobank data sets and were used in single-variant analyses with the iECAT-Score minP single variant test. We then performed region-based tests on 36,389 rare variants (MAF < 0.01) consisting of 7,640 genes which include at least two rare variants, adjusting for age, sex, and first four principal components, conditioning on top common variant (MAF > 0.01) with *p* value < 1e-06 from the single variant test within a 3kb region of the gene. The QQ plots from the tests integrating external control samples using both versions of iECAT-Score methods and using internal samples exclusively are shown in **Figure 3.4**. Compared to the method that exclusively used internal samples, both iECAT-Score methods controlled type I error rates; the patterns of the QQ plots that uses internal samples only and that uses iECAT-Score method are similar; iECAT-Score minP method, on the other hand, is more conservative than the other two methods, which is expected from applying the minimum *p* value method.

Table 3.3. Identification of variants showing significance (6.54e-06 level after Bonferroni correction) based on iECAT-Score minP method.
The iECAT-Score minP jointly tests the rare variants within each gene after conditioning on top common variant whose single-variant association *p* value less than 1e-06, within the 3kb region. Shown are conditional *p* values of analyses from exclusive usage of internal samples and using iECAT-Score minP method, total minor allele count (MAC), and top variant and its respective minor allele frequency (MAF) from single variant analyses, in different case groups.

No.	Gene	Chr	Set Size	Region Test p-value		Total MAC			Top variant	Top variant MAF		
				Internal	iECAT minP	Internal case	Internal control	External control	*p* value	Internal case	Internal control	External control
1	C3	19	29	2.16e-25	8.31e-24	576	273	7060	6.68e-24	1.23e-02	4.30e-03	5.25e-03
2	ASPM	1	15	2.80e-19	1.70e-17	739	868	12642	1.35e-10	6.71e-03	1.22e-02	9.41e-03
3	PRRC2A	6	32	1.50e-06	4.95e-07	1073	686	14535	1.52e-05	1.38e-02	9.91e-03	7.96e-03
4	CFHR5	1	9	6.50e-06	5.68e-07	470	331	6682	2.70e-06	9.68e-03	6.40e-03	4.87e-03
5	F13B	1	4	5.75e-07	1.11e-06	233	295	5261	1.40e-06	6.72e-03	1.03e-02	8.52e-03
6	DXO	6	9	7.38e-05	3.34e-06	342	209	4444	1.29e-04	2.59e-03	1.30e-03	1.47e-03
7	CFB	6	17	3.17e-06	3.49e-06	983	689	12934	1.29e-05	1.42e-02	1.04e-02	8.10e-03

Table 3.3 presents the top genes showing significance at 6.54e-06 after Bonferroni correction. The iECAT-Score method detected several AMD-associated genes including *C3* (Maller et al., 2007; Yates, 2007), *CFHR5* (Narendra, Pauer, & Hagstrom, 2009), *F13B* (Keenan et al., 2015), *CFB* (Sun, Zhao, & Li, 2012), which are well-known associations for the trait. The association of gene *DXO* was revealed by applying the iECAT-Score method (*p* value: 3.62e-06), but did not reach significance level (*p* value: 5.18e-05) with the sole usage of internal samples from the IAMDGC dataset. We show in **Figure 3.5** the single variant association *p* values within the top seven genes that showed association in the conditional rare-variant testing based on the iECAT-Score minP method. We also present the results from the unconditioned association testing results in the Supplementary materials (**Figure S2.1** and **Table S1**).

Figure 3.4: QQ plots for analysis of age-related macular degeneration (AMD).
QQ plots for analysis of AMD from internal study of International AMD Genomics Consortium (IAMDGC) and external control study of UK Biobank. For better visualization, the maximum of x and y axes in the plots are set to be 9, corresponding to *p* values of 1e-09.

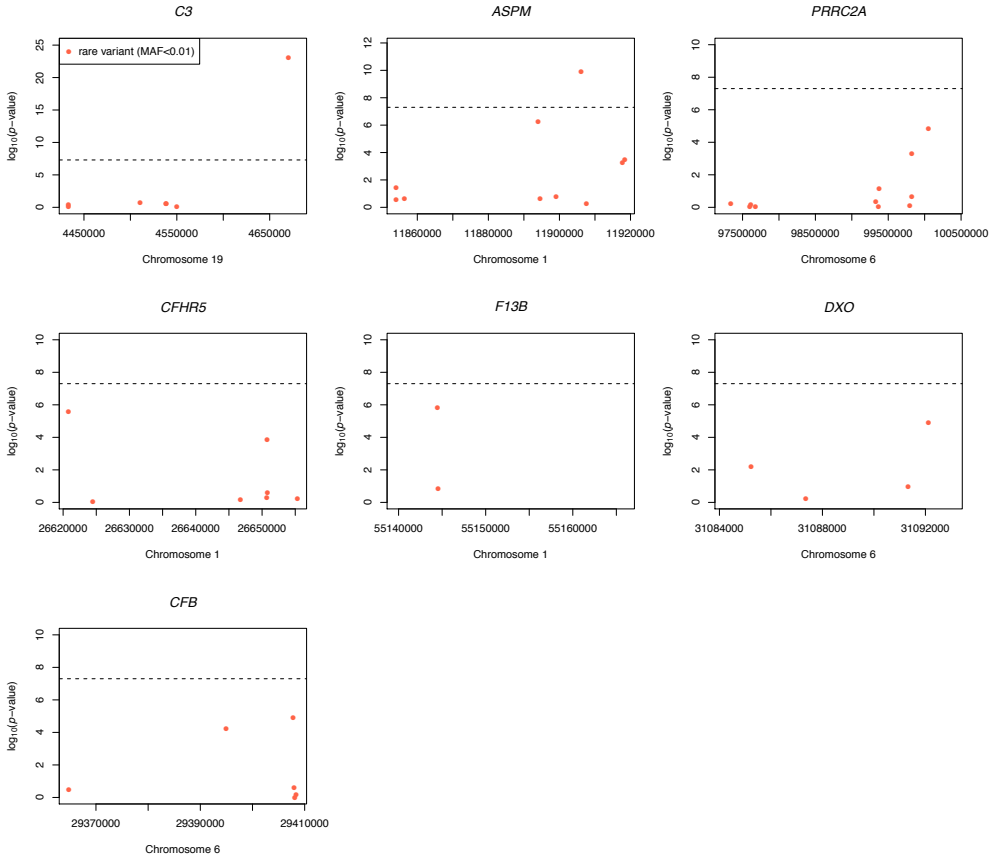


(a) (b) (c)

3.4. Discussion

Integrating External Controls into Association Testing (iECAT) is a powerful tool to increase power in genetic association test and discovering genetic variants that are predisposed to human disease. In this paper, we proposed a region-based iECAT-Score test, which allows for testing the joint effect of rare variants within a gene or a region when integrating genotyped data from external studies. We extended the original iECAT-Score single variant test to three variant-set tests: burden-, SKAT-, and SKATO-type tests. We also took advantage of the minimum *p* value method to further improve the performance of the iECAT-Score test. Our proposed iECAT-Score variant-set tests are not only able to adjust for covariates and population stratification but are computationally efficient when applied to large-scale genome-wide association studies. We showed through simulation studies that our proposed methods have controlled type I error rates and improved power for association testing compared to the methods that exclusively use internal samples. The analysis of the AMD from IAMDGC using UK Biobank exome data as external controls revealed associations that were not found by the sole usage of IAMDGC data.

Figure 3.5: *P* values in -log10 scale of single variants in top seven significant genes from the iECAT-Score minP conditioned rare-variant gene-based test.
The single variant *p* values are calculated using the iECAT-Score minP method, conditioned on significant common variant (*p* value < 1e-06) within 3kb region of the gene. Shown on the x-axis are the positions of each variant within each gene on their respective chromosomes.



Our type I error simulation studies showed that the both versions of the iECAT-Score variant set tests are conservative as compared to the methods that exclusively use internal samples. Such observations are expected as the single-variant iECAT-Score is conservative and both SKAT and burden type tests are conservative at low α levels (B. Li & Leal, 2008; Wu et al., 2011). However, the iECAT-Score tests still improve power for disease association testing. As the true proportion of the true causal rare variants varies in a region, either the SKAT-type or the burden-type test has higher power, consistent with their performance in the usual region-based rare variant tests. The optimal SKAT-O test attains highest power under all possible underlying causal effects of rare variants in the region.

In the data analysis of AMD data from IAMDGC using UK Biobank as external controls, we performed region-based tests on rare variants, conditioning on top common variant (MAF > 0.01) with p values < 1e-06 from the single variant test within a 3kb region of the gene. Although it is possible that there exists more than one independent common variants that are associated with the phenotypes from disjoint linkage disequilibrium blocks, it is less likely for multiple independent common variant associations to occur within a small region of 3kb length. Thus, our choice of conditioning on the top common variant within a 3kb region is reasonable and efficient. Our data analysis revealed a rare variant association of gene *DXO* that was not identified with the sole usage of the internal. The Decapping Exoribonuclease (*DXO*) gene is suggested as a housekeeping gene whose protein function is yet unknown (Jiao et al., 2017; Picard-Jean et al., 2018). However, its association with AMD has been reported through a study based on retina eQTL data (Ratnapriya et al., 2019). Although no definitive pathogenic *DXO* mutations have been found, our analyses shed light on further directions to investigate its roles in the prognosis of AMD.

In this article, we proposed the iECAT-Score region-based test that can improve power for rare-variant association test. The currently presented format of iECAT-Score tests uses genotyped data to assess the batch effect between internal and external samples and test for association. Hence, the performance of the iECAT-Score tests depends on the confidence of the comparison between the two sets of samples using genotyped data. As the sequencing cost continues to drop and large-scale biobanks become available, we ought to be cautious about the quality of genotyped data called from the sequencing data. The quality of genotype data is subject to many factors such as read depths, genotype-calling error rates, quality control (QC)

pipelines, etc., all of which could result in bias in the estimation of minor allele frequencies (MAFs) using genotyped data, leading to batch effect between the two sets of control samples. We will extend our method to account for these factors using sequence data to better account for batch effect between samples from different cohorts.

Through introducing the iECAT-Score region-based test, we expanded the iECAT framework to jointly test for the genetic effects within a gene or a region and improve power for rare-variant tests. When applying the iECAT methods, we require that individual level data including the genotypes and covariates to be available. This requirement on individual level data might pose some challenges when selecting external controls, but as more public biobank data become available, we believe that our iECAT methods will be easily applied. The method is implemented in the R-package "iECAT" available on the GitHub repository.

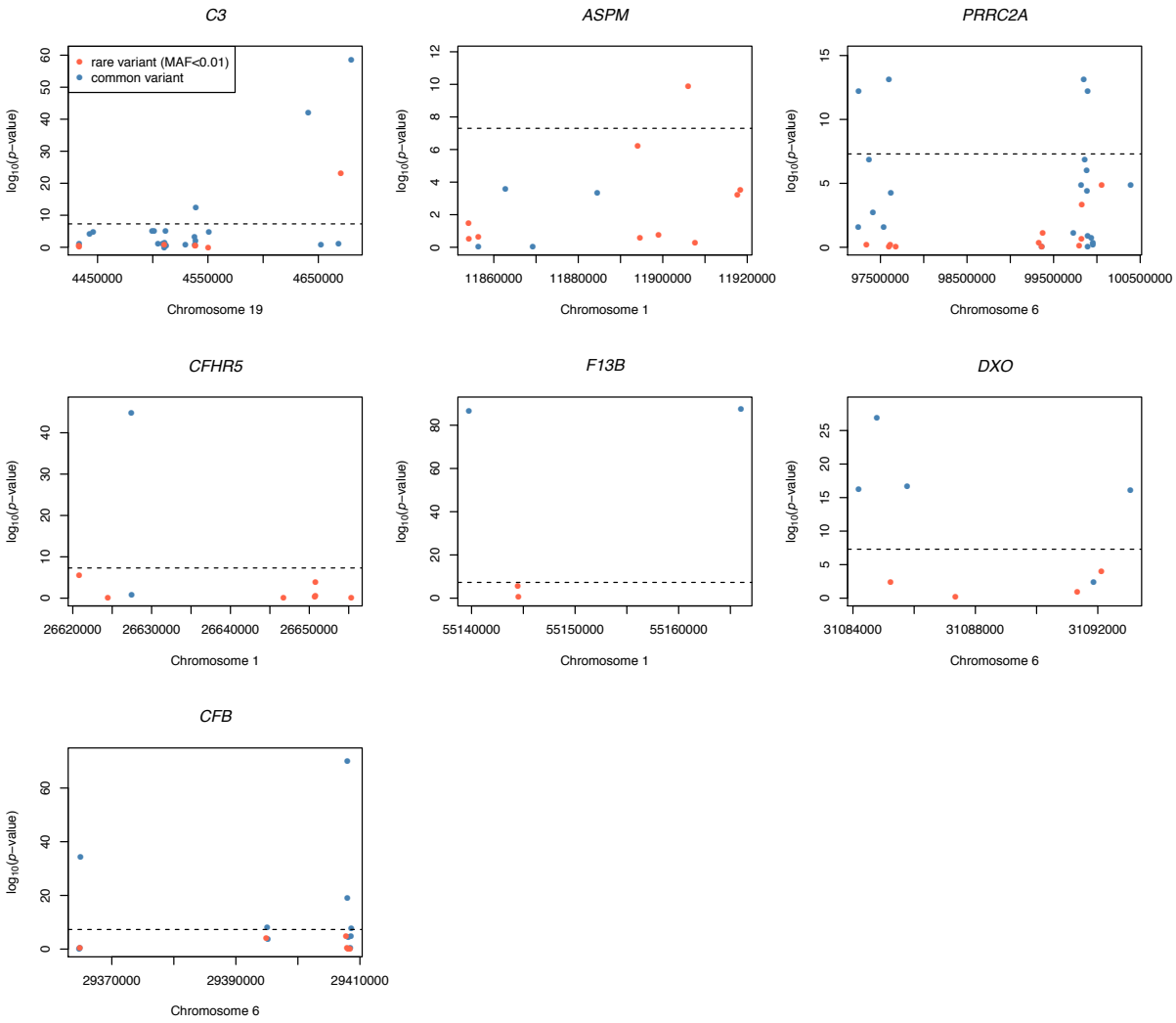
3.5. Supplementary Materials

3.5.1 Supplementary Tables and Figures

Table S3.1: Identification of variants showing significance (6.54e-06 level after Bonferroni correction) based on iECAT-Score minP method, jointly testing the rare variants within each gene.
Shown are marginal *p* values of analyses from exclusive usage of internal samples and using iECAT-Score minP method, total minor allele count (MAC), and top variant and its respective minor allele frequency (MAF) from single variant analyses, in different case groups.

No.	Gene	Chr	Set Size	Region Test p-value		Total MAC			Top variant	Top variant MAF		
				Internal	iECAT minP	Internal case	Internal control	External control	*p* value	Internal case	Internal control	External control
1	C3	19	29	2.28e-25	8.75e-24	576	273	7060	8.82e-24	1.23e-02	4.30e-03	5.25e-03
2	ASPM	1	15	2.80e-19	1.70e-17	739	868	12642	1.35e-10	6.71e-03	1.22e-02	9.41e-03
3	PRRC2A	6	32	1.24e-06	4.89e-07	1073	686	14535	1.46e-05	1.38e-02	9.91e-03	7.96e-03
4	CFHR5	1	9	6.50e-06	5.68e-07	470	331	6682	2.71e-06	9.68e-03	6.40e-03	4.87e-03
5	F13B	1	4	5.83e-07	1.14e-06	233	295	5261	1.40e-06	6.72e-03	1.03e-02	8.52e-03
6	CFB	6	17	3.37e-06	3.46e-06	983	689	12934	1.40e-05	1.42e-02	1.04e-02	8.10e-03
7	DXO	6	9	4.27e-05	6.59e-06	342	209	4444	1.17e-04	2.59e-03	1.30e-03	1.47e-03

Figure S3.1: *P* values in -log10 scale of single variants in top seven significant genes from the iECAT-Score minP marginal rare-variant gene-based test.
The single variant *p* values are calculated using the iECAT-Score minP method. Shown on the x-axis are the positions of each variants within each gene on their respective chromosomes. The single variant association *p*-values for both common (blue) and rare (red) variants within each gene are included.



Chapter 4 Integrating External Controls into Association Analysis Using Sequencing Data

4.1 Introduction

Genome-wide association studies (GWAS) predominantly use array-based genotyping techniques that detect single nucleotide polymorphisms (SNPs) in pre-selected sites (Maróti, Boldogkői, Tombácz, Snyder, & Kalmár, 2018). Recent advances in whole-genome sequencing (WGS) and whole-exome sequencing (WES) technologies have enabled GWAS with a larger number of variants, including many more rare variants. Such advances provide the potential for greater discovery of disease associated variants that might better explain missing heritability of complex traits (Eichler et al., 2010). There are large-scale consortium studies whose WGS/WES data are publicly available. For example, the UK Biobank has WES data on approximately 200,000 individuals from the United Kingdom (Bycroft et al., 2018); the 1000 Genome Project Consortium has reported the genomes of 1,092 individuals from 14 populations analyzed through both WGS and WES data (1000 Genomes Project Consortium et al., 2012). These databases serve as potentially untapped resources of additional control samples to increase the power of association test in case-control studies.

Despite the potential advantages of publicly available whole-genome/exome sequence data for use as external controls, they have predominantly been used for variant discovery before researchers develop microarrays to genotype the target populations (Derkach et al., 2014). Several factors hinder the use of sequence data in similar ways as the microarray data. The common practice of analyzing sequence data involves first aligning the sequence data to common reference genome to create BAM files using SAMtools (H. Li et al., 2009) and then applying variant calling algorithms using tools such as GATK (DePristo et al., 2011) to obtain called genotypes. In whole genome/exome sequencing, samples tend to be sequenced at lower depths overall (4-10x for WGS and ~30x for WES) as compared to microarray data. When

internal and external samples are sequenced at different depths, samples sequenced at lower read depths would be more prone to biased genotype calls. In addition, different genotype calling algorithms could further contribute to such bias. If typical quality control procedures are applied to remove variants with low read depths and low quality, large number of variants could be removed, including valuable causal rare variants.

Several recent methodologic developments have attempted to use sequence data from publicly available control groups for case-control association tests that allow for differential sequencing depths between internal cases and external controls. Derkach (Derkach et al., 2014) proposed a score test that replaces the called genotypes with their expected values given read data and developed a robust variant estimate to prevent type I error inflation. Derkach considered a retrospective setting by treating the true genotype as unknown, but instead built a joint likelihood model relating the phenotype and underlying genotype via the observed sequence data. As such retrospective setting requires the probability model between variables to be specified, it cannot adjust for covariates whose probabilistic function with the genotype is unknown. Extending Derkach's method, Chen and Lin (Chen & Lin, 2018) proposed regression calibration (RC)-based and maximum-likelihood (ML)-based methods to adjust for differential sequencing errors between internal cases and external controls. These methods allow for variant's effect size estimation, with the assumption that weak confounding from population stratification are the only potential confounders. However, similar to Derkach's method, Chen and Lin's method was based on a retrospective model and thus does not allow for covariate adjustment such as age or gender. Hu et al. (Hu, Liao, Johnston, Allen, & Satten, 2016) proposed a likelihood-based method that directly models sequencing reads without calling the genotypes. This method first estimates the single nucleotide variant (SNV) locations and then applies a burden-type test to assess the significance of the association between a SNV and the phenotype. This method, however, still does not allow for covariate adjustment, including population stratification.

One major shared drawback of the above-mentioned methods is their lack of ability for covariate adjustment. Confounders such as age, gender, population stratification could have major impact on the disease risk and thus are essential in assessing the relationship between genetic variants and the phenotype. In Chapter 2 and 3, we proposed the iECAT-Score methods, which are score-based methods that allow for covariate adjustment in the logistic regression

model while integrating external control samples. The iECAT-Score single variant and region-based tests use genotype data to assess the batch effect between internal and external control samples. Hence, the performance of the iECAT-Score tests depends on the confidence of the comparison between the two sets of samples using called genotype data. In sequence data, the quality of genotype calls is subject to many factors such as read depth, genotype-calling error rate, quality control (QC) pipelines, etc., all of which could result in bias in the estimation of minor allele frequencies (MAFs), leading to batch effect between the two sets of control samples. Although QC filters are often applied to include only genotype variants of "good quality" to use in the analysis, such an application does not guarantee that the bias in the estimation of MAFs is removed. It appears that in fact, the bias in the estimation of MAFs could be accentuated by applying stringent QC filters (Derkach et al., 2014). Thus, different QC pipelines applied between internal and external study samples could result in a more profound batch effect. In addition, applying QC filters could drastically reduce the number of variants available for analyses, especially for low-to-medium coverage sequencing (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012; Nielsen, Paul, Albrechtsen, & Song, 2011).

Compared to the hard called genotype, which is the most likely genotype given the read data, the expected genotype, or genotype "dosage", is calculated to be the weighted average of all possible genotypes given their respective posterior genotype likelihood and accounts for the uncertainty about the true genotype (Zheng, Li, Abecasis, & Scheet, 2011). Given sequence read data, genotype likelihood could be calculated as a function of read depth and genotype-calling error rate. Hence, genotype dosage, which inherently contains information on the uncertainty of underlying true genotypes given sequenced read data, offers a potential tool improve the performance of iECAT-Score methods when applied to whole genome/exome sequence data.

In this chapter, we first simulated the effect of various factors in the genotype calling pipeline on the estimation of MAFs (Sections 4.2.1-4.2.2). The major factors of interest included: read depth, genotype-calling error rate, and genotype quality score (often used in QC filtering procedure). Then we assessed the distributions of these factors in real data, the observed inconsistency of MAF estimation using called and expected genotypes (Section 4.2.3), and how these observations inspired us to integrate the genotype likelihood in our methods of iECAT-Score tests (Section 4.2.4). We discussed strategies to replace called genotypes with genotype dosages in the iECAT-Score testing framework (Section 4.3.1) and performed simulation studies

to compare the performance of the tests using called genotypes and genotype dosages in sequence data (Section 4.3.2). Then we applied our proposed methods to sequence data of the Myocardial Infarction Genetics Exome Sequencing Consortium (Myocardial Infarction Genetics Consortium, 2009) and the UK Biobank (Bycroft et al., 2018) (Section 4.3.3). We present the results from simulation studies and read data analysis in Section 4.4.

4.2 Exploratory Investigation: Variables in Genotype Calling Pipelines

4.2.1 Read depth, base-calling error rate, and genotype quality score

In this section, we describe the key steps in the sequencing and genotype calling procedures where several parameters have an impact on the final called genotype. For a given locus, the sequencer detects the existence of a certain base by reading multiple number of times (*read depth*). Then based on this read data, a posterior probability of the genotype at this locus is calculated, implying the relative likelihood of the true genotype being 0/0, 0/1, or 1/1. Finally, a called genotype is assigned based on the posterior probability of the genotype.

Consider a study of sample size n. We assume a true unobserved genotype G_{ij} for individual i at locus j. The joint likelihood model of the phenotype Y_i and the sequence data D_{ij} is given by $\Pr\left(\boldsymbol{Y} = (Y_1, \dots, Y_n),\ \boldsymbol{D} = \left(D_{1j}, \dots, D_{nj}\right)\right) = \prod_{i=1}^{n}\left(\sum_{g=0}^{2} \Pr(Y_i|G_{ij} = g)\Pr(G_{ij} = g,\ D_{ij})\right)$. The sequence read data D_{ij} consists of r_{ij} reads, where each read randomly picks up the base of one of two alleles at the locus G_1 or G_2. Hence, given the true genotypes G_{ij}, the likelihood of observed reads D_{ij} is defined as $P\left(D_{ij} = (g_1, \dots,\ g_r)\middle|G_{ij} = G_1G_2\right) = \prod_{k=1}^{r_{ij}}\left(\frac{1}{2}P(g_k|G_1) + \frac{1}{2}P(g_k|G_2)\right)$, where $P(g_k|G_l), l = 1, 2$, is the probability for a sequencer to detect a base given the allele G_l. When reading a base at a locus for the k-th time among the r reads, there is a chance that the true base is not being correctly detected with a *base-calling error rate*: the probability of the sequencer not picking up the base given the allele G_l is denoted as $e_{k_{10}}$; the probability of the sequencer picking up the base when the allele is not G_l is denoted as $e_{k_{01}}$. Thus, given a true allele G_l, the probability of seeing a base from a sequencer is $P(g_k|G_l) =$

$\begin{cases} e_{k01},\ g_k \neq G_l \\ 1 - e_{k10},\ g_k = G_l \end{cases}$. In a VCF file, the locus-level parameter *QUAL* is the Phred-scaled probably of the base-calling error rate. Hence, a value of 10 indicates one in ten chance of base-calling error.

Using the conditional probability described above, a posterior probability of genotype G_{ij} given the read data D_{ij} consisting of r reads, often referred to as the *posterior genotype likelihood,* is calculated using the Bayes rule $P\left(G_{ij} = g_{ij} \middle| D_{ij}\right) = \frac{P(G_{ij})P(D_{ij}|G_{ij})}{P(D_{ij})}$ (Nielsen et al., 2012). After obtaining the posterior probabilities for all possible genotypes $g_{ij} = 0/0, 0/1, 1/1$, we assign $G_{(1)}$ and $G_{(2)}$ to be the two genotypes with the two highest posterior likelihoods. To describe the confidence of the genotype being $G_{(1)}$ over $G_{(2)}$, a log-odds score calculated as the log10-scaled ratio of the largest two posterior probabilities by $\log_{10} \frac{\Pr\,(G_{(1)}|read\ data)}{\Pr\,(G_{(2)}|read\ data)}$. If this log-odds score is larger than a filtering threshold *R*, then the called genotype is assigned to be $G_{(1)}$; if this log-odds score is smaller than *R*, the called genotyped is assigned to be missing due to low confidence. In a VCF file, the sample-level parameter *PL* indicates the Phred-scaled posterior likelihood of each genotype; the *genotype quality score GQ* is the difference between the *PL* scores of the second most likely and the most likely genotypes. Hence, if $GQ > R$, the genotype is being called as $G_{(1)}$; if $GQ < R$, the genotyped is assigned to be missing.

Hence, the called genotypes a discrete function of *GQ*, taking values of either $G_{(1)}$ the "best guess" or missingness. In practice, a quality control filter is usually applied before using called genotype data for analyses (Lazaridis et al., 2014). Sometimes a VCF file include fields such as read depth, the Phred-scaled posterior likelihood of each genotype and/or the genotype quality scores for each sample at each locus. An alternative genotype dosage, or the "expected genotype" is calculated by taking the expectation over the posterior distribution of G_{ij} given read data D_{ij}: $\mathrm{E}\left(G_{ij} \middle| D_{ij}\right) = \sum_{g=0}^{2} g \times P\left(G_{ij} = g_{ij} \middle| D_{ij}\right)$. As compared to the called genotype which takes discrete values of 0, 1, 2, or *NA*, the genotype dosage takes continuous values ranging from 0 to 2.

4.2.2 Biased MAF estimation using called genotypes - effects of read depths, base calling and genotype quality score

To investigate whether MAF estimation using called genotypes could be influenced by read depths, base calling error rates, and genotype quality score filter, we performed simulation studies to assess the ratio between the MAF estimated using called genotypes and the "true" genotypes, as well as the ratio between the MAF estimated using expected genotype and the "true" genotypes by setting different values of these parameters. Specifically, we focused on rare variants, ranging the MAF from 0 to 0.01. The base-calling error rates were set to be 1e-02, 1e-03, and 1e-04, representing the site-level *QUAL* score of 20, 30, and 40, respectively. We set *R* to be 0, 1, 2, and 5 as the filter being applied to the *GQ* score: when the *GQ* score is below *R*, the genotype is assigned to be missing; when *R* is equal to 0, there is no missing genotype calls.

We present in **Figure 4.1** the ratio between the MAF estimated using either called genotype or genotype dosages and the "true" genotypes when the true MAF ranged from 0 to 0.01 for different base-calling error rates and read depths, fixing the *GQ* filter at 20. The MAFs calculated from called genotypes tended to underestimate the true genotypes, especially for sites of lower read depths and higher base-calling error rates; the bias of MAF estimation using genotype dosages, on the other hand, was smaller compared to using called genotypes. **Figure 4.2** shows the ratio of estimated MAF and "true" genotypes using called genotypes and genotype dosages, when various filter *R* was applied on the *GQ* score during genotype calling. As we used more stringent filter to call genotypes, higher read depths were required to achieve reliable estimation of MAF using called genotypes. Hence, given sequence data, applying filters on genotype quality does not necessarily make MAF estimation closer to the true MAF in called genotypes. When read depth was low and in rarer variants, applying a filter on the *GQ* score could further underestimate or overestimate the MAFs using called genotype.

Figure 4.1: Ratio of estimated MAF and "true" genotypes using called genotypes and expected genotypes, for different base-calling error rates and read depths.
MAFs were estimated based on 10,000 samples.

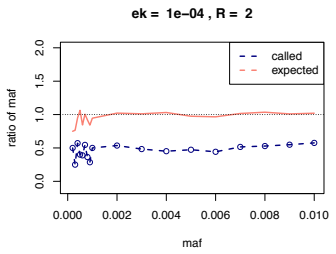
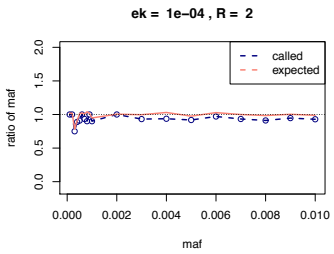
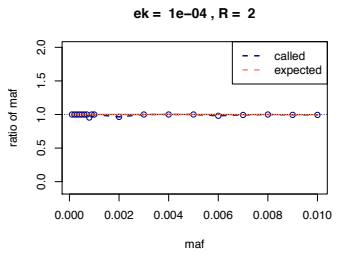
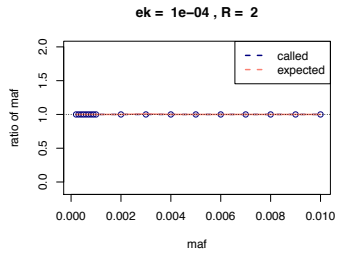
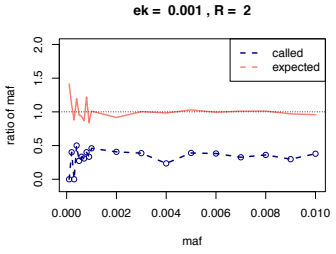
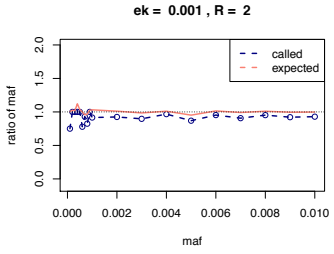
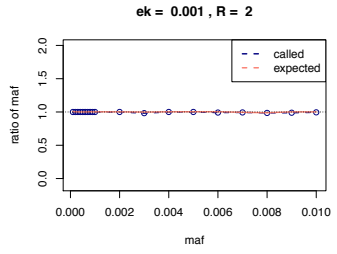
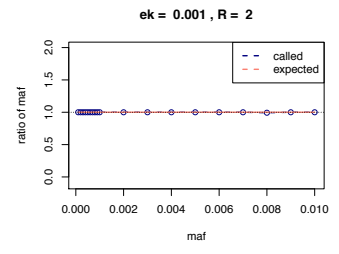
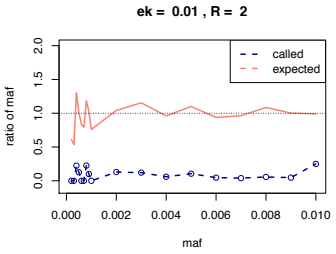
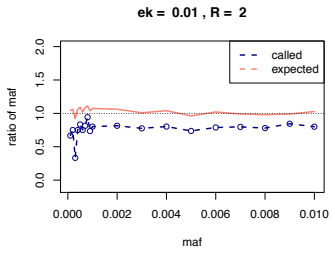
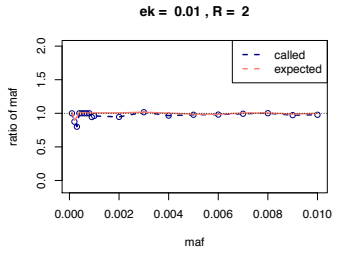
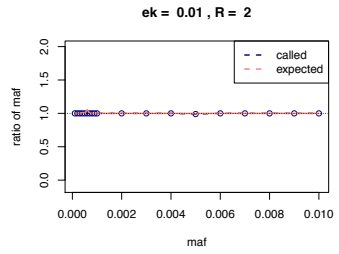
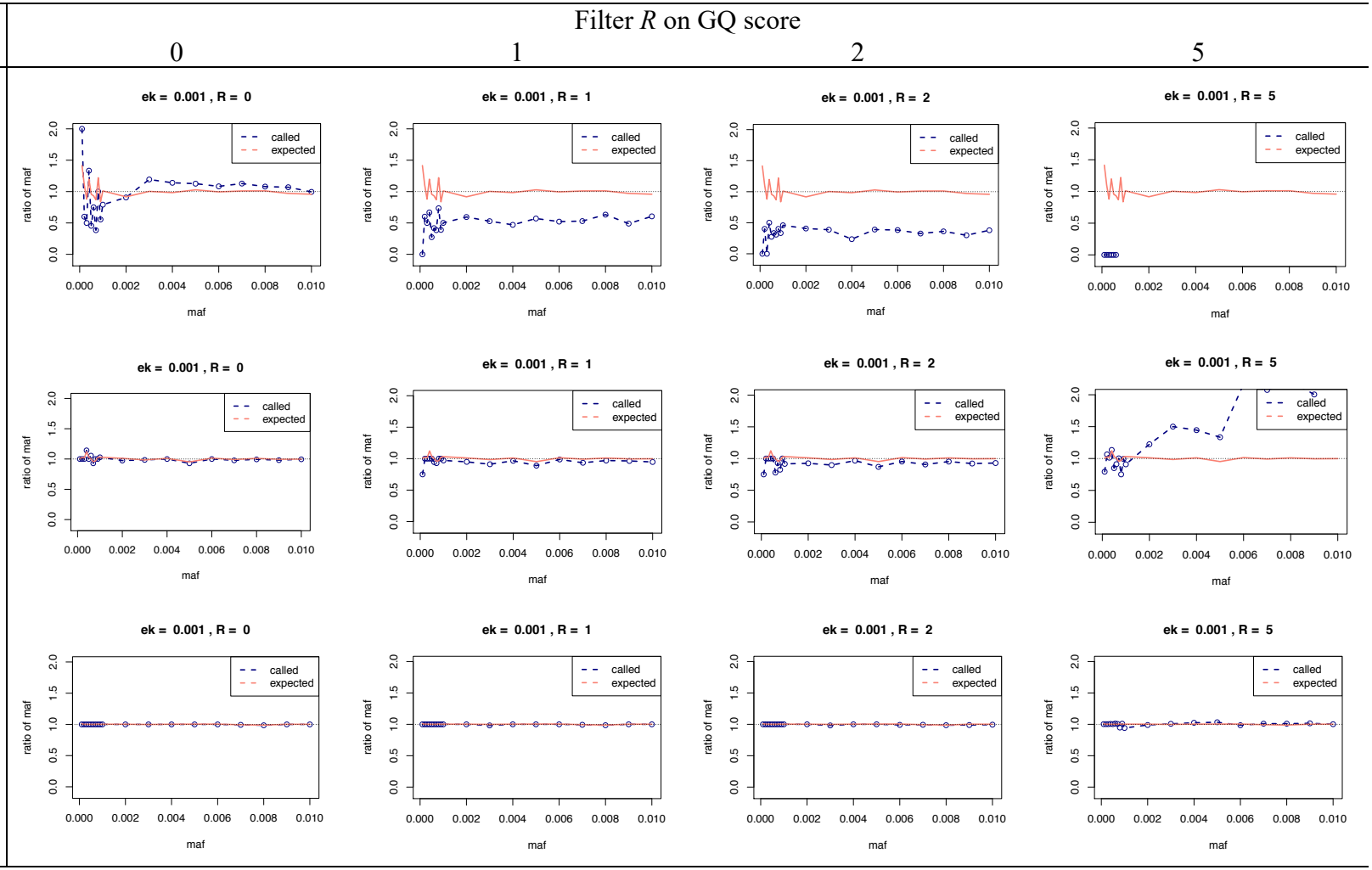
Base-calling error rate	Read Depth 4	10	20	40
1e-04				
1e-03				
1e-02				

Figure 4.2: Ratio of estimated MAF and "true" genotypes using called genotypes and expected genotypes, when varying filter *R* was applied on the *GQ* score when calling genotypes.
MAFs were estimated based on 10,000 samples.



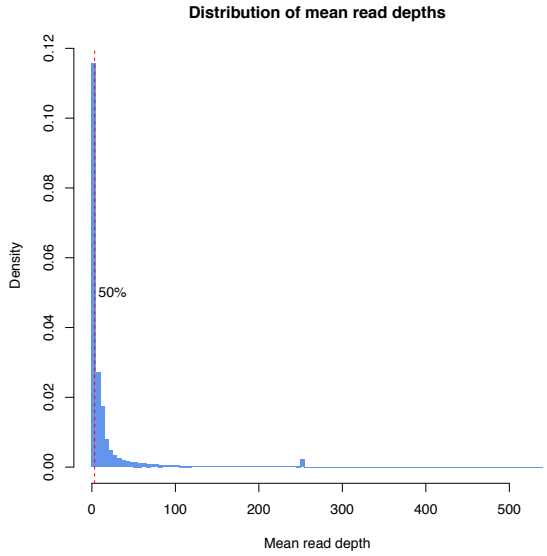
4.2.3 Distributions of genotype calling pipeline parameters and MAFs in real data

We used sequencing data of the Myocardial Infarction Genetics Exome Sequencing Consortium (Myocardial Infarction Genetics Consortium, 2009) downloaded from dbGaP and the GotCloud variant calling pipeline (Jun, Wing, Abecasis, & Kang, 2015) to investigate distributions of read depths, base calling error rates, and genotype quality score observed in real data. This case-control study consisted of 2,322 subjects with 1,452 cases of myocardial infarction and 1304 controls. Exome sequencing was performed on the Illumina HiSeq system.

For this investigation, we performed variant calling using GotCloud on sequencing data of 537 samples for chromosomes 21 and 22, and annotated variants of exonic regions using the ANNOVAR software (K. Wang et al., 2010). A total of 200,383 variants including 8,642 variants of exonic regions passed the SVM filter (Jun et al., 2015). We first checked the distributions of mean read depths at all variants and at variants of exonic regions, Phred-scaled base-calling error rates at all variants, and *GQ* scores at all samples (**Figure 4.3**). The median of the mean read depths was 3.29, showing a large proportion of variants with low average read depths across samples (**Figure 4.3**, panel a), which was expected from whole exome sequencing data. The median of the mean read depths in the exonic regions was 100 and the 5% quantile of the mean read depths in the exonic regions was 17 (**Figure 4.3**, panel b). The 5% quantile of *QUAL* scores was 28, i.e., more than 5% of all sites had an estimated base-calling error rate greater than 0.001 (**Figure 4.3**, panel c). The 10% quantile of *GQ* scores was 22, indicating that 10% called genotypes would be missing when we set a filter *R* of 2.2 in genotype calling (**Figure 4.3**, panel d).

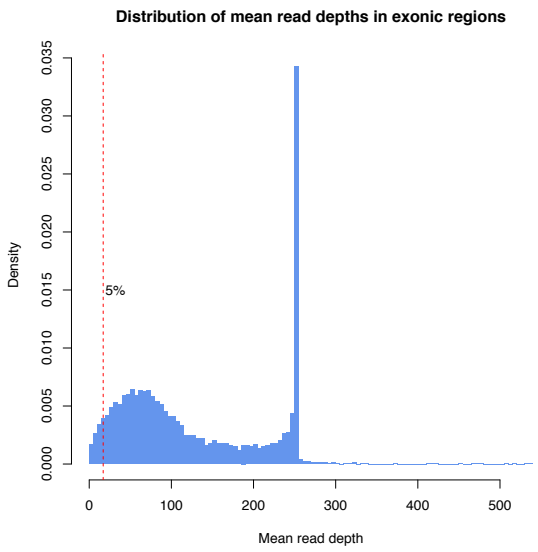
Figure 4.3: Distributions of mean read depths, base-calling error rates, and GQ scores in Myocardial Infarction Genetics Exome Sequencing Consortium data.
Distributions of mean read depths at all variants (a), mean read depths at variants of exonic regions (b), Phred-scaled base-calling error rates at variants (c), and *GQ* scores (d) at samples. In panel (a), the red dashed line represent the 50% quantile of mean read depths; in panel (b), the red dashed line represent the 5% quantile of mean read depths at variants of exonic regions; in panel (c), the red dashed lines represent *QUAL* scores corresponding to base calling error rates of 0.01 (left) and 0.001 (right); in panel (d), the red dashed line represent the 10% quantile of *GQ* scores of all variants and samples.

Mean read depth



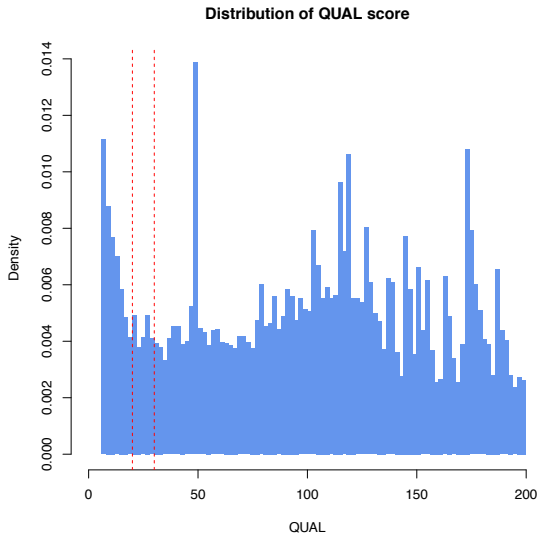
(a)

Mean read depth (exonic regions)



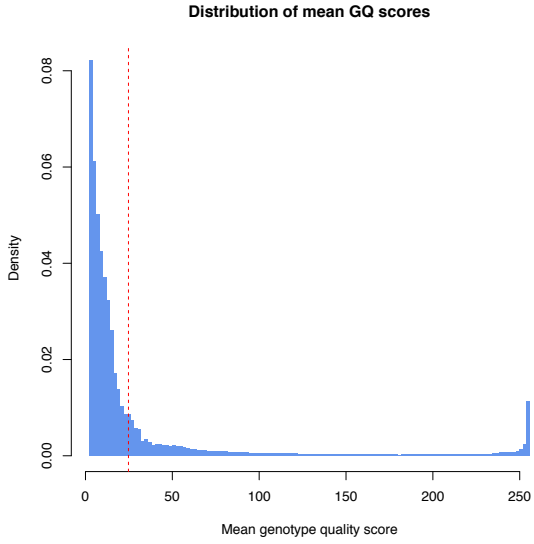
(b)

Phred-scaled base-calling error rate



©

Mean enotype quality (GQ) scores



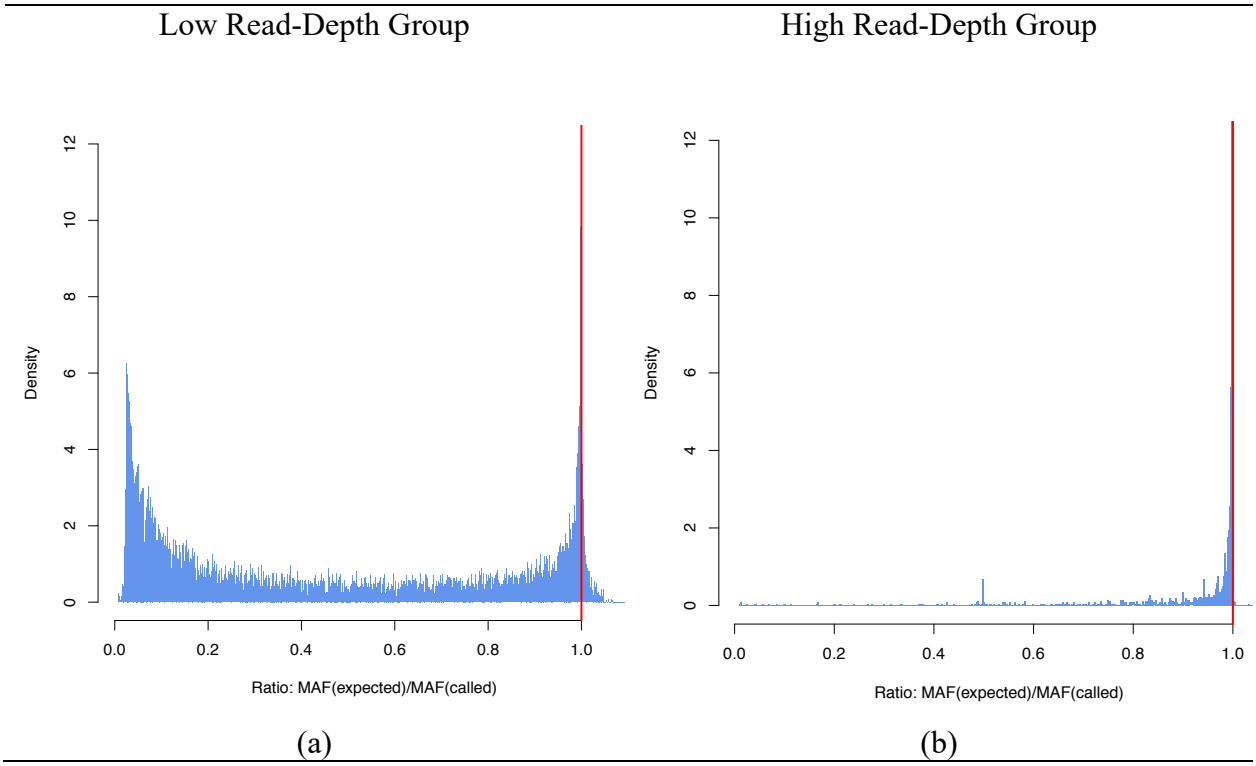
(d)

To check the MAF estimation between variants with low mean read depths as compared to those with high read depths, we first defined variants with mean read depths between five to twenty to be in the low read-depth group (Nielsen et al., 2011), and those with mean read depths greater than 40 to be in the high read-depth group. A total of 13,053 and 13,326 variants fell into the low read-depth group and high read-group, respectively. Within each group, we calculated the ratio of MAFs calculated using called genotypes and expected genotypes (called genotypes MAFs divided by expected genotypes MAFs) at each variant and checked the distribution of such ratios using histograms (**Figure 4.4**). Among the variants within the low read-depth group, the ratios of MAFs calculated using called genotypes and expected genotypes tended to be more scattered less than one; of variants within the high read-depth group, the ratios still tended to be less than one, although they were more centered towards one as compared to the low read-depth group. These observations indicated that the MAFs tend to be underestimated when called genotypes are used; such underestimation could be further accentuated among variants with low read depths.

4.2.4 Implications from the exploratory studies - a proposal for better iECAT-Score methods

Despite the observation that MAFs tend to be underestimated in variants with lower read depths, it would not be a good idea to filter out variants or genotype calls with low (mean) read depths or low (mean) genotype quality score (which is highly correlated with read depth). As we observed from the distributions of mean read depths across variants, most of the variants tend to have low mean read depths across samples. Although the mean read depths in the exonic regions tended to be higher, there were still a significant proportion of exonic variants whose mean read depths were lower than 20x. In fact, in our sample data set, about 6.6% of exonic variants had a mean read depth of 20 or lower, and applying a filter on read depths of 20 would significantly reduce the number of exonic variants that could be used for analyses.

Figure 4.4: Distributions of ratios of MAFs calculated using called genotypes and expected genotypes.
Ratios of MAFs between called genotypes and genotype dosages (called genotypes MAFs divided by expected genotypes MAFs) in low read-depth group (a) and high read-depth group (b).



Derkach *et al* (Derkach et al., 2014) suggested that one main source of bias in the estimation of MAF using called genotypes is the misclassification of rare homozygous genotypes into heterozygous calls. For a fixed read depth, there is a specific MAF, below which rare homozygotes can be misclassified as heterozygous. The higher the read depth is, the MAF threshold below which rare homozygotes are miscalled as heterozygotes decreases. When the read depth is lower, there are more common variants whose "minor" homozygotes get miscalled as heterozygotes. In the practice of analyses using called genotypes, a filter on the genotype quality score is often applied to each genotype calls and genotypes whose GQ scores are under the threshold are assigned to be missing. However, the higher the filtering threshold is, more of the miscalled heterozygous genotypes are assigned to be missing, increasing the weight of common homozygous and heterozygous genotypes, making the MAF even more underestimated.

Hence, in the setting of integrating external controls, in addition to the inherent bias in MAF estimation using called genotypes, the difference in read depths or quality control filters applied between the internal and external studies could lead to more differentiation in the MAF estimations based on called genotypes. When internal and external study samples are from the same population and we assume no population stratification exists, the minor allele frequencies at the same locus are expected to be indistinguishable. If the two samples are sequenced at different depths and disparate genotype called pipelines and quality control filters are applied, it would be possible to observe divergent MAF estimations using called genotypes from the two samples. Such differentiation could lead to more profound batch effect and potentially affect the type I error rate and power in association tests.

Equipped with the implications from the exploratory studies above, we propose to exploit genotype dosage in the iECAT-Score methods. Genotype dosages are calculated from the posterior genotype likelihood, which is a function of read depths, base-calling error rates; the use of genotype dosages eliminates the need for applying a *GQ* score filter. Genotype dosages also result in less bias in MAF estimation, especially in loci of low read depth and possible base-calling errors. By integrating genotype dosages, we hope to increase the number of variants available for analyses, reduce the bias in MAF estimation, obtain a better assessment of batch effect between internal and external samples, and thus improve the performance of the iECAT-Score method.

4.3 Methods

4.3.1 Applying the posterior genotype likelihood to iECAT-Score Tests

We incorporate the posterior genotype likelihood through the genotype dosage to the iECAT-Score test, which is a score test for the variant effect on the phenotype of case or control, combining external control samples. We propose three methods to integrate the genotype dosage in place of the called genotypes into the single-variant iECAT-Score statistics, and strategies to calculate their respective variance estimates and *p* values for association.

Logistic Regression Models and the iECAT-Score Test

The single variant score test for genetic effect in a study of sample size *n* are formulated from the logistic regression model

$$logit[\Pr(Y_i = 1 \mid \boldsymbol{X}_i, G_i)] = \boldsymbol{X}_i^T \boldsymbol{\alpha} + G_i \beta \qquad (4.1)$$

where $y_i = 0/1$ is the dichotomous phenotype for control/case, $\boldsymbol{X}_i = \left(X_{i1}, X_{i2}, .., X_{ip}\right)^T$ are the covariates, and $\boldsymbol{G} = (G_1, G_2, .., G_n)^T$ are the genotypes at a variant for *n* subjects ($G_i = 0, 1, 2$ represent 0, 1, 2 copies of the minor allele). In this equation, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of coefficients for p covariates including an intercept, and β is the genotype effect at the variant of interest. Assessing whether the association exists between the phenotype Y_i and the genotype at a variant is equivalent to testing $H_0: \beta = 0$ in Equation (4.1). To test the association between a single variant and the phenotype, we use the score test statistic

$$S = \boldsymbol{G}^T (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$$

where $\boldsymbol{\mu} = \{\mu_i\} = \{\Pr(Y_i = 1 \mid X_i)\}$ under H_0, and $\hat{\mu}_i$ is the maximum likelihood estimate of μ_i. Under the null hypothesis of no genetic effect, $E(S) = 0$ and $Var(S) = \sum_{i=1}^{n} \tilde{G}_i^2 \hat{\mu}_i (1 - \hat{\mu}_i))$, where $\widetilde{\boldsymbol{G}} = \{\tilde{G}_i\} = \boldsymbol{G} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{G}$ is the covariate-adjusted genotype vector, and $\boldsymbol{V} = diag\{\hat{\mu}_i(1 - \hat{\mu}_i)\}$. Then $\frac{S^2}{Var(S)}$ asymptotically follows χ_1^2, and a p value can be obtained as $p = P(\chi_1^2 > \frac{S^2}{Var(S)})$.

To integrate the external controls while controlling the type I error rates, the iECAT-Score test statistics takes the form of a weighted sum of the score statistics calculated using exclusively internal samples and using the combined samples

$$S_w = a\tau S_{int} + (1 - \tau) S_{all} \qquad (4.2)$$

where $S_{int} = \boldsymbol{G}_{int}^T (\boldsymbol{Y}_{int} - \widehat{\boldsymbol{\mu}}_{int})$, $S_{all} = \boldsymbol{G}_{all}^T (\boldsymbol{Y}_{all} - \widehat{\boldsymbol{\mu}}_{all})$, $a = \frac{(n_1^I + n_0^I)(n_1^I n_0^I + n_1^I n_0^E)}{n_1^I n_0^I (n_1^I + n_0^I + n_0^E)}$ to adjust for the sample sizes, and $\tau = \frac{\tau_1}{1 + \tau_1}$ with $\tau_1 = \frac{S_{IvE}^2}{Var(S_{IvE})}$ assesses the batch effect. The asymptotic distribution of $\frac{S_w^2}{Var(S_w)}$ approximately follows a χ_1^2. Thus, a p value can be approximated by $P(\chi_1^2 > \frac{S_w^2}{Var(S_w)})$. Details on deriving the iECAT-Score statistic and its variance can be found in Chapter 2 and Appendices.

Genotype Dosage in Place of Called Genotype

When posterior genotype likelihoods are available, the genotype dosage (expected genotype) can be calculated as $G_D = \sum_{g=0,1,2} g \times \Pr(G = g|D)$, where $\Pr(G = g|D)$ is the posterior probability of the true genotype is $g = 0, 1, 2$, given read data D. Consider a generic score test statistic $S = \boldsymbol{G}^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$. A natural choice to use the genotype dosage in the score test is to substitute the called genotype G by their genotype dosage G_D given the observed sequence data, i.e. $S_D = \boldsymbol{G}_D^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$. Variance of S_D is calculated under the logistic regression model assumption in Equation (4.1), given by $Var_N(S_D) = \sum_{i=1}^n \widetilde{G_{D_i}}^2 \hat{\mu}_i(1 - \hat{\mu}_i))$, where $\widetilde{\boldsymbol{G}_D} = \{\tilde{G}_{D_i}\} = \boldsymbol{G}_D - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{G}_D$ is the covariate-adjusted genotype vector, and $\boldsymbol{V} = diag\{\hat{\mu}_i(1 - \hat{\mu}_i)\}$. To apply the genotype dosage in the iECAT-Score test, we replace the called genotype with the genotype dosage to construct the three score statistics using internal samples exclusively $S_{D,int} = \boldsymbol{G}_{D,int}^T(\boldsymbol{Y}_{int} - \widehat{\boldsymbol{\mu}}_{int})$, the combined samples $S_{D,all} = \boldsymbol{G}_{D,all}^T(\boldsymbol{Y}_{all} - \widehat{\boldsymbol{\mu}}_{all})$, and the combined control samples $S_{D,IvE} = \boldsymbol{G}_{D,IvE}^T(\boldsymbol{Y}_{IvE} - \widetilde{\boldsymbol{\mu}}_{IvE})$. The iECAT-Score test statistic using genotype dosage is given by $S_{D,w} = a\tau_D S_{D,int} + (1 - \tau_D)S_{D,all}$ where $\tau_D = \frac{\tau_{D,1}}{1+\tau_{D,1}}$ with $\tau_{D,1} = \frac{S_{D,IvE}^2}{Var(S_{D,IvE})}$. The model-based variance estimates for $S_{D,int}$, $S_{D,all}$, and $S_{D,IvE}$ are $\widehat{Var}_N(S_{D,int})$, $\widehat{Var}_N(S_{D,all})$, and $\widehat{Var}_N(S_{D,IvE})$. Then a p value for association can be calculated following the iECAT-Score testing framework, with the iECAT-Score test statistic $S_{D,w}$ and its model-based variance estimate $\widehat{Var}_N(S_{D,w})$.

Derkach (Derkach et al., 2014) constructed a robust variance estimate for the score statistics by calculating the empirical variance of genotype dosages. As the genotype dosage is calculated as the expected genotype given the read data and the posterior probabilities of true genotypes, it is an empirical mean estimate of the true genotype G, which is unobserved. Hence, inspired by Derkach's method, we propose an alternative approach to estimate the variance of the generic score statistic $S_D = \boldsymbol{G}_D^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$ and estimate its empirical variance by $\widehat{Var}_E(S_D) = \sum_{i=1}^n \widehat{Var}(G_{D_i}) \times (Y_i - \hat{\mu}_i)^2$. This empirical variance estimate is similar to Derkach's method by treating the unobserved genotype as random. The variation of the empirical genotype dosage can be decomposed into the variation of the true genotype in the samples, and the additional variation introduced in the sequencing and genotype calling procedures due to sequencing errors, read depths, and other factors (Supplementary Materials **4.6.1.1**). After obtaining the empirical

variance estimate for the three score statistics $\widehat{Var}_E(S_{D,int})$, $\widehat{Var}_E(S_{D,all})$, and $\widehat{Var}_E(S_{D,IvE})$, we can calculate a p value for association using the iECAT-Score method based on the iECAT-Score test statistic $S_{D,w}$ and its empirical variance estimate $\widehat{Var}_E(S_{D,w})$.

4.3.2 Simulations

We carried out simulation studies under a range of scenarios to evaluate the performance of the proposed iECAT score test regarding type I error rates and power. Under each scenario, we implemented tests including the internal-sample-exclusive test, the combined-sample test, and the iECAT test, each using the simulated true genotypes, the called genotypes, and the genotype dosage with two types of the variance estimates $\widehat{Var}_N(S_{D,w})$, and $\widehat{Var}_E(S_{D,w})$.

For both type I error and power simulations, we first generated binary phenotypes of case/control from the true genotype from the logistic regression model:

$$logit[\Pr(Y = 1 \mid X, G)] = \alpha_0 + 0.5X_1 + 0.5X_2 + \beta G$$

where X_1 was a continuous covariate generated from a standard normal distribution, X_2 was a dichotomous covariate with the probability of 0.5 being 0 or 1, α_0 was chosen such that the disease prevalence was 0.05, G is the genotype at the variant of interest generated from a binomial (2, MAF) distribution, and β is the effect size of the variant. For internal samples, MAF was sampled from the MAF distribution in the AMD data; for external control samples, MAF was the corresponding minor allele frequency of the same variant in the MGI data.

After generating the true genotypes for both internal and external study samples, we simulated their sequence reads data at each locus for each sample. For internal samples, we simulated read depth from a normal distribution with mean 70 and a standard deviation 60% of its mean; for internal samples, we simulated 95% of the read depths from a normal distribution with mean 25 and a standard deviation 60% of the mean, and the rest 5% of the read depths from a $N(6, 3.75)$. Such choices of the read depth distributions mimic the read depth distributions in we observed in the internal study of MIGen and external study of UKBiobank data. We also simulated alternative scenarios of read depths from a normal $N(40, 24)$ distribution for both internal and external samples. The base-calling error rate was randomly generated from $N(10^{-3}, 0.025)$ and left truncated at 0. Using the simulated sequence read data, we calculated the posterior genotype likelihood based on the simple Bayesian genotyper as described in

Section 4.2, and obtained called genotypes using *GQ* filters 0, 2, and 5. We also calculated the genotype dosage using the posterior genotype likelihood.

As the number of samples whose data are available could vary between variants in real data, we chose to simulate sample sizes separately for each variant. The internal cases: internal controls: external controls sample size ratio is approximately 1,000: 1,000: 5,000 or 10,000, with the exact sample sizes generated from the following distributions: $\mathrm{Uniform}(600, 1400)$; $\mathrm{Uniform}(600, 1400)$, and $\mathrm{Uniform}(4000, 5000)$ or $\mathrm{Uniform}(8000, 10000)$, consistent with common scenarios where large external controls are available. In type one error simulations, $\beta = 0$. We generated 5×10^6 datasets to evaluate type I error rates at 5×10^{-3} and 5×10^{-5} level. In power simulations, β was set to values from a grid of $\log(1.1), \log(1.15), \ldots, \log(1.5)$, representing the odds ratio (OR) of 1.1, 1.15, …, 1.5 for the causal variant. We generated 10^6 data sets in each setting of effect size and case-control ratio to evaluate empirical power at the significance level of 10^{-4}.

4.3.3 Real data analysis

We applied our proposed methods using genotype dosage to exome sequence data from the Myocardial Infarction Genetics Exome Sequencing Consortium (MIGen) (Myocardial Infarction Genetics Consortium, 2009) downloaded from dbGaP. We used sraToolkit (Leinonen, Sugawara, Shumway, International Nucleotide Sequence Database Collaboration, 2011) and GotCloud variant calling pipeline (Jun et al., 2015) with the HapMap (McCarthy et al., 2016) panel to call variants. The MIGen dataset consists of 1,216 cases and 1,033 controls of myocardial infarction. As external controls, we used 9,210 unrelated samples from the UK Biobank (Bycroft et al., 2018). We used ICD-9 codes to select individuals from UK Biobank who are free from myocardial infarction. We applied the Fruposa software (Zhang et al., 2020) on called genotypes with the 1000 Genomes reference (The 1000 Genomes Project Consortium, 2015) to obtain genetic principal components. For ancestry matching of internal and external samples, we matched samples whose first two principal components fall into the mutual major cluster based on the Euclidean distance.

We first carried out quality control steps remove loci of low quality. Within each locus, we removed samples whose read depth was less than five; then we removed loci which failed either of the following filters: (1) monomorphic in internal or external samples; (2) sample size

of non-missing entries is fewer than 100; (3) the MAFs in internal and external samples differ beyond five folds. We performed analyses on the variants that passed the quality control filters to compare the performance of our proposed iECAT-Score applied to genotype dosage and called genotype. We used a logistic regression model to test for the association between the disease status of myocardial infarction and genetic variants that are shared by MIGen and UK Biobank data sets, adjusting sex and first ten principal components. We compared the performance of iECAT-Score, iECAT-Score minP, and methods that exclusively use internal samples and that naively combine control samples without adjusting for batch effect. When using genotype dosage, we exploited the three strategies as described in Section 3.1 to obtain the variance estimate of the score statistics; when using called genotypes, *GQ* filters of 0 or 50 were applied to obtain genotype calls below which genotype was set to be missing. We also compared the performance of our proposed methods using genotype dosage with the method proposed by Derkach without covariate adjustment.

4.4 Results

4.4.1 Type I error and power simulations

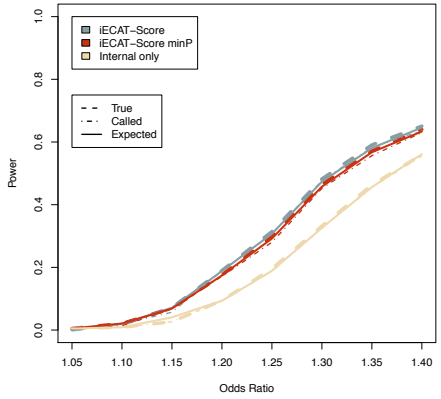
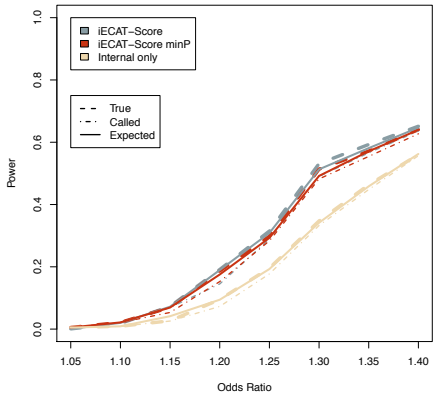
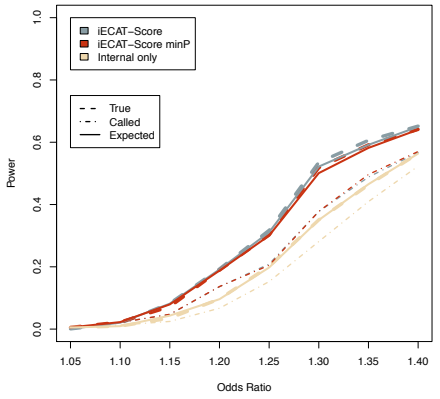
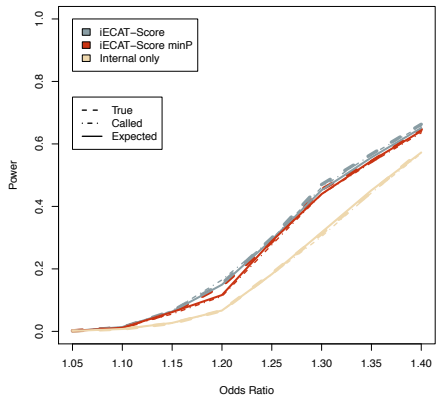
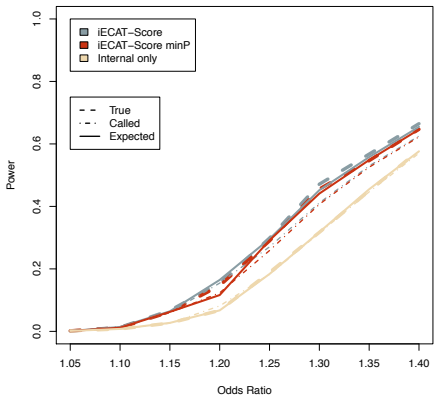
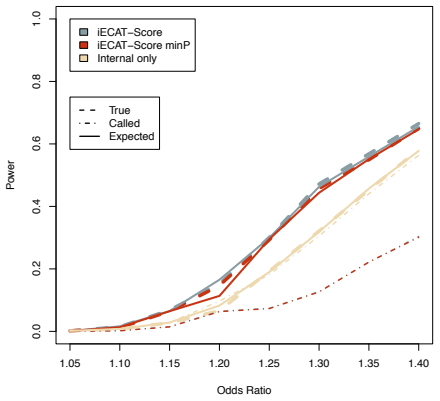
We present in **Table 4.1** the type I error rates of the tests that exclusively use internal samples, tests that naively combine control samples, and two versions of the iECAT-Score test at the significance level of 1e-04. For each test, we compare the type I error rates using the simulated true genotype, called genotypes with varying *GQ* filters, and genotype dosages with two variance estimate strategies. The results show that naively combining controls could result in type I error inflation secondary to batch effect, regardless of the type of genetic data used. The iECAT-Score methods using called genotypes have improved control of type I error rates; however, they still have risk some level of inflation, depending on the *GQ* filters applied during the genotype calling procedures. When applied to genotype dosages, the iECAT-Score methods with either variance estimating strategy demonstrate proper for type I error control in all settings.

Table 4.1: Type I error rates at 1e-04 level of comparing the following methods: method using exclusively internal samples, method that naively combines control samples, and various versions of the iECAT-Score method.
Shown under each method are type I errors using true genotypes, called genotypes with varying *GQ* filters, and expected genotypes using three variance estimate approaches. Values in the cells are ratios of their respective type I errors and the type I error using true genotype in the internal samples. The sample sizes of internal cases, internal controls, and external controls were 1000, 1000, 5000, respectively. A total of 5e06 datasets were generated to estimate the type I error rates at the significance level of 1e-04.

Mean Read Depth (internal cases: internal controls: external controls:	Type of Genotype	GQ Filter (called genotype only)	Internal	Naïvely combining controls	iECAT-Score	iECAT-Score minP	Internal vs. external
40: 40: 40	True	/	0.40	0.92	0.42	0.41	1.48
	Called	0	0.43	11.88	0.41	0.37	33.1
		20	0.42	12.72	0.62	0.52	2.13
		50	0.56	4.54	1.58	1.12	6.16
	Expected	$Var_N(S_E)$	0.46	12.08	0.56	0.47	1.90
		$Var_E(S_E)$	0.42	12.43	0.64	0.54	1.90
70: 70: 25	True	/	0.43	0.88	0.47	0.41	1.51
	Called	0	0.44	0.97	0.43	0.38	1.76
		20	0.42	10.85	0.93	0.74	36.7
		50	0.41	177.2	17.01	15.28	370.0
	Expected	$Var_N(S_D)$	0.42	1.05	0.51	0.45	1.86
		$Var_E(S_D)$	0.42	1.72	0.74	0.61	2.07

We compared the power of the iECAT-Score method, the iECAT-Score minP method, and the method that exclusively used internal samples, when applied to simulated true genotypes, called genotypes with *GQ* filters equal to 0, 2, 5, and genotype dosages at the empirical alpha level of 1e-04. **Figure 4.5** shows the power comparisons for two settings of internal versus external read depths $(d_i: d_e)$: (1) 40: 40 (top row); (2) 70: 25 (bottom row). In each scenario, the power of iECAT-Score minP test using genotype dosages reached powers almost as high as if the true genotypes were used. When called genotypes were used, the power was either similar to or lower than the case using genotype dosages, depending on relative read depths between internal and external samples and the *GQ* filters applied.

Figure 4.5: Empirical power comparisons between called genotypes and genotype dosages.
Empirical power comparison of the iECAT-Score method, the iECAT-Score minP method, and the method that exclusively used internal samples, when applied to simulated true genotypes, called genotypes with *GQ* filters equal to 0, 2, 5, and genotype dosages at the empirical alpha level of 1e-04. Shown are the power comparisons for two settings of internal versus external read depths $(d_i: d_e)$: (1) 40: 40 (top row); (2) 70: 25 (bottom row).

Read depth	R=0	R=2	R=5
40: 40			
70: 25			

4.4.2 Application to Myocardial Infarction Genetics Exome Sequencing (MIGen) Data

We analyzed the association between single variants and myocardial infarction from the Myocardial Infarction Genetics Exome Sequencing Consortium (MIGen), using samples from UK Biobank with exome sequencing data as external controls. We matched samples' genetic ancestry by matching samples whose first two principal components fall into the mutual major cluster with the Euclidean distance less or equal to 30 (**Figure 4.6**). After ancestry matching, the MIGen dataset consists of 1,145 cases and 1,025 controls of myocardial infarction; the UKBiobank consists of 8,606 controls. The female samples consist of 29.00% and 55.44% in internal cases and external controls, respectively; no female controls were present in the internal samples (**Table 4.2**).

Figure 4.6:First two genetic principal component scores of MIGen and UKBiobank study samples.
Shown in the gray circle are samples whose first two principal components fall into the mutual major cluster with the Euclidean distance less or equal to 30.

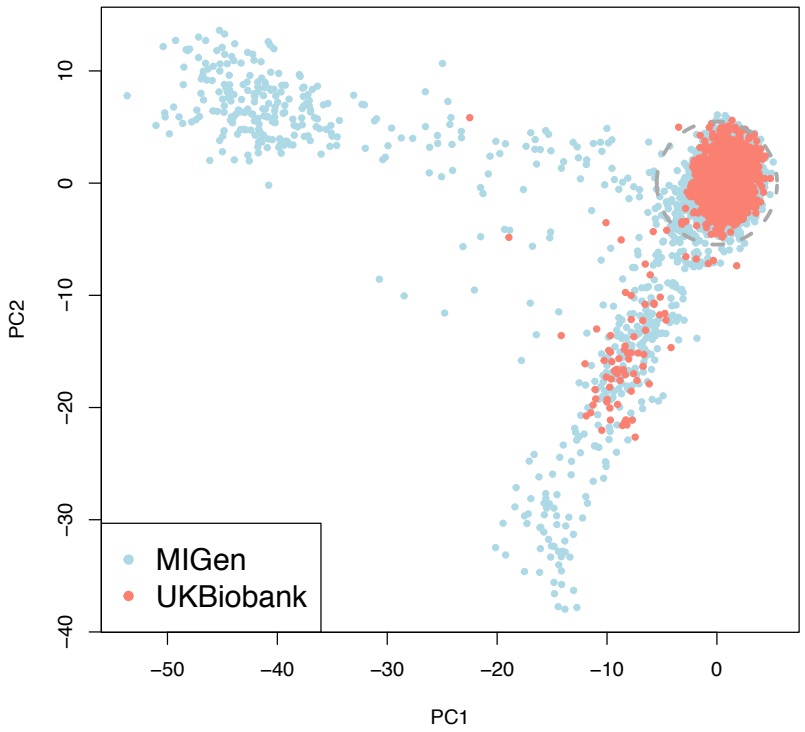


Table 4.2: Descriptive statistics of study subjects from internal (MIGen) and external (UKBiobank) studies.
Shown in the table are the sample sizes, the number (percentage) of female samples, and mean (standard deviation of sample age in years in MIGen and UKBiobank data.

Study	Sample Size N		Female N (%)			Total
	Cases	**Controls**	**Cases**	**Controls**	**Total**	
MIGen (internal)	1,145	1,025	332 (29.0)	0 (0)	332 (15.3)	2,170
UKB (external)		8,606		4,771 (55.4)	4,771 (55.4)	8,606
Total	1,145	9,631	332 (29.0)	4,771 (49.5)	5,103 (44.3)	10,776

A comparison of the variant-specific median read depths for chromosome 22 in the internal samples with those in the external samples show that the internal study samples were sequenced at higher read depths with median read depths 67 (**Figure 4.7**); the median read depths in the external samples are 23 with some variants being sequenced at very low depths (<10). Comparing the estimated minor allele frequencies for corresponding loci between the two studies, we observed that the two studies tend to give similar MAF estimates; however, the two studies could give differentiating estimates at a large number of loci (**Figure 4.7**).

After applying the quality control filters, 79,470 single variants were tested for association with myocardial infarction. **Figure 4.8** compares the *p* values from the single variant association analyses using the iECAT-Score minP testing methods with genotype dosages and called genotypes with varying *GQ* filters (R = 0, 5). The results show that iECAT-Score with genotype dosages control type I error rates. When more stringent *GQ* filters are applied to obtain called genotypes, there are increased level of type I error inflation. We note that the type I error inflation using called genotypes is marginal from these analyses, especially when no *GQ* filters are applied (R = 0). We also examined the performance of the iECAT-Score method using genotype dosage and using called genotypes in various subsets of the data, stratified by read depths, minor allele frequencies, and sample sizes. Specifically, we made the following observations from the QQ plots based on the stratified data: (1) when the variant-specific median read depths in the external samples are low (ranging from five to 20, **Figure 4.9**), methods using called genotypes could result in inflation in type I error rates; (2) when available external sample

size is small (fewer than 500 external control samples, **Figure 4.10**), the iECAT-Score tests using called genotypes tended to be conservative as compared to using genotype dosages; (3) among rare variants with minor allele frequencies less than 0.01 (**Figure 4.11**), the iECAT-Score test using genotype dosages could be more powerful than using called genotypes. We further compared our methods of iECAT-Score in data analysis with Derkach's method, which does not require internal controls to be available. As Derkach's method does not adjust for covariates, we applied the iECAT-Score method without including the covariates or population principal component scores for a fair comparison. Contrary to the iECAT-Score method which controlled for type I error, both Derkach's tests resulted in significant type I error inflation (**Figure 4.12**).

Figure 4.7: Distributions of median read depths and minor allele frequencies (MAFs) in internal and external samples.
Distributions of variant specific median read depths of Chromosome 22 in internal samples (left panel), external samples (center panel), and minor allele frequencies (right panel). In the right panel, shown in red color are variants whose internal and external sample minor allele frequencies are within five-fold difference.

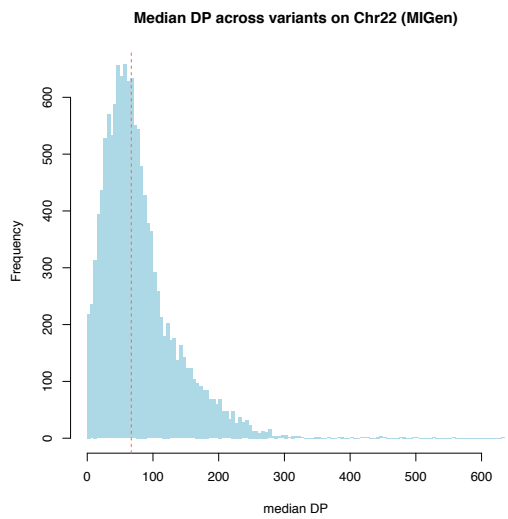
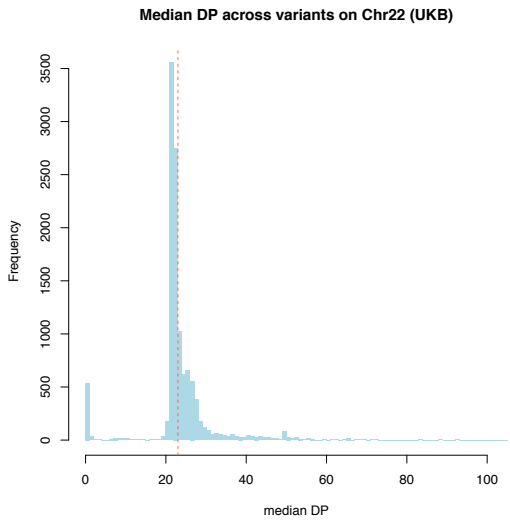
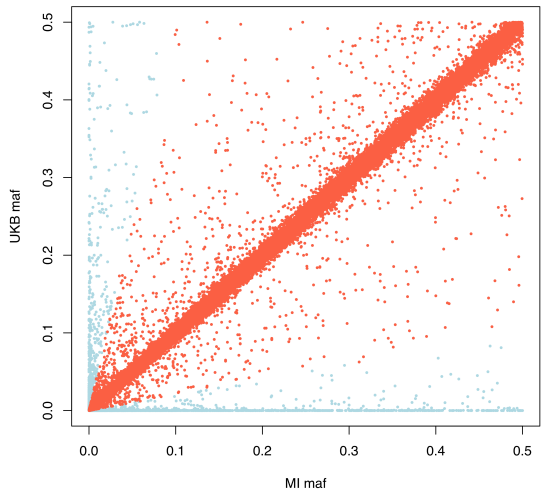
Internal (median 67)	UKB (median 23)	MAF comparison
		

Figure 4.8: QQ plots for *p* values from analysis of MIGen and UKBiobank data using the iECAT-Score minP method.
P values calculated using the iECAT-Score minP methods applied to genotype dosages (left panel) and called genotypes with varying *GQ* filters (R = 0, 5, middle and right panel).

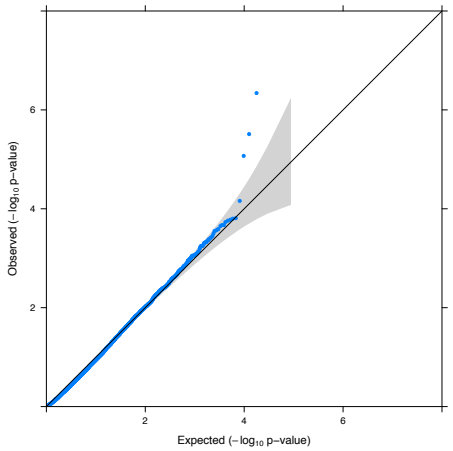
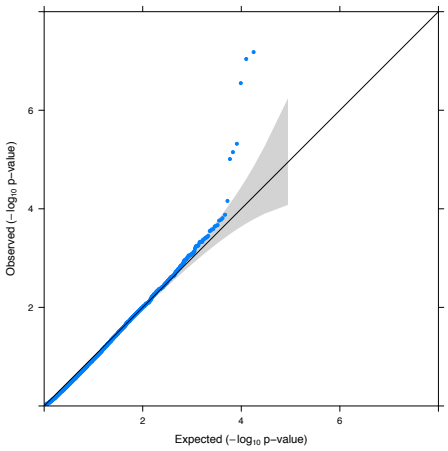
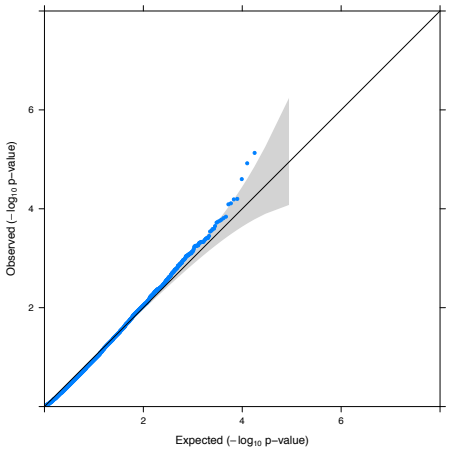
Genotype Dosage	Called Genotypes	
	R = 0	R = 5
		

Figure 4.9: QQ plots for *p* values from analysis of MIGen and UKBiobank data using the iECAT-Score minP method in variants of low median read depths.
P values are calculated applied to genotype dosages and called genotypes with varying *GQ* filters (R = 0, 5). Shown are variants whose variant-specific median read depths are fewer than 20 in UKBiobank samples.

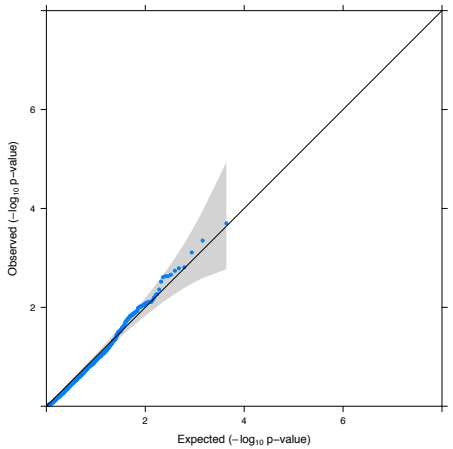
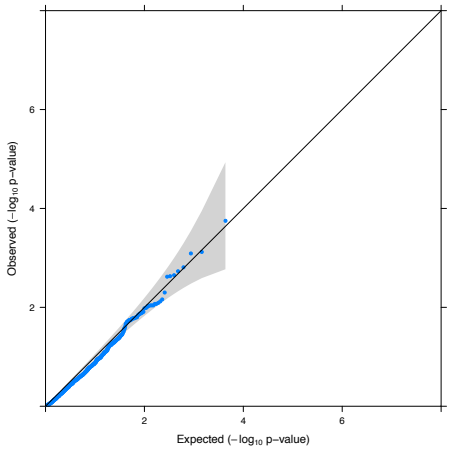
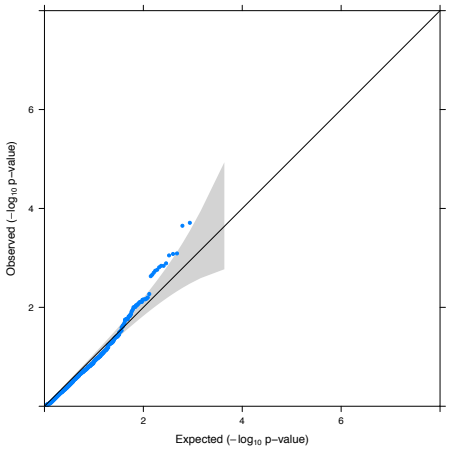
Genotype Dosage	Called Genotypes	
	R = 0	R = 5
		

Figure 4.10: QQ plots for *p* values from analysis of MIGen and UKBiobank data using the iECAT-Score minP method in variants of small sample sizes.
P values are calculated applied to genotype dosages and called genotypes with varying *GQ* filters (R = 0, 5). Shown are variants whose variant-specific available UKBiobank sample size is fewer than 500.

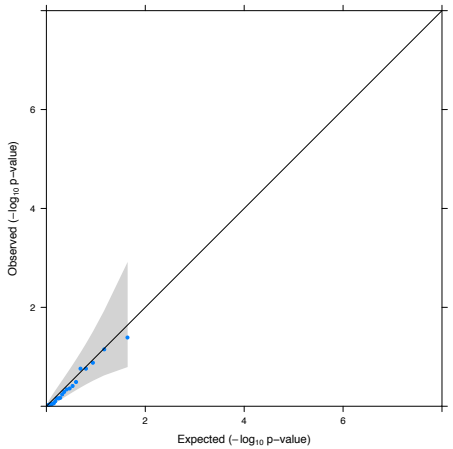
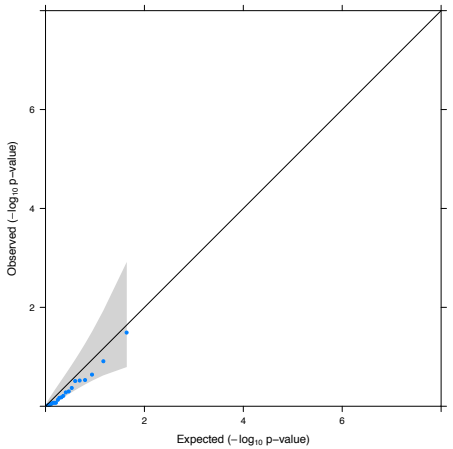
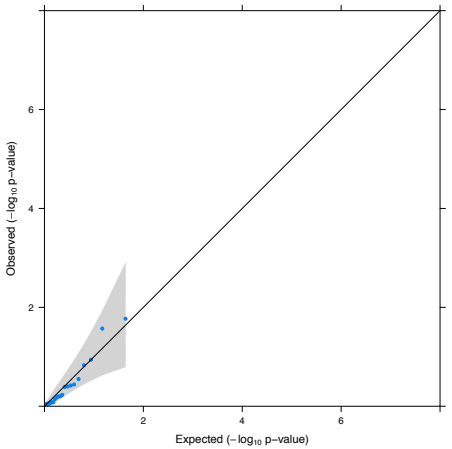
Genotype Dosage	Called Genotypes	
	R = 0	R = 5
		

Figure 4.11: QQ plots for *p* values from analysis of MIGen and UKBiobank data using the iECAT-Score minP method in rare variants.

P values are calculated applied to genotype dosages and called genotypes with varying *GQ* filters (R = 0, 5). Shown are variants whose minor allele frequencies in UKBiobank samples is less than 0.01.

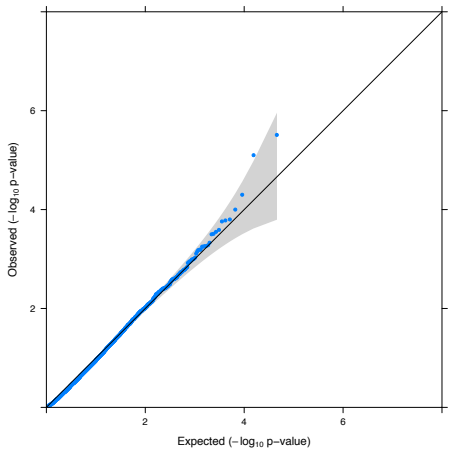
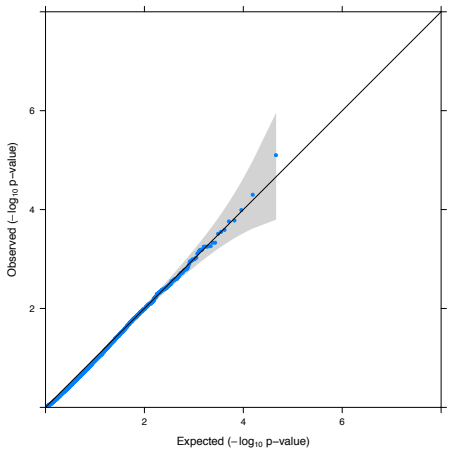
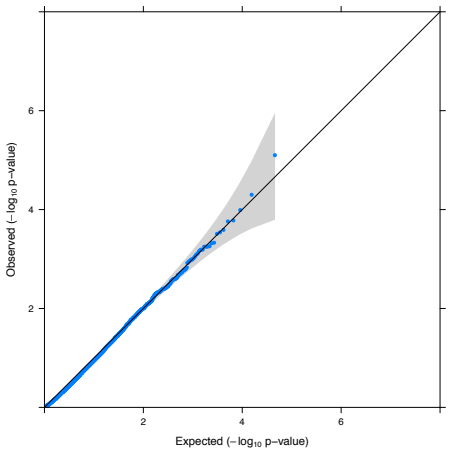
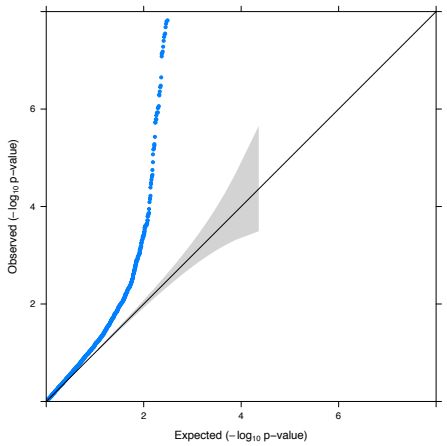
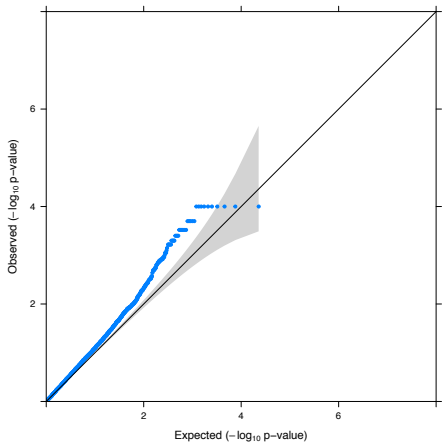
Genotype Dosage	Called Genotypes	
	R = 0	R = 5
		

Figure 4.12: QQ plots for *p* values using Derkach's method.
P values are calculated from analysis of MIGen and UKBiobank data using Derkach's method without using internal controls.

Derkach's method without using internal controls

Asymptotic test	Permutation test
	

We present in **Table 4.3** the top four variants using the iECAT-Score method applied to genotype dosages, among which two variants within genes *RBPJ* and *MUC2* reached or were close to the genome-wide significance level of 5e-08. On the contrary, no SNPs reached the genome-wide significance level of 5e-08 using either the iECAT-Score with called genotypes or tests that exclusively used internal samples. We note that none of the presented top variants have comparable *p* values when internal samples were exclusively used or if called genotypes were used in the iECAT-Score tests. However, the genes linked to the top variants have been reported to be associated with myocardial infarction or cardiac functions, indicating their biologic plausibility. Specifically, *RBPJ* (recombination signal binding protein for immunoglobulin kappa J region) is a high-level modulator of cardiomyocytes in myocardial angiogenesis (He et al., 2018; Scimia et al., 2019); the expression of secretory mucins *MUC2* is associated with important clinicopathological characteristics in patients with cardiac myxoma (P.-H. Chu, Jung, Yeh, Lin, & Chu, 2005); the epigenetic dysregulation of clustered *PCDHs* has been observed in Williams-Beuren syndrome (WBS), which is characterized by cardiovascular abnormalities and other impairments (Gurda, Handschuh, Kotkowiak, & Jakubowski, 2015); *SNRPB2* (ribonucleoprotein B) is a splicing-related gene that may have a global activating or inhibiting effect on genes whose expression is related to left ventricular hypertrophy (LVH), which is associated with hypertension and is a cardiovascular risk factor (Cerutti et al., 2006).

4.5 Discussion

Using publicly available sequenced data as external controls is a cost-effective approach to increase statistical power in case-control studies. However, using called genotypes generated from sequence data could result in inflation in type I error rates in methods that integrates external controls, including our own iECAT-Score method. In this chapter, we explored the effects of various factors in the sequencing and genotype calling process that could result in biased minor allele frequency estimation. Based on our findings, we proposed to replace the called genotypes with genotype dosages in the iECAT-Score methods to better adjust for the batch effect between internal and external samples. Compared to the iECAT-Score methods that use called genotypes, the iECAT-Score coupled with genotype controls for type I error rates improves power to detect rare variant associations.

Table 4.3: Top four variants from analysis of association with myocardial infarction based on iECAT-Score minP method using genotype dosages.

Shown are allele frequencies estimated using called and expected genotypes, minor allele frequencies in European population of corresponding variants in the dbSNP database, median read depths in internal and external samples, *p* values of analyses from exclusive usage of internal samples with genotype dosage and using the iECAT-Score method with genotype dosage and called genotype.

Signal Number	Index variant				Minor allele frequency (called genotype, genotype dosage)					dbSNP EUR MAF	Median Read Depth (internal, external)	*p* values		
	Name	dbSNP ID	Chr:Pos	Major / minor allele	Internal		External	Combined				iECAT-Score minP (genotype dosage)	Internal (genotype dosage)	iECAT-Score minP (called genotype)
					Case	Control	Control	Control	All					
1	*RBPJ*	rs3113014	4:26406256	T/C	8.22e-04 8.22e-04	3.87e-03 3.87e-03	2.99e-03 5.05e-03	3.08e-03 4.92e-03	2.84e-03 4.49e-03	7.00e-03	61 12	2.27e-16	1.26e-01	1.00e-01
2	*MUC2*	rs9735156	11:1099733	T/C	1.64e-03 2.09e-03	3.87e-03 4.90e-03	9.23e-04 4.50e-03	1.24e-03 4.54e-03	1.28e-03 4.27e-03	1.45e-03	92 17	5.08e-08	1.72e-01	1.76e-01
3	*PCDHB7*	rs116101007	5:141174295	T/C	1.52e-02 1.52e-02	5.81e-03 5.81e-03	5.05e-03 5.65e-03	5.13e-03 5.66e-03	6.20e-03 6.68e-03	4.45e-03	221 123	1.06e-07	8.38e-03	8.69e-04
4	*SNRPB2*	rs141440350	20:16732309	A/G	1.64e-03 1.64e-03	4.85e-03 4.85e-03	1.64e-03 3.75e-03	1.90e-03 3.84e-03	1.87e-03 3.61e-03	2.35e-03	50 12	1.78e-06	1.26e-01	1.93e-01

The simulation studies showed that iECAT-Score methods using genotype dosages better control for type I error rates and have improved empirical power in the case of low read depths compared to using called genotypes. Analysis of the myocardial infarction from the Myocardial Infarction Genetics Exome Sequencing Consortium and UK Biobank data revealed that iECAT-Score can improve power for association discovery for variants with low minor allele frequencies.

The quantile-quantile plots from real data analysis showed increased level of inflation when more stringent filters were applied to the genotype quality scores to obtain called genotypes; however, such inflation was less profound as compared to what we observed from the simulation studies. In simulation studies, we aimed to examine the effect of sample sizes, read depths, and genotype quality filters on the performance of our iECAT-Score method. Although we attempted to mimic the distributions of various parameters in the simulation studies as in the real data, the real data tend to contain a mixture of scenarios that are challenging to fully imitate via simulations. We examined the performance of the iECAT-Score method using genotype dosage and using called genotypes in various subsets of the data, stratified by read depths, minor allele frequencies, and sample sizes (**Figures 4.9**, **Figure 4.10**, **Figure 4.11**). The stratified QQ plots revealed that the trend of inflation or deflation, which might be present in certain strata of the data, became less obvious when the strata are collapsed.

In our first project (Y. Li & Lee, 2021) where we proposed the single-variant iECAT-Score test using called genotypes, we noticed via simulation studies, that when the number of internal or external control samples is limited, our method could deliver conservative or anti-conservative performance. As the iECAT-Score method assesses the existence of batch effect via a control vs. control comparison, it would not be unusual that having abundant control samples, or the lack thereof, could affect the performance of the iECAT-Score tests. Our stratified QQ plots showed that, on the contrary, using genotype dosages could rescue the deflation of the iECAT-Score test among variants where the external control sample size was limited. Hence, genotype dosages could offer more reliable assessment of the batch effect between internal and external controls and potentially improve the power for association test when there are fewer control samples available.

The top hit variants from the iECAT-Score test using genotype dosages indicated a few association signals with myocardial infarction. These variants have been discussed in multiple references to be involved in important pathways of the cardiac related diseases; however, they did not show significance if the traditional called genotypes were used, or if the internal samples were exclusively used. One common feature of these top significance variants is that their minor alleles are present in low frequencies in the population (< 1%). In fact, the stratified QQ plots focusing on the rare variants (MAFs < 1%) also revealed that the iECAT-Score test using genotype dosage could offer improved power in discovering association signals when the minor allele frequency is low in the population. The results of additional simulation studies focusing on rare variants (Supplementary **Figure S4.1**) were consistent with our observations in real data.

Both the internal study of Myocardial Infarction Genetics Exome Sequencing Consortium and the external study of the UK Biobank consist of British samples. Surprisingly, when we compared the minor allele frequencies at same variants between the two study samples, we observed a decent number of variants whose minor allele frequencies differ significantly between the two sets of data (**Figure 4.7**, right panel). While the internal samples are sequenced at higher read depth overall, we do not believe that the differences in their estimated minor allele frequencies were a mere result of differences in the read depths. Using the MAFs in the European samples in the dbSNP database as reference, either internal or external samples could have MAF closer to the reference MAF (**Table 4.3**). Such observation indicates that the differential MAFs between different study samples are intrinsic properties of specific populations and thus are a result of underlying biology (e.g., unaccounted for genetic drift in the sampled populations) in addition to the technical factors (sequencing platforms, genotype calling pipelines etc.). Hence, it is necessary to include internal control samples from the same population of interest (i.e., the internal controls), as they serve as reference samples to be compared with the external control samples, informing us important biological differentiations as well as technical batch effect when integrating external study samples. Based on our simulations and data analyses from the previous chapters, a minimum of several hundred internal controls provides a reasonable starting point as reference.

We compared our methods of iECAT-Score in data analysis with Derkach's method that does not require internal controls to be available. Derkach constructed a robust variance estimate for the score statistics calculated from genotype dosages to avoid false discoveries, and yet we

still observed significant type I error inflation (**Figure 4.12**). Such inflation is likely the result of not having internal controls samples as reference and instead relying on the average read depths in internal cases and external controls to decide on whether the robust variance estimate is deployed. However, as we discovered in our data, the differences in MAFs between two study samples could be attributed to both disparate read depths and intrinsic biological differences. Therefore, we believe that it is crucial to rely on the internal controls versus external controls comparison to examine the level of batch effect before integrating external controls, to better avoid false discoveries secondary to batch effect.

In summary, we extended the iECAT framework to be applicable to both whole genome/exome sequence data. As the sequencing cost continues to drop and large-scale biobanks become available, publicly available sequence data provide valuable resources to augment control sample size to assist association discoveries in case-control studies. When analyzing sequence data of moderate read depth, we recommend applying genotype dosages instead of using called genotypes with quality control filters. Using genotype dosages offer consistent MAF estimation, controls type I error rate, and improves power for association discovery especially in rare variants; opting to use genotype dosages also preserves more variants available for association testing. Our proposed method is extremely accessible, as it does not require SRA or BAM files to be available, instead only requiring posterior genotype likelihood, which is often provided in the VCF files through Phred-scaled genotype likelihood. Through the incorporation of the strategy to use genotype dosages, we develop a complete framework of integrating external controls that is applicable to both genotyped and sequencing data, further honing the statistical methods needed to identify disease-causing variants within the human genome.

4.6 Supplementary Materials

4.6.1 Validation of Theoretical Results

4.6.1.1 Empirical variance $\widehat{Var}_E(S_{D,w})$

The genotype dosage $\phi(D_i) = \sum_{g=0,1,2} g \times \widehat{Pr}(g|D_i)$ is an *estimator* of the unobserved true genotype G_i. We assume that $E(\phi(D_i)|G_i) = G_i$; marginally we assume $E(G_i) = \mu_i = \mu$ and $Var(G_i) = \sigma_i^2 = \sigma^2$ for all i.

Let $S = \boldsymbol{G}^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$ be a generic score test statistic. Replacing the true genotype $\boldsymbol{G}$ with genotype dosage $\phi(\boldsymbol{D})$, the score of interest is $S_D = \boldsymbol{G}_D^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) = \sum_i E(G_i|D_i)(Y_i - \hat{\mu}_i) = \sum_i \phi(D_i)(Y_i - \hat{\mu}_i)$. To estimate the variance of S_D, we want to estimate the variance of the estimator $\phi(D_i)$, i.e., $\widehat{Var}\big(\phi(D_i)\big)$.

Consider

$$\begin{aligned} Var\big(\phi(D_i)\big) &= Var(E[\phi(D_i)|G_i]) + E\big(Var(\phi(D_i)|G_i)\big) \\ &= Var(G_i) + E\big(Var(\phi(D_i)|G_i)\big) \end{aligned} \tag{4.1}$$

where $Var(E[\phi(D_i)|G_i]) = Var(G_i) = \sigma^2$. An estimator of variance of $\phi(D_i)$ would be

$$\widehat{Var}\big(\phi(D_i)\big) = \widehat{\sigma^2} + \widehat{Var}\big((\phi(D_i)|G_i)\big) \tag{4.2}$$

Some algebra:

$$E\big(\phi(D_i)\big) = E(E(\phi(D_i)|G_i) = E(G_i) = \mu$$

$$\begin{aligned} E\big(\phi^2(D_i)\big) &= E\big[E\big(\phi^2(D_i|G_i)\big)\big] = E\{Var(\phi(D_i)|G_i) + [E(\phi(D_i)|G_i)]^2\} \\ &= E\{Var(\phi(D_i)|G_i) + G_i^2\} = E\{Var(\phi(D_i)|G_i)\} + \mu^2 + \sigma^2 \end{aligned}$$

$$E\big(\overline{\phi(D)}^2\big) = E\left[\left(\frac{1}{n}\sum_i \phi(D_i)\right)^2\right] = \frac{1}{n^2}\left\{E\left[\sum_i \phi^2(D_i) + \sum_{i \neq i'} \phi(D_i)\phi(D_{i'})\right]\right\}$$

$$= \frac{1}{n^2}\left\{\sum_i (E[Var(\phi(D_i)|G_i)] + \mu^2 + \sigma^2) + n(n-1)\mu^2\right\}$$

$$= \frac{1}{n^2}\left\{\sum_i E[Var(\phi(D_i)|G_i)] + n^2\mu^2 + n\sigma^2\right\}$$

$$E\left[\sum_i \left(\phi(D_i) - \overline{\phi(D)}\right)^2\right] = E\left[\sum_i \phi^2(D_i) - n\overline{\phi(D)}^2\right] = \sum_i E\left(\phi^2(D_i)\right) - n \times E\left(\overline{\phi(D)}^2\right)$$

$$= \left\{\sum_i \left\{E\left[Var\left(\phi(D_i)|G_i\right)\right] + \mu^2 + \sigma^2\right\}\right\}$$

$$-\frac{1}{n}\left\{\sum_i E[Var(\phi(D_i)|G_i)] + n^2\mu^2 + n\sigma^2\right\}$$

$$= \frac{n-1}{n}\sum_i E[Var(\phi(D_i)|G_i)] + (n-1)\sigma^2 \Longrightarrow E\left\{\frac{1}{n-1}\sum_i \left(\phi(D_i) - \overline{\phi(D)}\right)^2\right\}$$

$$= \frac{1}{n}\sum_i E[Var(\phi(D_i)|G_i)] + \sigma^2$$

Hence, an unbiased estimator of σ^2 is given by

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_i \left(\phi(D_i) - \overline{\phi(D)}\right)^2 - \frac{1}{n}\sum_i \widehat{Var}(\phi(D_i)|G_i) \tag{4.3}$$

Substituting $\widehat{\sigma^2}$ into Equation (4.3), we have

$$\widehat{Var}\left(\phi(D_i)\right) = \frac{1}{n-1}\sum_i \left(\phi(D_i) - \overline{\phi(D)}\right)^2 - \frac{1}{n}\sum_i \widehat{Var}(\phi(D_i)|G_i) + \widehat{Var}\left((\phi(D_i)|G_i)\right) \tag{4.4}$$

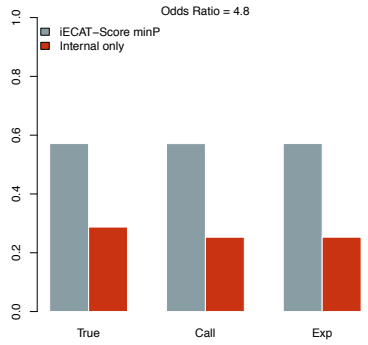
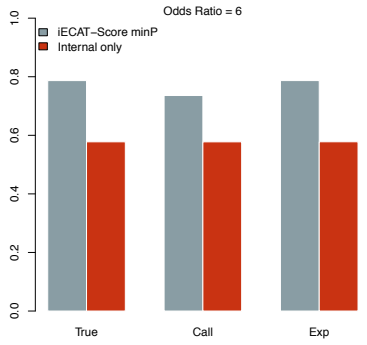
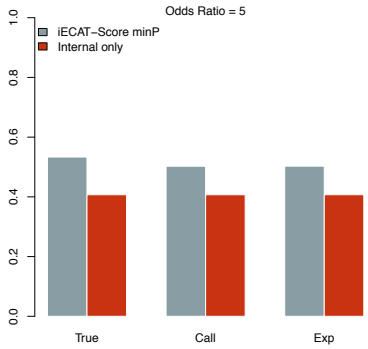
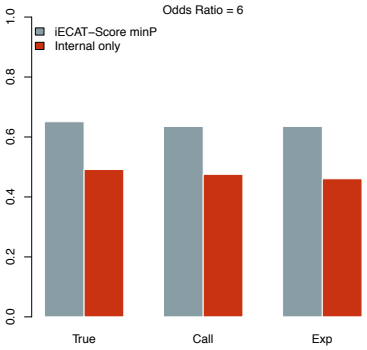
where $\widehat{Var}(\phi(D_i)|G_i)$ can be estimated using the posterior likelihood, as $\phi(D_i)|G_i$ is a multinomial variable with probabilities approximated by the posterior genotype probabilities.

Finally, the empirical variance estimator for $S_D = \boldsymbol{G}_D^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) = \sum_i \phi(D_i)(Y_i - \hat{\mu}_i)$ is given by

$$Var(S_D) = \sum_i \widehat{Var}\left(\phi(D_i)\right) \times (Y - \hat{\mu})^2$$

4.6.2 Supplementary Tables and Figures

Figure S4.1: Power comparison for rare causal variants (MAF < 0.01) of the iECAT-Score minP method and the method that exclusively used internal samples.
Shown are power comparisons when the methods are applied to simulated true genotypes, called genotypes with GQ filters equal to 0, and genotype dosages at the alpha level of 1e-04. Read depths indicates the simulated mean read depths of variants in internal and external samples, respectively.

	Odds Ratio	
Read depth	5	6
40: 40		
70: 25		

Chapter 5 Conclusion

In this dissertation, we proposed single variant and region-based score tests that allow for the integration of external controls for improved power in association tests. In Chapter 2, we developed a single-variant score test, iECAT-Score, based on the insight of the original iECAT test, that allows for covariate adjustment and constructs a shrinkage score statistic that is a weighted sum of the score statistics using exclusively internal samples and uses both internal and external control samples. Extending the iECAT-Score test, we constructed a region-based test in Chapter 3 to achieve improve power for rare-variant association test. In Chapter 4, we investigated the effect of multiple parameters during the sequencing and genotype calling process on the estimated minor allele frequencies; we suggested replacing called genotypes with genotype dosages when applying iECAT-Score tests to sequence data to reduce bias in minor allele frequency estimation and obtain improved power for detecting rare variants associations. In all three chapters, we showed via simulation studies and data analyses that our proposed methods control for type I error rates and have improved power to detect disease associated variants and genes.

The iECAT-Score tests are powerful tools for investigators to take advantage of the publicly available consortium genotype and sequence data to assist the association studies for the phenotype and population of interest. We should note that, our methods offer an alternative strategy to combine separate studies to achieve higher power as compared to meta-analysis. Meta-analysis uses summary statistics to combine results of individual studies to increase statistical power and precision in estimating genetic effects. There exist methods of meta-analysis to assess inter-study heterogeneity and identify sources of heterogeneity. Contrary to meta-analysis, the iECAT-Score methods focus on improving the power for the study of interest through integrating existing resources as additional controls. Meta-analysis combines effect sizes or *p* values from independent studies, each consisting of cases and controls, whereas the iECAT

methods combines individual level data using the internal/primary case and control samples, and external controls. Hence, meta-analysis and iECAT provide different approaches to achieve improved power to detect disease-susceptible genes under distinct study settings. A direct comparison of the power between the two methods is outside the scope of this dissertation.

Our methods illuminate several major challenges that arise in the effort to combine study samples as large-scale consortia and biobanks become widely available to the public. Heterogeneity between studies including technical batch effect and population stratification could cause spurious discoveries if left unaccounted for. Our methods rely on the comparison between internal and external controls to assess the level of heterogeneity between the two studies and decides on a weight of external samples to be included. Such comparison detects heterogeneity such as differential distributions of allele frequencies, covariates, technical batch effect, and population stratification. Thus, the performance the iECAT testing suite depends on the quality and sample size of both internal and external data. Several other recently developed methods allow for exploiting external controls without requiring internal controls to be available. These methods use sequence data to adjust for the technical batch effect, but are not able to adjust for covariates such as age, gender. Thus, new methods that are able to adjust for covariates and heterogeneity of various sources without requirement on the control sample size (both internal and external) would be beneficial, as some studies primarily focus on genotyping/sequencing disease cases.

The current set up of the iECAT framework involves the internal study of cases and controls and one source of external controls. One area of future work involves extending our methods to allow for multiple sources of external controls. As iECAT uses a logistic regression model to assess the level of batch effect between the two sets of controls, it would be important to extend the method to account for the batch effect between internal samples and external samples, as well as the heterogeneity among the external samples. One possible strategy could be using propensity score matching to select subsets of external control samples and apply the iECAT testing methods. An alternative proposal would be to pool the external control samples, calculate propensity scores using internal and external controls, form bucketing case and control groups based on propensity scores, apply the iECAT methods within each bucket, and aggregate the results from the disjoint buckets with appropriate weights. Another extension that would worth pursuing involves taking account of related samples. As the sample size of large biobanks

continue to grow, it would not be unusual that related individuals exist in the same study. Thus, it would be beneficial for our methods to address relatedness when using such data. Mixed model association tests such as GMMAT (H. Chen et al., 2016), SMMAT (H. Chen et al., 2019), SAIGE (Zhou et al., 2018), and SAIGE-GENE (Zhou et al., 2020) provide helpful starting point for this extension.

In summary, in this dissertation we developed a powerful testing suite for integrating external controls into genetic association tests that are applicable to both genotyped and sequencing data. Our iECAT methods address some important challenges that arise as researchers leverage the existing genetic consortia and biobank resources that are widely available today. Through further research and validation, we hope that more methodological improvements will help us expand the toolkit for genetic association studies and the discovery of disease susceptible genes within the human genome.

Bibliography

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. http://doi.org/10.1038/nature11632

Ancient human genomes suggest three ancestral populations for present-day Europeans. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409–413. http://doi.org/10.1038/nature13673

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 1–25. http://doi.org/10.1038/s41586-018-0579-z

Cerutti, C., Kurdi, M., Bricca, G., Hodroj, W., Paultre, C., Randon, J., & Gustin, M.-P. (2006). Transcriptional alterations in the left ventricle of three hypertensive rat models. *Physiol Genomics*, *27*(3), 295–308. http://doi.org/10.1152/physiolgenomics.00318.2005

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., et al. (2019). AR TICLE Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *The American Journal of Human Genetics*, *104*(2), 260–274. http://doi.org/10.1016/j.ajhg.2018.12.012

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., et al. (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *The American Journal of Human Genetics*, *98*(4), 653–666. http://doi.org/10.1016/j.ajhg.2016.02.012

Chen, S., & Lin, X. (2018). Analysis in Case–Control Sequencing Association Studies with Different Sequencing Depths. *Biostatistics*, 1–17. http://doi.org/10.1093/biostatistics/kxy073

Chu, P.-H., Jung, S.-M., Yeh, T.-S., Lin, H.-C., & Chu, J.-J. (2005). MUC1, MUC2 and MUC5AC expressions in cardiac myxoma. *Virchows Archiv*, *446*(1), 52–55. http://doi.org/10.1007/s00428-004-1147-5

Conneely, K. N., & Boehnke, M. (2007). So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests. *The American Journal of Human Genetics*, *81*(6), 1158–1168. http://doi.org/10.1086/522036

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. http://doi.org/10.1038/ng.806

Derkach, A., Chiang, T., Gong, J., Addis, L., Dobbins, S., Tomlinson, I., et al. (2014). Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics*, *30*(15), 2179–2188. http://doi.org/doi:10.1093/bioinformatics/btu19

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *The American Journal of Human Genetics*, *101*(1), 37–49. http://doi.org/10.1016/j.ajhg.2017.05.014

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, *11*(6), 446–450. http://doi.org/10.1038/nrg2809

Fagerness, J. A., Maller, J. B., Neale, B. M., Reynolds, R. C., Daly, M. J., & Seddon, J. M. (2008). Variation near complement factor I is associated with risk of advanced AMD. *European Journal of Human Genetics*, *17*(1), 100–104. http://doi.org/10.1038/ejhg.2008.140

Freedman, M. L., Monteiro, A. N. A., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics*, *43*(6), 513–518. http://doi.org/10.1038/ng.840

Fritsche, L. G., Chen, W., Schu, M., Yaspan, B. L., Yu, Y., Thorleifsson, G., et al. (2013). Seven new loci associated with age-related macular degeneration. *Nature Genetics*, *45*(4), 433–9–439e1–2. http://doi.org/10.1038/ng.2578

Fritsche, L. G., Igl, W., Bailey, J. N. C., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., et al. (2015). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, *48*(2), 134–143. http://doi.org/10.1038/ng.3448

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, *102*(5), 717–730. http://doi.org/10.1016/j.ajhg.2018.04.002

Gold, B., Merriam, J. E., Zernant, J., Hancox, L. S., Taiber, A. J., Gehrs, K., et al. (2006). Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nature Genetics*, *38*(4), 458–462. http://doi.org/10.1038/ng1750

Guan, W., Liang, L., Boehnke, M., & Abecasis, G. R. (2009). Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genetic Epidemiology*, *33*(6), 508–517. http://doi.org/10.1002/gepi.20403

Gurda, D., Handschuh, L., Kotkowiak, W., & Jakubowski, H. (2015). Homocysteine thiolactone and N-homocysteinylated protein induce pro-atherogenic changes in gene expression in human vascular endothelial cells. *Amino Acids*, 1–21. http://doi.org/10.1007/s00726-015-1956-7

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., et al. (2018). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, *47*(D1), D853–D858. http://doi.org/10.1093/nar/gky1095

He, Y., Pang, S., Huang, J., Zhu, K., Tong, J., Tang, Y., et al. (2018). Research Article Blockade of RBP-J-Mediated Notch Signaling Pathway Exacerbates Cardiac Remodeling after Infarction by Increasing Apoptosis in Mice. *BioMed Research International*, 1–8. http://doi.org/10.1155/2018/5207031

Helgason, H., Sulem, P., Duvvari, M. R., Luo, H., Thorleifsson, G., Stefansson, H., et al. (2013). A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. *Nature Genetics*, *45*(11), 1371–1374. http://doi.org/10.1038/ng.2740

Hendricks, A. E., Billups, S. C., Pike, H. N. C., Farooqi, I. S., Zeggini, E., Santorico, S. A., et al. (2018). ProxECAT: Proxy External Controls Association Test. A new case-control gene

region association test using allele frequencies from public controls. *PLOS Genetics*, *14*(10), e1007591–14. http://doi.org/10.1371/journal.pgen.1007591

Hu, Y.-J., Liao, P., Johnston, H. R., Allen, A. S., & Satten, G. A. (2016). Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. *PLOS Genetics*, *12*(5), e1006040–19. http://doi.org/10.1371/journal.pgen.1006040

Ioannidis, J. P. A., Thomas, G., & Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, *10*(5), 318–329. http://doi.org/10.1038/nrg2544

Jiao, X., Doamekpor, S. K., Bird, J. G., Nickels, B. E., Tong, L., Hart, R. P., & Kiledjian, M. (2017). 5′ End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell*, *168*(6), 1015–1027.e10. http://doi.org/10.1016/j.cell.2017.02.019

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, *25*(6), 918–925. http://doi.org/10.1101/gr.176552.114

Keenan, T. D. L., Toso, M., Pappas, C., Nichols, L., Bishop, P. N., & Hageman, G. S. (2015). Assessment of Proteins Associated With Complement Activation and Inflammation in Maculae of Human Donors Homozygous Risk at Chromosome 1 CFH-to- F13B. *Investigative Opthalmology & Visual Science*, *56*(8), 4870–10. http://doi.org/10.1167/iovs.15-17009

Lee, S., Fuchsberger, C., Kim, S., & Scott, L. (2015). An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. *Biostatistics*, *17*(1), 1–15. http://doi.org/10.1093/biostatistics/kxv033

Lee, S., Kim, S., & Fuchsberger, C. (2017). Improving power for rare-variant tests by integrating external controls. *Genetic Epidemiology*, *41*(7), 610–619. http://doi.org/10.1002/gepi.22057

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, *13*(4), 762–775. http://doi.org/10.1093/biostatistics/kxs014

Leinonen, R., Sugawara, H., Shumway, M., International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19–21. http://doi.org/10.1093/nar/gkq1019

Li, B., & Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics*, *83*(3), 311–321. http://doi.org/10.1016/j.ajhg.2008.06.024

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Li, Y., & Lee, S. (2021). Novel score test to increase power in association test by integrating external controls. *Genetic Epidemiology*, *2021*(45), 293–304. http://doi.org/10.1002/gepi.22370

Maller, J. B., Fagerness, J. A., Reynolds, R. C., Neale, B. M., Daly, M. J., & Seddon, J. M. (2007). Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nature Genetics*, *39*(10), 1200–1201. http://doi.org/10.1038/ng2131

Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M. J., & Seddon, J. M. (2006). Common variation in three genes, including a noncoding variant in CFH, strongly

influences risk of age-related macular degeneration. *Nature Genetics*, *38*(9), 1055–1059. http://doi.org/10.1038/ng1873
Maróti, Z., Boldogkői, Z., Tombácz, D., Snyder, M., & Kalmár, T. (2018). Evaluation of whole exome sequencing as an alternative to BeadChip and whole genome sequencing in human population genetic analysis, 1–13. http://doi.org/10.1186/s12864-018-5168-x
McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. http://doi.org/10.1038/ng.3643
Myocardial Infarction Genetics Consortium. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics*, *41*(3), 334–341. http://doi.org/10.1038/ng.327
Narendra, U., Pauer, G., & Hagstrom, S. (2009). Genetic analysis of complement factor H related 5. *Molecular Vision*, *15*, 731–736.
Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*, *7*(7), e37558–11. http://doi.org/10.1371/journal.pone.0037558
Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451. http://doi.org/10.1038/nrg2986
Picard-Jean, F., Brand, C., Tremblay-Létourneau, M., Allaire, A., Beaudoin, M. C., Boudreault, S., et al. (2018). 2'-O-methylation of the mRNA cap protects RNAs from decapping and degradation by DXO. *PLoS ONE*, *13*(3), e0193804–14. http://doi.org/10.1371/journal.pone.0193804
Ratnapriya, R., Sosina, O. A., Starostik, M. R., Kwicklis, M., Kapphahn, R. J., Fritsche, L. G., et al. (2019). Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nature Genetics*, 1–13. http://doi.org/10.1038/s41588-019-0351-9
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*, 1576–1583.
Scimia, M. C., Bushway, P., Tran, D., Monosov, A., Monosov, E., Peterson, K., et al. (2019). Notch-independent RBPJ controls angiogenesis in the adult heart. *Nature Communications*, 1–10. http://doi.org/10.1038/ncomms12088
Seddon, J. M., Reynolds, R., Yu, Y., & Rosner, B. (2014). Three New Genetic Loci (R1210C in CFH, Variants in COL8A1 and RAD51B) Are Independently Related to Progression to Advanced Macular Degeneration. *PLoS ONE*, *9*(1), e87047–11. http://doi.org/10.1371/journal.pone.0087047
Seddon, J. M., Yu, Y., Miller, E. C., Reynolds, R., Tan, P. L., Gowrisankar, S., et al. (2013). Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. *Nature Genetics*, *45*(11), 1366–1370. http://doi.org/10.1038/ng.2741
Sun, C., Zhao, M., & Li, X. (2012). CFB/C2Gene Polymorphisms and Risk of Age-Related Macular Degeneration: A Systematic Review and Meta-Analysis. *Current Eye Research*, *37*(4), 259–271. http://doi.org/10.3109/02713683.2011.635401
Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *2021*(590), 290–299. http://doi.org/10.1038/s41586-021-03205-y

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, *337*(6090), 64–69. http://doi.org/10.1126/science.1219240

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. http://doi.org/10.1038/nature15393

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., et al. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. *Nature Genetics*, *46*(4), 409–415. http://doi.org/10.1038/ng.2924

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. http://doi.org/10.1093/nar/gkq603

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. http://doi.org/10.1038/nature05911

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). AR TICLE Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics*, *89*(1), 82–93. http://doi.org/10.1016/j.ajhg.2011.05.029

Yates, J. R. W. (2007). Complement C3 Variant and the Risk of Age-Related Macular Degeneration. *The New England Journal of Medicine*, 1–9. http://doi.org/10.1056/NEJMoa072618)

Zhan, X., Larson, D. E., Wang, C., Koboldt, D. C., Sergeev, Y. V., Fulton, R. S., et al. (2013). Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nature Genetics*, *45*(11), 1375–1379. http://doi.org/10.1038/ng.2758

Zhang, D., Dey, R., & Lee, S. (2020). Fast and robust ancestry prediction using principal component analysis. *Bioinformatics*, *36*(11), 3439–3446. http://doi.org/10.1093/bioinformatics/btaa152

Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L. G., & Lee, S. (2020). UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. *The American Journal of Human Genetics*, *106*(1), 3–12. http://doi.org/10.1016/j.ajhg.2019.11.012

Zheng, J., Li, Y., Abecasis, G. R., & Scheet, P. (2011). A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic Epidemiology*, *35*(2), 102–110. http://doi.org/10.1002/gepi.20552

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 1–12. http://doi.org/10.1038/s41588-018-0184-y

Zhou, W., Zhao, Z., Nielsen, J. B., Fritsche, L. G., LeFaive, J., Taliun, S. A. G., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature Genetics*, 1–20. http://doi.org/10.1038/s41588-020-0621-6