Biostatistics (2016), **17**, 1, *pp.* 1–15
doi:10.1093/biostatistics/kxv033
Advance Access publication on September 11, 2015

An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies

SEUNGGEUN LEE*, CHRISTIAN FUCHSBERGER

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

leeshawn@umich.edu

SEHEE KIM

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

LAURA SCOTT

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Summary

For aggregation tests of genes or regions, the set of included variants often have small total minor allele counts (MACs), and this is particularly true when the most deleterious sets of variants are considered. When MAC is low, commonly used asymptotic tests are not well calibrated for binary phenotypes and can have conservative or anti-conservative results and potential power loss. Empirical p-values obtained via resampling methods are computationally costly for highly significant p-values and the results can be conservative due to the discrete nature of resampling tests. Based on the observation that only the individuals containing minor alleles contribute to the score statistics, we develop an efficient resampling method for single and multiple variant score-based tests that can adjust for covariates. Our method can improve computational efficiency $>$1000-fold over conventional resampling for low MAC variant sets. We ameliorate the conservativeness of results through the use of mid-p-values. Using the estimated minimum achievable p-value for each test, we calibrate QQ plots and provide an effective number of tests. In analysis of a case–control study with deep exome sequence, we demonstrate that our methods are both well calibrated and also reduce computation time significantly compared with resampling methods.

Keywords: Rare variants; Next generation sequencing; Resampling methods.

1. Introduction

Recent advances in sequencing technologies have made it possible to investigate the role of rare variants in complex diseases, and numerous statistical methods have been developed to identify rare

*To whom correspondence should be addressed.

variant associations. Many of the currently popular gene- or region-based multiple variants tests are based on individual variant score statistics, which provide rapid computation and natural adjustment for covariates (Lee *and others*, 2014). For example, variance component tests use the weighted sum of squared individual variant score statistics as in C-alpha (Neale *and others*, 2011), SSU (Pan, 2009), and SKAT (Wu *and others*, 2011). Many versions of burden tests (Li and Leal, 2008; Lin and Tang, 2011; Madsen and Browning, 2009) are essentially equivalent to collapsing the individual variant score statistics. Other examples include SKAT-O (Lee, Emond, *and others*, 2012; Lee, Wu, *and others*, 2012) and Fisher method (Derkach *and others*, 2012; Sun *and others*, 2013).

For a given gene or region, the number of variants tested together and the total of their minor allele counts (MACs) can vary due to the sequence or genotyping coverage of the gene, the class of variants tested, and the sample size. In the context of gene-based tests, we use MAC to refer to the total MAC of all variants in a tested set (i.e. the sum of the MAC of the rare, low, and common frequency variants in the set) and in the context of single variant tests, MAC refers to the single variant MAC.

In exome-sequencing studies, one approach (among many) is to test disruptive or predictively damaging variants (Zuk *and others*, 2014). These tend to be very rare and tests based on these variants often have sets of variants with very small total MACs ($\text{MAC} \leqslant 40$). Asymptotic-based score tests for a single variant with small MAC, however, yield conservative results under a balanced case–control design, and anti-conservative results under an unbalanced case–control design (Ma *and others*, 2013). This lack of calibration can lead to lack of calibration in gene- or region-based asymptotic score tests. A moment-based adjustment (MA) was developed to improve the Type I error control when testing variant sets with low MAC; however, this approach is also based on the asymptotic properties of the tests (Lee, Emond, *and others*, 2012; Lee, Wu, *and others*, 2012) and may be less well calibrated, when testing for very low MAC variant sets. An alternative approach would be to perform experiment-wise permutation to control family wise error rate by obtaining the empirical distribution of asymptotic p-values across variant sets (Kiezun *and others*, 2012). However, because the degree of miscalibration for asymptotic p-values can vary by MAC, this approach may have reduced power to detect specific classes of causal multiple variant sets.

Resampling methods, such as permutation tests, do not rely on the asymptotic properties of the test (Efron and Tibshirani, 1994). Permutation tests for genetic data often permute case and control status without regard to differential odds of individual being a case based on covariates. This approach can result in the inflated Type I error rates in the presence of confounding covariates, such as population stratification (Epstein *and others*, 2012). In a more nuanced approach, permutations can be performed within strata of one or more covariates, such as geographical region, so the underlying null distribution provides a better match to the observed test statistic (Purcell *and others*, 2007). In the presence of continuous covariates, such as principal components which are used to adjust for population stratification, Fisher's noncentral hypergeometric distribution-based permutation can be performed, allowing for individuals to have different odds of being selected as a case (Efron and Tibshirani, 1994; Fog, 2008). The major limitation of the permutation approach is that disease status is permuted across all study participants, requiring significant computational cost, which increases as sample sizes become larger. Adaptive permutation procedures can reduce computational time for the estimation of large or moderate p-values (Efron and Tibshirani, 1994), but substantial time is still required to estimate highly significant p-values. In addition, permutation p-values tend to be conservative for binary traits with small MAC, since test statistics are discrete (Lancaster, 1961).

In this paper, we develop an efficient resampling (ER) method for score statistic-based single and multiple variant tests that improves computational efficiency. Our method is based on the insight that only individuals with minor alleles (assuming the minor allele is coded as one) contribute to the score test. Instead of permuting case–control status across all individuals, resampling can be performed by resampling the case–control status of individuals with a minor allele at a given variant (for a single variant test), and similarly, individuals with minor alleles at any included variants (for a multiple variant test). Within the group of individuals with minor alleles, we allow for covariate adjustment through the use of Fisher's non-central hyper-geometric distribution (Epstein *and others*, 2012; Fog, 2008). The computational time for the ER method increases as the MAC increases, so we developed a method for moderate to high variant set MAC (MAC $>$ 40) in which quantiles of the test statistics are estimated through ER (based on a more limited number of permutations) and then used to better-calibrate our moment-matching approximation quantile adjustment (QA).

Furthermore, we develop statistical approaches to calibrate the discrete nature of test statistics. Using the ER method, we obtain mid-p-values (Lancaster, 1961). We estimate the lower limit of p-values for each variant set (minimum achievable p-values (MAP)) (Kiezun *and others*, 2012), using the exact resampling distribution. We use the MAP to estimate the effective number of tests and to calibrate quantile–quantile (QQ) plots. Through simulation-based work and analysis of deep exome sequencing data, we demonstrate that the ER-based methods and calibration approaches are computationally efficient, control the false-positive rate (FPR) and can improve power.

2. Methods

2.1 *Statistical model and rare variant tests*

To understand currently used rare variant tests, suppose that n subjects are sequenced with n_{case} diseased individuals. The region being tested has p variant loci. For the ith subject, let y_i denote a binary phenotype, $\boldsymbol{G}_i = (g_{i1}, \ldots, g_{ip})'$ the number of copies of the minor allele ($g_{ij} = 0, 1, 2$), and $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{iq})'$ the covariates. MAC is defined as the sum of all genotype values, MAC $= \sum_{i=1}^{n} \sum_{j=1}^{p} g_{ij}$. To relate genotypes to binary phenotypes, we posit the logistic regression model, $\text{logit}(\pi_i) = \alpha_0 + \boldsymbol{X}_{\boldsymbol{i}}'\alpha + \boldsymbol{G}_{\boldsymbol{i}}'\beta$, where π_i is a disease probability, α_0 is the intercept, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are regression coefficients of covariates and genetic variants, respectively. A score statistic from a marginal model of variant j is

$$S_j = \sum_{i=1}^{n} (y_i - \hat{\pi}_i) g_{ij} \tag{2.1}$$

where $\hat{\pi}_i$ is an estimate of π_i under the null model $\text{H}_0 : \boldsymbol{\beta} = \boldsymbol{0}$. For single variant tests, S_j^2 is the score test statistic of variant j and follows a (scaled) χ^2 distribution with $df = 1$. Many popular gene- or region-based tests are also based on S_j. For example, Burden and SKAT test statistics can be written as a weighted linear and quadratic sum of S_j

$$Q_{\text{Burden}} = \left(\sum_{j=1}^{p} w_j S_j \right)^2; \quad Q_{\text{SKAT}} = \sum_{j=1}^{p} (w_j S_j)^2$$

where w_j is a weight for variant j. SKAT-O combines Burden test and SKAT using the following framework as $Q_\rho = (1 - \rho) Q_{\text{SKAT}} + \rho Q_{\text{Burden}}$. Since the optimal ρ is not known in prior, SKAT-O uses the minimum p-values over a grid of ρ as a test statistic.

2.2 *ER method*

In this section, we present the ER method for rare variant score tests with binary traits. We describe the generation of B resamples to estimate the following four probabilities of the gene or region-based association test statistic Q, which is a function of S_j ($j = 1, \ldots, p$), given genotypes (G), phenotypes (Y) and covariates (X):

(1) ER p-value: $P_{\text{ER}} = Pr(Q \geqslant \hat{Q}|Y, G, X)$
(2) ER mid-p-value: $P_{\text{ER-mid}} = Pr(Q \geqslant \hat{Q}|Y, G, X) - 0.5Pr(Q = \hat{Q}|Y, G, X)$
(3) ER minimum achievable p-value: $\text{MAP}_{\text{ER}} = Pr(Q = \hat{Q}_{\max}|Y, G, X)$
(4) ER minimum achievable mid-p-value: $\text{MAP}_{\text{ER-mid}} = Pr(Q = \hat{Q}_{\max}|Y, G, X)/2$

where $\hat{Q}$ is a test statistic from the original phenotype, and $Q_{\max}$ is the maximum of all possible permutation test statistics. Let m ($<n$) be the number of individuals with minor alleles in the gene or region, $m = \sum_{i=1}^{n} I(\sum_{j=1}^{p} g_{ij} > 0)$, where $I(\cdot)$ is an indicator function. It is apparent that m is smaller than or equal to MAC. From Equation (2.1), only individuals with minor alleles contribute to S_j, since the remaining individuals have zero genotype values for all of their loci. This observation allows us to reduce the computation time by restricting resampling to the case–control status of those m individuals only, rather than using all n individuals. To estimate ER p-values, we use a two-step approach that is based on the fact that p-value can be factorized as

$$Pr(Q \geqslant \hat{Q}|Y, G, X) = \sum_{d=0}^{m} Pr(Q \geqslant \hat{Q}|D = d, Y, G, X)Pr(D = d|Y, G, X)$$

where D is the number of cases among m individuals carrying a minor allele in the tested region.

Step 1 is to estimate $Pr(D = d|Y, G, X)$. If there are no covariates to adjust for, D follows the central-hypergeometric distribution. When there are covariates to adjust for, we use Fisher's noncentral hypergeometric distribution, which allows each individual to have different odds of being a case (Fog, 2008). Since estimating $Pr(D = d|Y, G, X)$ while allowing all individuals to have different odds is computationally challenging, we propose to stratify the m individuals into groups based on $\hat{\pi}_i$ and to assume an average common odds for all individuals within the same stratum. The $n - m$ individuals without variants are treated as a single group (Supplementary Appendix A). The only use of this stratification is to estimate $Pr(D = d|Y, G, X)$ for the m individuals in Step 1. We used 10 strata for the m individuals, for a total of 11 strata.

In Step 2, we estimate $Pr(Q > \hat{Q}|D = d, Y, G, X)$ by generating $B_d = B \times Pr(D = d|Y, G, X)$ permutations of the case–control status of m individuals. Suppose $S_{jd}^{(b)}$ is the bth resample of S_j given $D = d$, and $Q_d^{(b)}$ is the resulting test statistic Q. Examples of $Q_d^{(b)}$ include the resampled Burden and SKAT test statistics, $Q_{\text{Burden},d}^{(b)} = (\sum_{j=1}^{p} w_j S_{jd}^{(b)})^2$ and $Q_{\text{SKAT},d}^{(b)} = \sum_{j=1}^{p}(w_j S_{jd}^{(b)})^2$. The probability for the bth resample given $D = d$, say P_{db}, is also calculated using Fisher's noncentral hypergeometric distribution at the level of each individual in m (rather than the level of strata as in Step 1). Then the estimator of $Pr(Q \geqslant \hat{Q}|D = d, Y, G, X)$ is $\sum_{b=1}^{B_d} I(Q_d^{(b)} \geqslant \hat{Q})P_{db}$ and the ER p-value is

$$P_{\text{ER}} = \sum_{d=0}^{m} \sum_{b=1}^{B_d} I(Q_d^{(b)} \geqslant \hat{Q})P_{db}Pr(D = d|Y, G, X).$$

The estimator of ER-mid p-value is

$$P_{\text{ER-mid}} = P_{\text{ER}} - 0.5 \sum_{d=0}^{m} \sum_{b=1}^{B_d} I(Q_d^{(b)} = \hat{Q})P_{db}Pr(D = d|Y, G, X),$$

where the second term is an estimator of the tie probability. Suppose $Q_{\max}$ is the maximum of over all b and d (i.e. $Q_{\max} = \max Q_d^{(b)}$). Then, estimators of MAP_{ER} and $\text{MAP}_{\text{ER-mid}}$ are

$$\text{MAP}_{\text{ER}} = \sum_{d=0}^{m} \sum_{b=1}^{B_d} I(Q_d^{(b)} = Q_{\max}) P_{db} Pr(D = d | Y, G, X); \quad \text{MAP}_{\text{ER-mid}} = \text{MAP}_{\text{ER}}/2.$$

The detailed derivations of Steps 1 and 2 are given in Supplementary Appendix A.

The computational complexity of the proposed method is $O(Bmp)$ for SKAT and SKAT-O, and $O(Bm)$ for single variant and Burden tests, respectively. The computation complexity can be further reduced if the total number of configurations of case–control status (C_T) is small. For example, the total number of configurations of case–control status is 1024 when $m = 10$, indicating that we only need to evaluate 1024 possible configurations to obtain the exact resampling distribution. We note that we estimate MAPs when the exact resampling distribution is obtained (i.e. $B = C_T$); otherwise, the MAP estimates are not accurate. Since the computational cost of ER increases as m increases, it may not be practical to use ER for variant sets with moderate or large MAC. We develop ER-based QA moment matching (Supplementary Appendix B) for these variant sets, which produces more accurate p-values than the moment matching adjustment and yet provides fast computation for moderate or large MAC variant sets.

Because Bonferroni correction and QQ plots assume that p-values have a uniform distribution, they cannot correctly account for the fact that resampling p-values have lower limits, i.e., the MAPs. Kiezun *and others* (2012) proposed a heuristic approach in which to first identify variant sets with $\text{MAP} < 0.001$, and to count only these variant sets as the effective number of tests. We developed an alternative statistical approach to estimate the effective number of test and calibrating QQ plots using MAP (Supplementary Appendix C).

2.3 *Numerical simulations*

We generated 10 000 sequence haplotypes for an ~250 kbps region using a coalescent simulator FTEC (Reppell *and others*, 2012) with a faster-than-exponential growth model. In order to make variant sets having wide-ranges of MAC, we randomly selected a regions ranging from 125 to 12 500 bps, and then generated genotypes of variant sets using the simulated haplotypes. Three different case–control ratios were considered (1000:1000, 500:1500, and 500:1500). The binary phenotypes were generated from the logistic regression model:

$$\text{logit} P(Y = 1) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + G'_{\text{causal}} \beta_{\text{causal}} \tag{2.2}$$

where G_{causal} is a genotype vector containing causal variants, β_{causal} is a vector of genetic effect coefficients, X_1 was a binary covariate of Bernoulli (0.5), and X_2 was a continuous covariate of $N(0, 1)$. The intercept α_0 was chosen for the disease prevalence of 0.05. The non-genetic covariate coefficients α_1 and α_2 were 0 without covariates and 0.5 with covariates.

We applied five different methods to compute p-values for each of the Burden, SKAT and SKAT-O tests: (i) ER with a p-value (ER); (ii) ER with a mid-p-value (ER-mid); (iii) QA moment matching; (iv) moment matching adjustment (MA); and (v) unadjusted (UA) asymptotic tests. To verify that ER and the whole-sample permutation methods produce essentially identical p-values, we generated 20 000 variants sets and compared the p-values from ER and the permutation methods with and without covariates by generating 10^5 resamples (Supplementary Appendix E). We also compared computation times of SKAT-ER with whole-sample permutation for $m = 40$ and total sample sizes ranging from 100 to 50 00 0 (Supplementary Appendix E).

To compare the FPR for different ranges of total MAC, we considered six total MAC bins: MAC $\leqslant 10$; $10 < \text{MAC} \leqslant 20$; $20 < \text{MAC} \leqslant 40$; $40 < \text{MAC} \leqslant 100$; $100 < \text{MAC} \leqslant 200$; and $200 < \text{MAC} \leqslant 500$. For each bin, we used ranges of the number of variant sets $K = 5$ to 20 000, corresponding to candidate gene studies to genome-wide studies. In addition to FPR simulations, we carried out simulations to evaluate the power of ER and other tests. Details of FPR and power simulations can be found in Supplementary Appendix E.

3. Results

3.1 *Numerical simulations*

We examine the FPR control, power, and computational time of two existing approaches, the MA and UA p-value, and three newly developed ER-based methods, ER with p-value (ER), ER with mid-p-value (ER-mid), and the ER-based quantile adjustment (QA) for single variant and multiple variant tests across a range of MAC and case–control imbalance. For simulation-based data, we generated sequence haplotypes with a European demographic model that mimics the MAF spectrum and linkage-disequilibrium (LD) structure of the current European population (Reppell *and others*, 2012). The MAF spectrum of simulated haplotypes was similar to that observed for the GoT2D exome sequencing data (Supplementary Figure S1).

3.1.1 *Comparison of p-values obtained using ER or whole-sample permutations.* We compared SKAT p-values for 20 000 variant sets with total MAC $\leqslant 40$ using the ER method to those obtained from whole-sample-based permutation, either in the absence of covariates (permutation of case–control status) or in the presence of covariates (using Fishers noncentral hypergeometric distribution). The $-\log 10$ p-values were very highly correlated ($r > 0.99$) for tests with and without covariates, indicating that the ER-based results mirror those obtained from whole-sample-based permutation methods (Figure 1). We observed equally concordant p-values for Burden and SKAT-O tests (data not shown).

3.1.2 *Comparison of computational times for the estimation of a significant gene-based p-value.* To compare the computation times for a significant gene-based p-value (0.05/20 000 genes), we generated 10^7 resamples for each method for a single variant set. This allows us to estimate a $p\text{-value} = 2.5 \times 10^{-6}$ with a standard error $\sim$0.2 of 2.5×10^{-6}. When 40 individuals have minor alleles (MAC equal or slightly higher than 40), SKAT-ER with no covariates ran in $\sim$10 s and the computation times were invariant to sample size (100–50 000 samples). In contrast, for SKAT whole-sample permutations (SKAT-Perm), the computation time increased linearly with total sample size, from 0.35 to 10 h for 2000 and 50 000 samples, respectively (Figure 2(a)). With covariates, SKAT-ER also ran in $\sim$10 s and was invariant to sample size, whereas SKAT Fisher's noncentral hypergeometric distribution-based whole-sample permutations (SKAT-FNHPerm) using the BiasedUrn R-package took $>$10 h for 2000 samples (Figure 2(b)). The running times for SKAT-ER-mid were nearly identical to those for SKAT-ER (data not shown). In existing programs, 10^7 resamples of 2000 (50 000) samples with no covariates took 6 min (3.6 h) for C-alpha in PLINK/SEQ (and substantially longer for SKAT), and with covariates, took 6.4 h ($>$240 h) in SCORE-Seq using the offered set of 5 gene-based tests (Supplementary Table S1).

In contrast to the invariance by sample size, the computation time for ER increased with increasing number of individuals with minor alleles. For a single test with covariates, when the number of individuals with minor alleles $m = 10$, 40, 100, and 500, SKAT-ER took 0.01, 10, 58, and 310 s; the burden test was faster and SKAT-O slower (Figures 2(c) and (d); Supplementary Table S2). When $m \leqslant 20$, computation took substantially less time because the total number of configurations of cases and controls among those

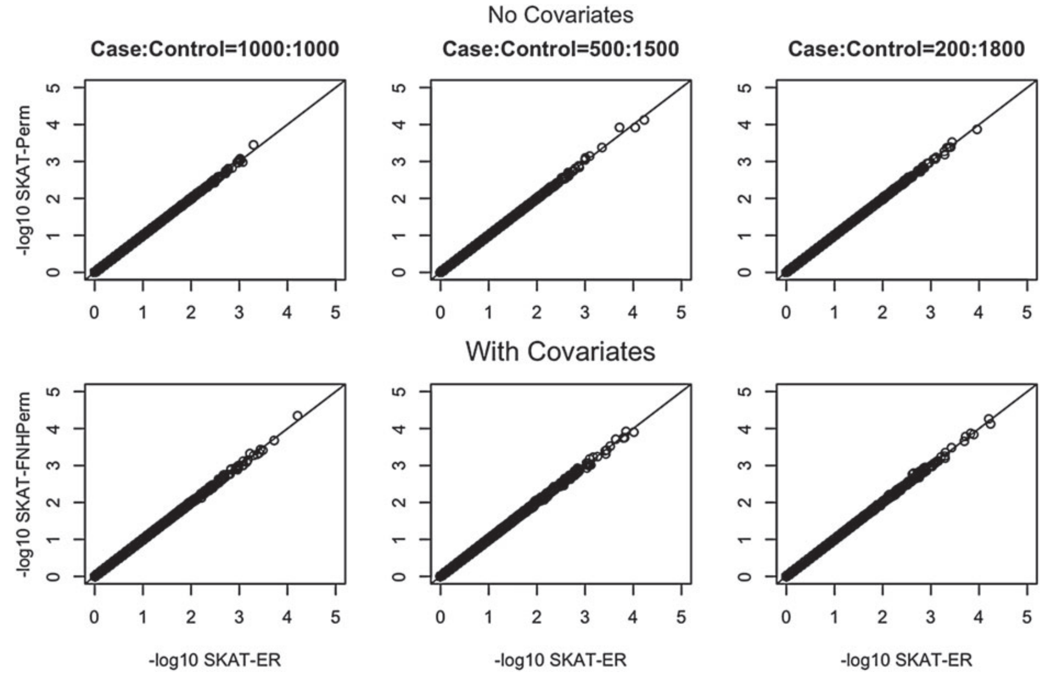

Fig. 1. Comparison of SKAT p-values obtained using ER or whole-sample permutations. In the absence of covariates, SKAT p-values were obtained through ER or whole-sample permutation (Perm) of disease status (top panel). In the presence of covariates, SKAT p-value were obtained through ER or Fisher's noncentral hypergeometric distribution based whole-sample permutation (FNHPerm) implemented in the BiasedUrn R-package (bottom panel). From left to the right, the plots consider case:control = 1000:1000, 500:1500, and 200:1800, respectively. The x-axis represents $-\log_{10}$ SKAT-ER p-values and y-axis represents $-\log_{10}$ SKAT-Perm or SKAT-FNHPerm p-values. Variant sets were randomly simulated, 20 000 sets with MAC $\leqslant 40$ selected, and 10^5 resamples were generated to compute p-values for each method.

m individuals was $<10^7$. The increase in computation time with increasing m led us to develop a substantially faster ($\sim$6- to 18-fold) QA asymptotic method based on ER (QA) (Figure 2(d) and Supplementary Table S2). QA was essentially linear in m and invariant to sample size (data not shown). For comparison, with covariates for $m = 40$ and sample size of 2000, the existing MA method for Burden, SKAT and SKAT-O took <0.2 s (and was invariant to m), and UA for Burden, SKAT and SKAT-O took <0.02 s (and was invariant to m) (data not shown).

3.1.3 *FPRs for existing and ER-based methods.* We compared empirical FPRs for variant sets for these five methods. We define the best-calibrated test as the one that had the FPR closest to but, at most, slightly exceeding the expected FPR at the Bonferroni corrected level α. Figure 3 shows the FPRs for SKAT in the presence of covariates using Bonferroni corrected $\alpha = 0.05$ for 5–20 000 sets of variants and MAC $\leqslant 40$. Over the MAC and case–control imbalance scenarios, ER-mid had the best-calibrated FPRs, though it was conservative when MAC $\leqslant 10$ for balanced case–control studies. ER was slightly more conservative than ER-mid when MAC $\leqslant 10$, but otherwise behaved similarly. QA was designed to speed the computation for moderate or large MAC. For MAC between 10 and 40 QA was conservative for balanced studies, and slightly anti-conservative for imbalanced studies. MA had conservative or anti-conservative FPRs

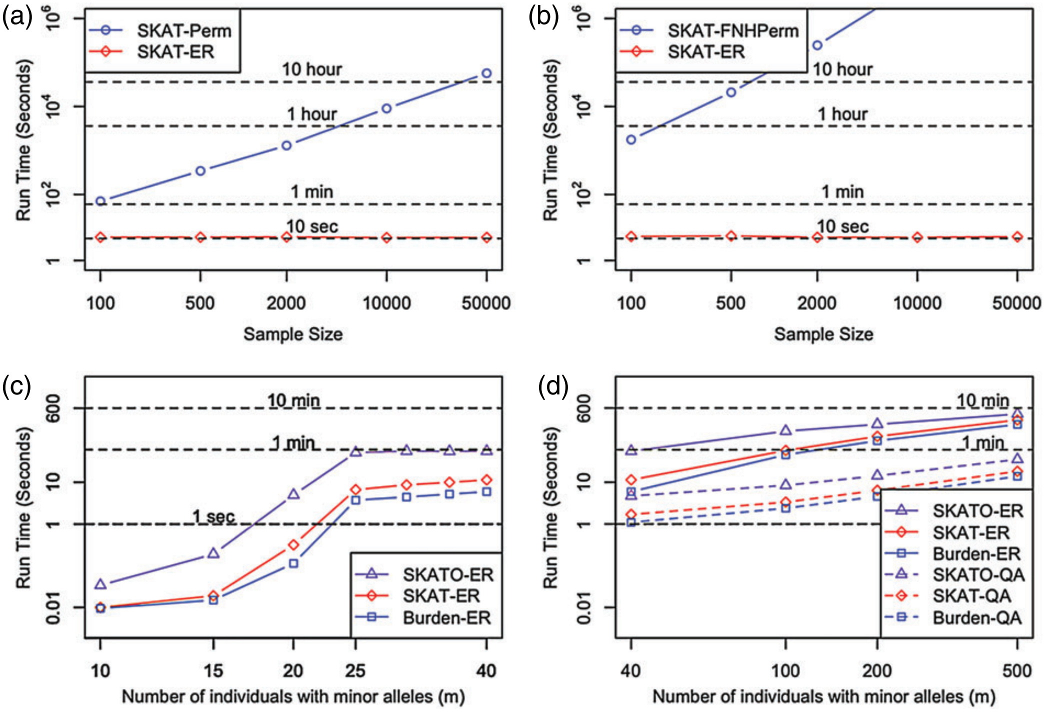

Fig. 2. Comparison of computation times for the estimation of a significant gene-based p-value using ER and existing methods. Estimated computation time for 10^7 resamples of a single variant set for 40 individuals with minor alleles ($m = 40$) and varying numbers of total samples (balanced case:control) using SKAT-ER or SKAT-Perm in the absence of covariates (a) or using SKAT-ER or SKAT-FNHPerm in the presence of covariates (b). The BiasedUrn R-package was used for SKAT-FNHPerm. Estimated computation time for 10^7 resamples of a single variant set for 2000 samples (balanced case:control) in the presence of covariates for SKAT-O, SKAT, or Burden test for $10 \leqslant m \leqslant 40$ individuals with minor alleles using ER (c) or for $40 \leqslant m \leqslant 500$ individuals with minor alleles using ER and QA (d). Each point represents a median of 10 experiments. When $m \leqslant 20$, the number of all possible configurations of the case–control status of individuals with minor alleles was smaller than 10^7; ER, therefore, obtained the exact resampling p-values. The number of variant loci was 30 when $m \geqslant 30$, otherwise, it was the same as m.

depending on the scenario, and UA was both the most conservative for balanced studies at MAC $\leqslant 10$, and the most anticonservative for imbalanced studies. We observed similar trends for the Burden test (Supplementary Figure S2) and SKAT-O (Supplementary Figure S3).

ER-mid based p-values are conservative for variant sets with MAC < 20 because many of the variant sets cannot reach Bonferroni-corrected thresholds. To improve the calibration of ER-mid, we used a mixture model (Supplementary Appendix C) to estimate the effective number of tests (K_{eff}) defined as the number of independent tests that yields the expected Bonferroni corrected FPR (Figure 4). For SKAT-ER-mid, when MAC $\leqslant 10$, K_{eff} was substantially smaller than the number of variant sets, especially for balanced studies. The K_{eff}-based Bonferroni correction had a slightly anti-conservative FPR for balanced case–control samples but well-calibrated FPRs for imbalanced case–control samples. The computation time for the K_{eff}-based multiple test adjustment are essentially the sum of the computation time to test each variant set, as fitting the mixture model requires little additional computation. We observed similar patterns of results for Burden test (Supplementary Figure S4) and SKAT-O (Supplementary Figure S5).

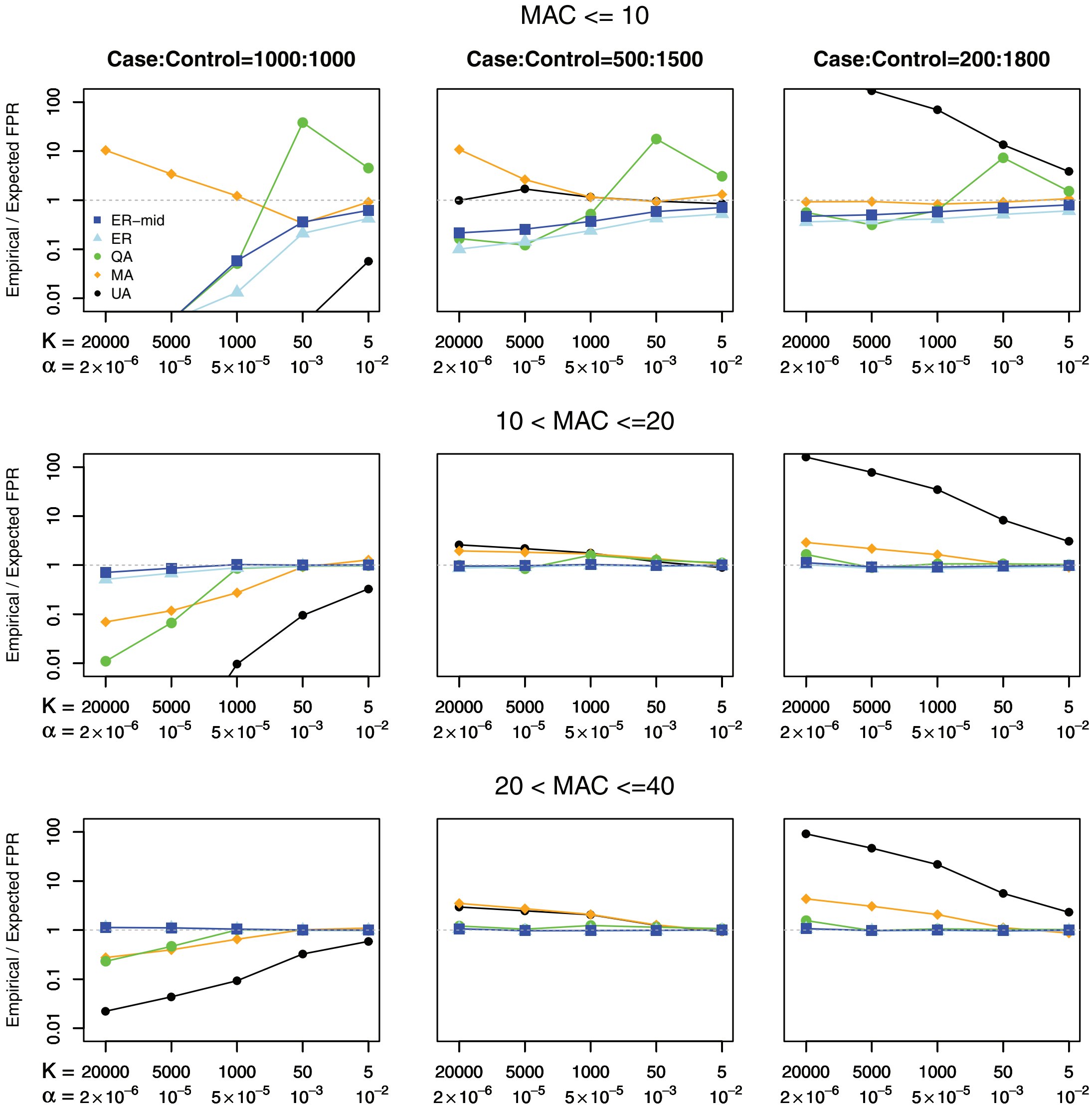

Fig. 3. False positive rates (FPRs) for SKAT using ER-based and existing methods to compute p-values for variant sets with MAC $\leqslant 40$. From top to bottom the plots show variant sets with MAC $\leqslant 10$; $10 <$ MAC $\leqslant 20$ and $20 <$ MAC $\leqslant 40$. From left to the right, the plots consider case:control = 1000:1000, 500:1500, and 200:1800. In each plot, the x-axis is the number of variant sets (K) and their corresponding Bonferroni corrected level $\alpha (= 0.05/K)$, and the y-axis is the empirical FPRs divided by the expected FPR. A well-calibrated test should have empirical/expected FRP = 1 (gray dashed line).

Next, we examined the FPRs for sets of variants with $40 <$ MAC $\leqslant 500$ in the presence of covariates. SKAT-ER-mid was generally well calibrated, although it was slightly conservative or anti-conservative at $\alpha = 2.5 \times 10^{-6}$ (Supplementary Figure S6). SKAT-QA was slightly conservative for balanced studies and slightly anti-conservative for studies with case–control imbalance. SKAT-MA was well calibrated or slightly anti-conservative for balanced studies, and was anti-conservative for imbalanced studies. SKAT-UA was not well calibrated in any of these scenarios. For Burden tests, all methods had close to the expected

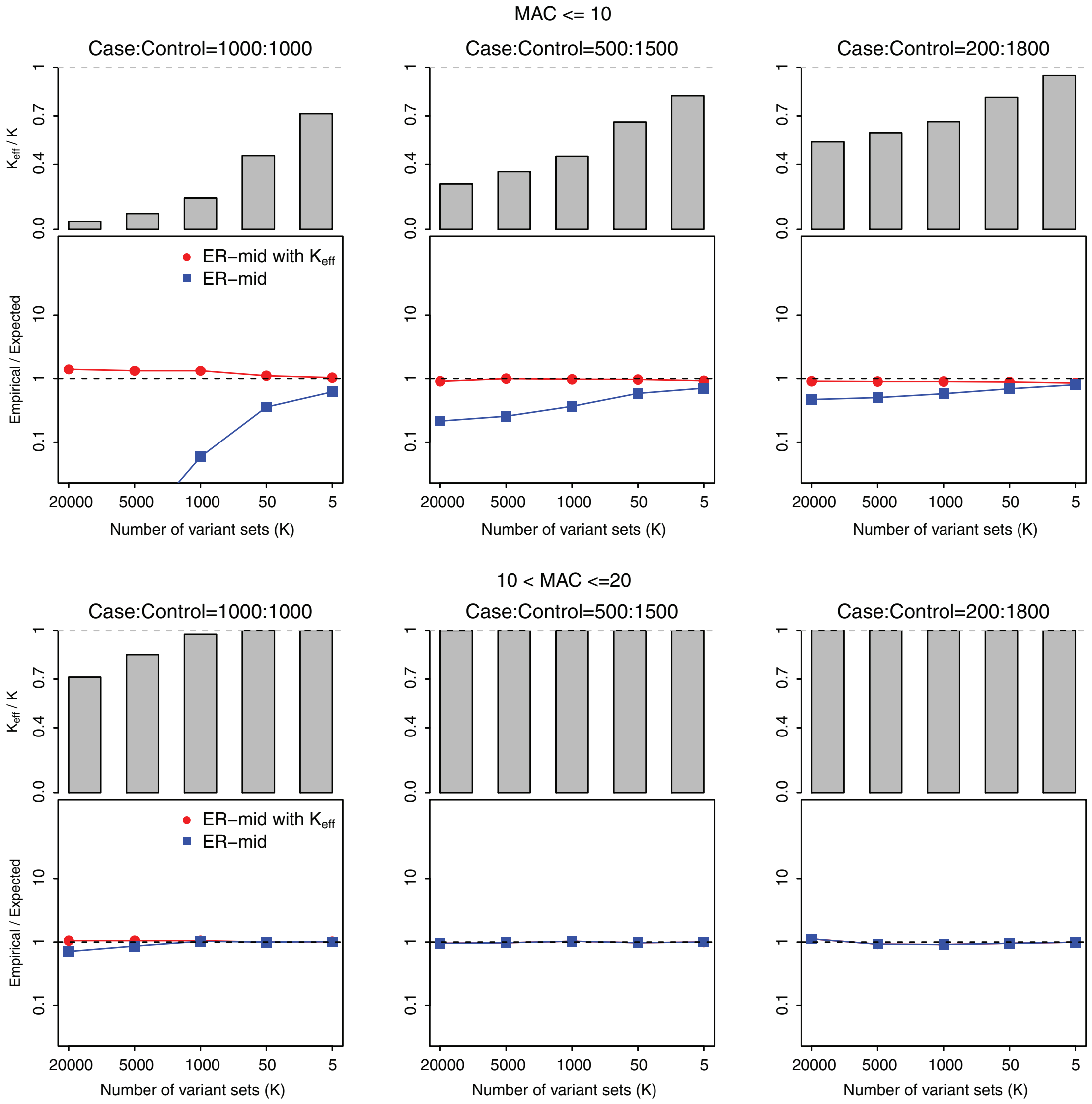

Fig. 4. Estimated effective number of tests (K_{eff}) and FPRs for SKAT-ER-mid for variant sets with MAC $\leqslant$ 20. Variant sets with MAC $\leqslant$ 10 (top row) and 10 < MAC $\leqslant$ 20 (bottom row) are shown. From left to the right, the plots consider case:control = 1000:1000, 500:1500, and 200:1800. In each plot, the top panel shows a bar plot of the estimated effective number of tests (K_{eff}) divided by the number of variant sets (K), and the bottom panel shows the empirical false positive rate (FPR) divided by the expected FPR of SKAT-ER-mid based on K (square) or K_{eff} (circle). A well-calibrated test should have empirical/expected FRP = 1 (black dashed line). The x-axis shows the number of variant sets (K).

FPRs for balanced studies and Burden-QA was best calibrated for unbalanced studies (Supplementary Figure S7). We observed similar patterns of results for SKAT-O (Supplementary Figure S8).

Overall, the results were quantitatively the same in the absence of covariates or when, instead of testing a set of variants, we tested single variants (a test which very similar to a Burden test with equal weights

Table 1. *Number of genes by MAC of selected variants in NHLBI-ESP whole-exome data and in chromosome 2 GoT2D-exome data*

	$1 \leqslant \text{MAC} \leqslant 10$	$10 < \text{MAC} \leqslant 20$	$20 < \text{MAC} \leqslant 40$	$40 < \text{MAC} \leqslant 100$	$100 < \text{MAC}$	Total
NHLBI ESP						
Disruptive	7261 (62%)	1425 (12%)	1313 (11%)	1306 (11%)	485 (4%)	11 790
Disruptive + potentially damaging	4250 (25%)	2636 (15%)	3135 (18%)	4034 (23%)	3185 (18%)	17 240
All nonsynonymous	1699 (9%)	1579 (9%)	2568 (14%)	4791 (27%)	7371 (41%)	18 008
GoT2D Chr2						
Disruptive	312 (92%)	17 (5%)	5 (1%)	6 (2%)	0 (0%)	340
Disruptive + potentially damaging	481 (46%)	174 (17%)	186 (18%)	161 (15%)	37 (4%)	1039
All nonsynonymous	284 (26%)	165 (15%)	208 (19%)	330 (30%)	123 (11%)	1110

Each cell has the number (percent) of genes in each MAC bin for genes with $\geqslant$1 variant. "Total" indicates the total number of genes with $\geqslant$1 variant. Nonsense, splicing, and frame-shift variants are classified as "disruptive" variants, and possibly and probably damaging variants by Polyphen2 and disruptive variants together are classified as "disruptive + potentially damaging" variants.

for all variants) (data not shown). To test for the robustness of our methods in the presence of population stratification, we simulated African American and European ancestry samples with a differential disease risk and adjusted for stratification in the analysis. The Type 1 error rates (Supplementary Appendix F and Supplementary Figures S9–S11) were quantitatively similar to those in Figure 3 and Supplementary Figures S2, S3 for European ancestry only.

Over a range of MAC and case–control ratios, no approach yielded an optimal mix of control of FPR and efficient computation. Based on our findings, we propose an ER-based hybrid approach (ER-mid when variant set $\text{MAC} \leqslant 40$; MA when variant set $\text{MAC} > 40$ and balanced case–control; and QA when variant set $\text{MAC} > 40$ and imbalanced case–control) to provide a balance of well-calibrated FPRs and computation time.

3.1.4 *Comparison of power to identify associations between low MAC variant sets and binary phenotypes.*

We next compared power for the ER-based hybrid approach using either experiment-wide permutations of the total sample or the effective number of tests (K_{eff}) based Bonferroni correction, and power for the MA or UA tests using experiment-wide permutations. We estimated the power to detect one causal variant set ($\text{MAC} = 20$) out of a background of 19 999 non-causal variant sets with the MAC distribution of disruptive + potentially damaging variants observed in NHLBI ESP data (Supplementary Appendix D and Table 1). Our causal variant set had 50% causal variants, either all increasing risk or with half the variants increasing and half decreasing risk. Over the different gene-based tests approaches and varying case control ratios, we observed similar power for ER-based hybrid approach using experiment wide permutations or K_{eff}-based Bonferroni correction (Supplementary Figure S12). For SKAT and SKAT-O, the ER-based hybrid approach had higher power than MA or UA. For the burden test, MA or UA had similar or slightly higher power to the ER-hybrid approach, but neither test was consistently higher power. We observed similar trends for causal $\text{MAC} = 40$ (Supplementary Figure S13).

3.2 *GoT2D data analysis*

We performed single and multiple variant tests using GoT2D chromosome 2 deep exome sequence data (1326 cases and 1331 controls) (Supplementary Appendix G). 35 576 (84%) of 42 045 chromosome 2 variants had $\text{MAF} < 0.01$ (corresponding $\text{MAC} \leqslant 53$). For single variant tests of $\text{MAF} < 0.01$ variants, the

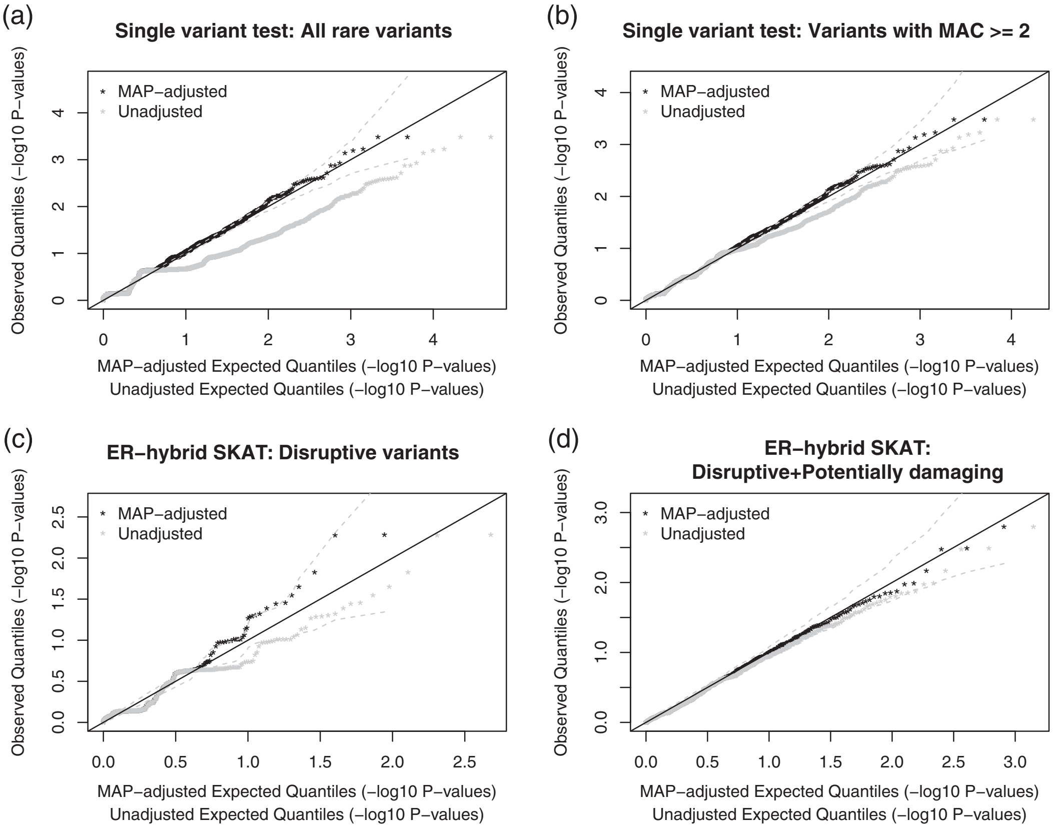

Fig. 5. MAP-adjusted and un-adjusted QQ plots of single variant and SKAT-ER-hybrid p-values from analysis of GoT2D chromosome 2 exome data. QQ plots of single variant tests with all rare variants (MAF < 0.01) (a) and rare variants with MAC $\geqslant 2$ (b). QQ plots of ER-hybrid SKAT p-values with disruptive variants (c) and disruptive + potentially damaging variants (d). In each plot, the x-axis is the MAP-adjusted or un-adjusted expected quantile of $-\log_{10}$ p-values, and the y-axis is observed quantiles of $-\log_{10}$ p-values. Observed p-values are plotted against the MAP-adjusted expected quantiles (black dots) and un-adjusted expected quantiles (gray dots). The dashed line represents a 95% confidence band based on 500 random draws from the MAP-based mixture distribution.

estimated effective number of tests (K_{eff}) was 2762, giving an order of magnitude less stringent threshold than the family-wise error rate 0.05. No variants were significant at K_{eff}-based Bonferroni-corrected $\alpha = 0.05$. The unadjusted QQ plot for single variant results showed a substantial p-value deflation compared with the expected p-value (Figure 5(a)); though the deflation was less pronounced when testing was restricted to variants with MAC $\geqslant 2$ (Figure 5(b)). In contrast, in QQ plots based on a mixture model of the minimum achievable p-values, no p-value deflation was observed (Figures 5(a) and (b)).

In the chromosome 2 GoT2D data, 334 of 340 (98%) genes with at least one disruptive variant had $1 \leqslant$ MAF $\leqslant 40$, and 841 of 1039 (81%) genes with at least one disruptive + potentially damaging variant had $1 \leqslant$ MAC $\leqslant 40$ (Table 1). Even in the whole-exome data from the larger NHLBI-ESP sample, 85% and 58% of genes with at least one disruptive or disruptive + potentially disruptive variant, respectively, had $1 \leqslant$ MAC $\leqslant 40$ (Supplementary Appendix D and Table 1). We used SKAT-ER-hybrid to perform gene-based tests for disruptive and disruptive + potentially damaging variants ($K_{\text{eff}} = 44$ and 540, respectively) in the chromosome 2 GoT2D exome data. No gene was significant at the K_{eff}-based Bonferroni corrected

$\alpha = 0.05$. In unadjusted QQ plots, we observed deflation of the gene-based p-values, whereas in MAP adjusted QQ plots the p-values were not deflated and results for disruptive variants were near the upper 95% confidence bound (Figures 5(c) and (d)). We observed similar results for ER-hybrid Burden and SKAT-O tests (Supplementary Figures S14 and S15).

Within the disruptive + potentially damaging variant tests, YSK4 Sps1/Ste20-related kinase homolog (*YSK4*) was the most significant gene for the Burden-ER-mid test (p-value $= 1.7 \times 10^{-3}$, MAC $= 27$) and the second most significant gene for SKAT-O-ER-mid (p-value $= 5.2 \times 10^{-3}$). Recent large-scale meta-analysis has shown that a common variant in *YSK4* is associated with fasting insulin (Scott *and others*, 2012).

To assess the ER method using dosage data, we compared the results of ER and whole-sample permutations for variant set-based testing using dosage data from non-exomal GOT2D low-pass sequencing and found very similar p-values (Supplementary Appendix H and Supplementary Figure 16).

4. Discussion

In this paper, we develop an ER method for binary traits for score statistic-based tests of variant sets with low MAC that allows inclusion of covariates in analysis. The ER methods are necessary because the existing asymptotic (UA) or asymptotic-based adjustment methods (MA) have poor calibration of FPRs at lower MAC and imbalanced case control ratios. As in whole-sample permutations, the ER method preserves the correlation structure or LD among variants in the tested set. Across almost all tested MAC bins and case–control ratios, we found that one or more of the ER-based methods were well calibrated. Based on these observations and the computational time considerations, we recommend a hybrid approach using ER-mid for small variant set MAC (MAC $\leqslant 40$); MA for moderate or large variant set MAC with balanced case–control and QA for moderate or large variant set MAC with unbalanced case–control. Use of a threshold of MAC $= 40$ is a practical compromise between computational time and Type 1 error rate; a slightly lower threshold would result in faster computation time but at the risk of slightly higher Type 1 error rate, particularly for the SKAT and SKAT-O. If a permutation approach is desired, then ER-mid is (substantially) faster than whole-sample permutations even for large MAC.

Estimation of the effective number of tests, K_{eff}, using MAP is a simple and fast alternative to performing experiment-wise permutation of the total sample to control the family-wise error rate. One limitation of the MAP approach is that it cannot account for correlations among tests, and may result in conservative FPRs in the presence of the strong correlations of variants between genes. However, we expect that gene-based tests will be less correlated than single variant tests, since they involve multiple variants and genes located further away from each other than individual variants.

When MAC is extremely small, MAP is unlikely to reach genome-wide significance. One approach to increase power would be to construct larger sets by combining adjacent regions or including more classes of potentially functional variants.

The ER method can be used for imputed dosage, as well as genotype data; permutations are performed within the individuals with non-zero genotype or dosage values. If many individuals have very small dosage values (e.g. <0.1), the number of individuals with minor alleles can be larger than MAC (i.e. MAC $< m$). Thus, for the same MAC, computational time can be higher with dosage data than with genotype data; however, the ER method still takes substantially less time than whole-sample permutation method.

QQ plots comparing observed vs. expected p-value distributions are used in genetic association studies to assess both the presence of confounding (or misimplimented/misspecified test) and the presence of significant association signals. However, when MAC is small, the expected p-value distribution of the resampling-based test is not uniform (0,1), and hence the (unadjusted) QQ plot cannot be used to accurately assess the concordance (or departure) of the observed p-value distribution from the expected. In the spirit of experiment wide permutations (Kiezun *and others*, 2012), we use the MAP-adjusted p-value distribution to model the expected distribution of ER-hybrid p-values. In the MAP-adjusted QQ plot, the GoT2D gene-based p-value distribution for disruptive variants lies near the top of the 95% confidence band. This view allows better assessment of potentially interesting results than the unadjusted QQ plot in which the p-value distribution is deflated.

Most of variant sets in whole-exome or whole-genome data will not require 10^7 resampling since their p-values will be substantially higher than exome-wide (or genome-wide) significant levels. Hence, an adaptive resampling procedure, which reduces the number of resamples when a test has a moderate or large p-value, can substantially reduce computation time and has been implemented for the ER method. However, the use of adaptive resampling precludes the calculation of the effective number of test and the use of MAP-adjusted QQ plots, and thus we recommend the adaptive resampling procedure only for the case where case–control combinations among individuals with minor alleles are substantially larger than the number of resamples performed (for example, MAC > 20 for 10^7 resamples).

Our work has focused on providing well-calibrated gene-based tests for single studies across a range of MAC and case–control imbalance. Meta-analysis of gene-based tests can increase the power to detect genes of interest, but meta-analysis is sensitive to the calibration of the underlying tests (Ma *and others*, 2013), and may be particularly sensitive to the inclusion of studies with highly imbalanced case–control ratios. Further work will be needed to determine how best to combine results or data from across studies with a variety of case–control ratios.

5. Software

ER-mid, ER, QA, and MA methods are implemented in the SKAT R-package.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank investigators of GoT2D project for access to the chromosome 2 exome sequence data. We also thank M. Boehnke for discussion and insightful comments and Phoenix Kwan for her initial insights into the behavior of gene-based tests in the GOT2D data. *Conflict of Interest*: None declared.

Funding

This work was supported by grants R00 HL113264 (S.L.), the Austrian Science Fund (F.W.F.) grant J-3401 (C.F.), R01 HG000376 and RC2 DK088389 (L.S.).

References

Derkach, A., Lawless, J. F. and Sun, L. (2012). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology* **37**, 110–121.

Efron, B. and Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. CRC press.

Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S. and Satten, G. A. (2012). A permutation procedure to correct for confounders in case–control studies, including tests of rare variation. *American journal of human genetics* **91**, 215–223.

Fog, A. (2008). Calculation methods for Wallenius' noncentral hypergeometric distribution. *Communications in Statistics—Simulation and Computation* **37**, 258–273.

Kiezun, A., Garimella, K., Do, R., Stitziel, N. O., Neale, B. M., McLaren, P. J., Gupta, N., Sklar, P., Sullivan, P. F. and Moran, J. L. (2012). Exome sequencing and the genetic basis of complex traits. *Nature Genetics* **44**, 623–630.

Lancaster, H. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association* **56**, 223–234.

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

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M. and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics* **91**, 224–237.

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

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

Lin, D. Y. and Tang, Z. Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics* **89**, 354–367.

Ma, C., Blackwell, T., Boehnke, M. and Scott, L. J. (2013). Recommended joint and meta-analysis strategies for case–control association testing of single low-count variants. *Genetic Epidemiology* **37**, 539–550.

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

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K. and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics* **7**, e1001322.

Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* **33**, 497–507.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. and Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575.

Reppell, M., Boehnke, M. and Zöllner, S. (2012). FTEC: a coalescent simulator for modeling faster than exponential growth. *Bioinformatics* **28**, 1282–1283.

Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J. A. and Gustafsson, S. (2012). Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature Genetics* **44**, 991–1005.

Sun, J., Zheng, Y. and Hsu, L. (2013). A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genetic Epidemiology* **37**, 334–344.

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

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R. and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455–E464.

[*Received December 16, 2014; revised August 3, 2015; accepted for publication August 10, 2015*]