Supplementary Materials

Seunggeun Lee¹,², Christian Fuchsberger¹,², Sehee Kim¹, and Laura Scott¹,²

¹ Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109
² Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109
Appendix

A. ER method: details of the step 1 and step 2

Overview of Step1 and Step2:
The overall goal of this procedure is to perform resampling of case/control status within $m$ individuals with minor alleles based on probability that $d$ individuals are cases, and to use the resulting permutations to estimate the probability of obtaining a statistic as or more significant than the observed test statistic. To resample case status within the $m$ individuals, we cannot simply permute the observed case and controls status within the $m$ individuals as the number of case individuals with the minor alleles would remain constant across permutations. To address this, in Step 1 we estimate the probability of $d$ cases in $m$ individuals as the conditional probability $\Pr(D=d \mid Y, G, X)$ given genotypes ($G$), phenotypes ($Y$) and covariates ($X$) using Fisher’s noncentral hypergeometric distribution. We then use the probability $D=d$ to determine the number of permutations to perform in each strata of $d$ given a total of $B$ permutations. In Step 2 we estimate a conditional tail probability of the test statistic $Q$, $\Pr(Q > \hat{Q} \mid D=d, Y, G, X)$, given ($Y,G,X$) and $D=d$ by generating $B_d$ permutations of the case-control status of $m$ individuals, where $\hat{Q}$ is a test statistic from the original phenotype. Using the results of Step 1 and 2 we calculate the ER p-values and MAPs as described in the main manuscript in Section 2.2.

Definition of Fisher’s noncentral hypergeometric distribution:
Fisher’s noncentral hypergeometric distribution can be described using the following urn model. Suppose there are $S$ different colors of balls in the urn, and each has $h_s$ balls. The odds of selecting the $s$th colored ball is $\omega_s$. We randomly select a fixed number of balls
without replacement, and among them \( a_s \) balls are the color \( s \). The probability mass function of this process for \( a=(a_1, ..., a_S) \) is

\[
f(a; h, \omega) = \phi_0 \prod_{s=1}^{S} \binom{h_s}{a_s} \omega_s^{a_s}, \tag{S.1}
\]

where \( h=(h_1, ..., h_S) \), \( \omega=(\omega_1, ..., \omega_S) \) and \( \phi_0 \) is a normalizing constant making \( \sum_a f(a; h, \omega) = 1 \).

**Detailed description of Step 1:**

When no covariates require adjustment, \( D=d \mid (Y, G) \) follows a central hypergeometric distribution with the probability mass function:

\[
Pr(D=d \mid Y, G) = \binom{m}{d} \binom{n-m}{n-d}, \tag{S.2}
\]

which is equivalent to \( f(a; h, \omega) \) in (S.1) when \( a = (d, n_{case}-d) \), \( h = (m, n - m) \), and \( \omega = (\omega_1, \omega_2) \) with \( \omega_1=\omega_2 \). We note that (S.2) is also equivalent to \( f(a; h, \omega) \) when \( a = (d, m-d) \), \( h = (n_{case}, n - n_{case}) \) and \( \omega = (\omega_1, \omega_2) \) with \( \omega_1=\omega_2 \).

When covariates are included in the analysis we estimate probability of being a case for each individual \( \pi_i \) for \( i = 1, ..., n \), from the following logistic regression model under the null hypothesis of no genetic effects,

\[
\logit(\pi_i) = \alpha_0 + X_i'\alpha,
\]

where \( X_i = (x_{i1}, ..., x_{iq})' \) is a vector of covariates, \( \alpha_0 \) is the intercept, and \( \alpha=(\alpha_1, ..., \alpha_q)' \) are regression coefficients of covariates.

Because the calculation of the exact probabilities of \( Pr(D=d \mid Y, G, X) \) are computationally expensive when \( m \) is large, we stratify \( m \) individuals into \( S-I \) groups in which individual \( i \) belongs to group \( s \) \((s=1, ..., S-I)\) if \( \hat{\pi}_i \) is between \((s-1)/(S-I)\) and \( s/(S-I)\)
1). We chose to stratify $\pi_i$ in this way rather than by equally sized groups in order to preserve the effect of any extreme values of $\pi_i$ on the probability of $D=\text{d}$. The remaining $n-m$ individuals belong to the last group $S$. Suppose the group $s$ has $h_s$ individuals. Let $\bar{\mu}_s$ be the group-specific average of $\pi_i$ in the group $s$, and $\omega_s = \bar{\mu}_s/(1 - \bar{\mu}_s)$. Then, $Pr(D = \text{d}\mid Y, G, X) = \sum_{a_1 + \cdots + a_{S-1} = \text{d}} f(a; h, \omega)$, where $a_s$ is the number of cases in the group $s$. This conditional probability can be estimated numerically by enumerating all possible combinations of $(a_1, \ldots, a_S)$. We use $S=11$ in our simulations and in analysis of data.

The estimate of $Pr(D = \text{d}\mid Y, G, X)$ is used to determine the number of resamplings from each strata $d$ when the total number of configurations of case-control status ($C_T$) in the $m$ individuals is larger than $B$. We set $B_d$ to be proportional to $Pr(D=\text{d}\mid Y, G, X)$, i.e. $B_d = B \times Pr(D=\text{d}\mid Y, G, X)$. $Pr(D=\text{d}\mid Y, G, X)$ is also used to calculate ER $p$-values and MAPs as described in the main manuscript Section 2.2.

**Detailed description Step 2:**

Without loss of generality, the first $m$ individuals are assumed to have minor alleles. When $B = C_T$, we generate all possible configurations of case-control status. When $B < C_T$, within each strata of $D=\text{d}$, we perform $B_d$ permutations of the case-control status of the first $m$ individuals without regard to the individual’s odds of being a case. Let $y_{d}^{(b)} = (y_{1d}^{(b)}, \ldots, y_{md}^{(b)})^T$ be the $b$th permutation sample, where $b=1, \ldots, B_d$. To be able to take the different probabilities of observing each permutation into account in our analysis, for each permutation sample $b$ we calculate $P_{db} = f(y_{d}^{(b)}; h, \omega)$, the probability
of having \( y_d^{(b)} \) using the equation (S.1). Unlike Step 1, we enter each of the \( m \) individuals separately into Fisher’s noncentral hyper-geometric distribution, with \( h_i = 1 \), and estimate each individuals’ odds of being a case, as \( \omega_i = \hat{\alpha}_i / (1 - \hat{\alpha}_i) \). In an alternative approach, \( y_d^{(b)} \) can be generated from the Fisher’s noncentral hyper-geometric distribution (S.1) with \( h_i = 1 \) and \( \omega_i = \hat{\alpha}_i / (1 - \hat{\alpha}_i) \), for \( i = 1, \ldots, m \). In this approach, \( P_{db} = 1 \).

The \( b \)th resample of \( S_j \) for \( D = d \) is

\[
S_{jd}^{(b)} = \sum_{i=1}^{m} \left( y_{id}^{(b)} - \hat{\alpha}_i \right) g_{ij},
\]

and let \( Q_d^{(b)} \) be the resulting resampled test statistic \( Q \). Examples of \( Q_d^{(b)} \) include the resampled Burden test statistics \( Q_{Burden,d}^{(b)} = \left( \sum_{j=1}^{p} w_j S_{jd}^{(b)} \right)^2 \) and the resampled SKAT test statistics \( Q_{SKAT,d}^{(b)} = \sum_{j=1}^{p} \left( w_j S_{jd}^{(b)} \right)^2 \). Finally, the conditional probability \( Pr \left( Q > \hat{Q} \mid D = d, Y, G, X \right) \) is estimated by

\[
Pr \left( Q > \hat{Q} \mid D = d, Y, G, X \right) = \sum_{b=1}^{B_d} I \left( Q_d^{(b)} \geq \hat{Q} \right) P_{db}.
\]

**B. ER-based quantile-adjusted moment matching for Burden test, SKAT and SKAT-O**

Since the computational cost of ER increases as the number of individuals with variant alleles increases, it may not be practical to use ER for variant sets with moderate or large MAC. In this section, we propose a quantile-adjusted moment matching for these variant sets. Moment matching adjustment is a method to adjust for the asymptotic null distributions by estimating the small sample mean, variance and kurtosis (Lee and others,
2012) from a generated moderate number of resampled test statistics. Since it is based on moments, it cannot accurately approximate the tail area of the distribution, especially in the presence of imbalanced case-control samples. Therefore, we further calibrate the moment matching adjustment using quantiles of test statistics estimated through ER.

Suppose that $\mu_Q$, $\nu_Q$, and $\gamma_Q$ are the estimated small sample mean, variance and kurtosis of the test statistic $Q$, respectively, which can be from either SKAT or Burden test. The moment-matching adjustment transforms $Q$ to

$$Q_{mm} = (Q - \mu_Q) \frac{2\delta}{\nuar_Q} + \delta,$$

where $\delta=12/\gamma_Q$ is an estimated degree of freedom, and assumes that $Q_{mm}$ follows $\chi^2$, the chi-square distribution with d.f. $= \delta$ (Lee and others, 2012). For the extra adjustment, we generate $B$ resamples of $Q_{mm}$ through ER and compare the empirical quantiles of $Q_{mm}$ and $\chi^2$. We fit a linear interpolation function between empirical quantiles of $Q_{mm}$ and $\chi^2$, and evaluate the observed test statistics on this function to transform it to have the same quantiles of $\chi^2$. For SKAT-O, this quantile adjustment is used to compute a $p$-value of each $Q_\rho$. In the simulation and sequence data analysis, we generated 500,000 resamples to obtain empirical quantiles.

C. Calibration of the number of test and quantile-quantile (QQ) plot

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 et al (Kiezun and others, 2012) proposed a heuristic approach in which to first identify variant sets with MAP $< 0.001$, and to count only these
variant sets as the effective number of tests. Here we developed an alternative statistical approach to estimate the effective number of test using MAP.

Suppose that we are conducting $K$ tests, (i.e., $K$ variant sets). Let $P_{val,k}$ and $MAP_{k}$ be the $p$-value and MAP of the $k$th test, respectively. They can be either $p$-values or mid-$p$-values and their MAPs. To account for the fact that the $p$-value of the $k$th test has a low limit at $MAP_{k}$ with a point mass $MAP_{k}$, we assume that $P_{val,k}$ follows a mixture distribution

$$r_{k}MAP_{k} + (1 - r_{k})U(MAP_{k}, 1), \quad (S.3)$$

where $r_{k}$ is a Bernoulli random variable with success probability $MAP_{k}$, and $U(MAP_{k}, 1)$ is a uniform distribution between $MAP_{k}$ and 1. The mixture distribution (S.3) will be used to estimate the effective number of tests as well as for QQ plot adjustment.

Let $K_{eff}$ be the number of tests for the Bonferroni correction in which Bonferroni corrected level $\alpha$ is $\alpha / K_{eff}$. Then the family-wise error rate with $K_{eff}$ is approximately

$$\sum_{k=1}^{K} Pr\left(P_{val,k} < \frac{\alpha}{K_{eff}}\right) \approx \sum_{k=1}^{K} Pr\left(P_{val,k} < \frac{\alpha}{K_{eff}} | MAP_{k} < \frac{\alpha}{K_{eff}}\right)I\left(MAP_{k} < \frac{\alpha}{K_{eff}}\right)$$

$$\approx \frac{\alpha}{K_{eff}} \sum_{k=1}^{K} I\left(MAP_{k} < \frac{\alpha}{K_{eff}}\right),$$

which is equal to $\alpha$ when $K_{eff} = \sum_{k=1}^{K} I(MAP_{k} < \alpha/ K_{eff})$. We estimate $K_{eff}$ as the solution of this equation. The $K_{eff}$ is smaller than or equal to $K$, and equality holds when all MAPs are smaller than $\alpha / K$.

To adjust the QQ plot, we assume that the expected distribution of $p$-values is the mixture distribution (S.3), and estimate theoretical quantiles of $p$-values by generating random variables from this mixture distribution. In particular, we generate a random
variable for each \( k (k=1, \ldots, K) \), using the equation (S.3), and calculate their quantiles. We repeat this procedure 500 times and obtain the median of each obtained quantile, and use them as theoretical quantiles of \( p \)-values to compare with their empirical counterparts. When \( MAP_k \) estimate is not available due to large MAC at the \( k \)th test or the use of asymptotic based methods in calculating the \( k \)th \( p \)-value, we use \( MAP_k =0 \), which may increase the effective number of tests, and may cause Bonferroni correction and QQ plot adjustment slightly conservative.

D. NHLBI ESP data

We downloaded the 4300 NHLBI ESP European ancestry samples (Tennessen and others, 2012) from the exome variant server and calculated the total MAC in each gene as the sum of minor alleles for variants with MAF<.01 ((MAC < 86) for three different categories of variants, We defined the three categories of variants to reflect their potential impact on protein function: 1) “disruptive” for nonsense, splicing and frame-shift variants (Zuk and others, 2014); 2) “disruptive+potentially damaging” for disruptive variants plus variants classified as possibly and probably damaging by Polyphen2 (Adzhubei and others, 2010); and 3) “nonsynonymous” for all protein code changing variants. When the total number of allele counts of a variant was < 8600 due to missing genotypes, MAC was estimated as observed MAF \( \times 8600 \).

E. Details of false positive rate and power simulations
To verify that ER and the whole sample permutation methods produce essentially identical $p$-values, we generated 20000 variants sets and compared the $p$-values from ER and the permutation methods with and without covariates by generating $10^5$ resamples.

Variant sets were generated as described above and only sets with MAC $\leq$ 40 were included. The phenotypes were generated from equation (2) in Section 2.3 with $\beta_{causal} = 0$. For the whole sample permutations, in the absence of covariates we permuted case/control status (Perm). In the presence of covariates we used the Fisher’s noncentral hypergeometric distribution based permutation (FNHPerm) from BiasedUrn R-package (ver 1.06) (Epstein and others, 2012).

We compared computation times of SKAT-ER to whole sample permutation for $m=40$ and total sample sizes ranging from 100 to 50,000 (balanced case-control ratio). Binary phenotypes were generated with and without covariates as described above. ER method is implemented in R with core functions written in C. Whole sample permutations with no covariates were implemented in C, and BiasedUrn R-package was used for the Fisher’s noncentral hypergeometric distribution based permutation. We also estimated computation times using existing software packages: PLINK/SEQ (ver 1.09) and SCORE-seq (ver 5.2). In addition, for a sample size of 2000 (balanced case-control), we estimated the computation time for the Burden, SKAT, and SKAT-O tests using ER ($10 \leq m \leq 40$) and QA ($40 \leq m \leq 500$). When $m > 30$, the number of variant loci ($p$) was fixed at 30, otherwise it was the same as $m$.

To compare the false positive rate for different ranges of total MAC, we considered six total MAC bins: MAC $\leq 10$; $10 <$ MAC $\leq 20$; $20 <$ MAC $\leq 40$; $40 <$ MAC $\leq 100$; $100 <$ MAC $\leq 200$; and $200 <$ MAC $\leq 500$. For each bin, we used ranges of the
number of variant sets $K=5$ to $20000$, corresponding to candidate gene studies to genome-wide studies. To reduce the computational burden, we first generated independently $100 \times K$ variant sets in randomly selected regions (e.g., $2 \times 10^6$ variant sets when $K=20000$), and then generated 1000 phenotype and covariate sets per each variant set. The phenotypes were generated from equation (2) with $\beta_{\text{causal}} = 0$ in Section 2.3. Bonferroni corrected $\alpha=0.05 \times (0.05/K)$ was used to identify significant $p$-values per each phenotype and covariates set. We obtained the empirical false positive rate as the ratio of the average number of significant $p$-values to the number of variant sets ($K$). For ER, we evaluated all possible resamples if the total number of combinations was smaller than $10^7$; otherwise, we generated $10^7$ sets of resamples to compute the $p$-values.

For power comparisons, we considered that there exists only one low MAC causal variant set among 20000 variant sets. Non-causal variant sets were generated independently using the same approach in false positive simulations, and had the same MAC distribution of disruptive+potentially damaging variants in NHLBI ESP in Table 1. The causal variant set was randomly selected, and binary phenotypes were generated from the equation (2) in Section 2.3. Two different values of MAC (20 and 40) were considered for causal variants set. 50% of variants in the causal variant set were selected as causal variants, and log odds ratio $\beta = c|\log_{10}(MAF)|$ with $c=1.15$. We considered two scenarios in which all non-zero $\beta$s were positive, and only 50% of non-zero $\beta$s were positive. The significance levels for family-wise error rates = 0.05 were estimated empirically using the experiment-wise permutation. Causal variant sets were generated 1000 times and the empirical power was obtained as a fraction of $p$-values of the causal variant sets lower than the estimated significance level.
F. False positive rate simulations in the presence of population stratification

To evaluate false positive rate control in the presence of population stratification, we generated 10,000 European (EUR)-like and 10,000 African American (AA)-like sequence haplotypes using a coalescent simulator, COSI (Schaffner and others, 2005). We used COSI instead of FTEC because COSI provides parameters for generating AA-like sequence haplotypes. Genotypes of EUR and AA samples were generated by randomly matching haplotypes, and the probability of being EUR was 0.5. The binary phenotypes were generated from the following logistic regression model:

\[
\text{logit } P(Y = 1) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2,
\]

where \(X_1\) is a binary covariate taking a value of 0 for EUR samples and a value of 1 for AA samples, and \(X_2\) was a continuous covariate following \(N(0,1)\). The coefficient \(\alpha_1\) was set to be 0.5, assuming that an AA is more likely to have a higher risk of being cases. The intercept \(\alpha_0\) was chosen for the disease prevalence of 0.05, and \(\alpha_2\) was 0.5.

We applied the proposed method to compute p-values for each of the Burden, SKAT and SKAT-O tests. We considered that the true ancestry (i.e. \(X_1\)) was known and included as a covariate to adjust for. The Figures S9-S11 show false positive rates of each method when MAC < 40. The overall performances of the proposed method were as good as results without population stratification. We also achieved the same good performances when MAC > 40 (data not shown).

G. Details of GoT2D data analysis

We applied our tests to deep exome sequence data from the Genetics of Type-2 Diabetes (GoT2D) sample of 1,326 European T2D cases and 1,331 European controls (drawn from
the DGI, FUSION, UKT2DGO, and KORA studies) (Gaulton and others, 2013). For illustration, we analyzed variants with MAF < 0.01 in the chromosome 2 deep exome sequence dataset. We performed single variant tests on all variants (regardless of their functional classes), and carried out gene-based tests based on the three functional variant classes described in Web Appendix D. We included covariates sex, two principal components and two batch effect variables.

For single variant tests, we used ER to obtain the mid-$p$-values of score test when $\text{MAC} \leq 20$, and used the Firth biased corrected likelihood ratio test (Firth, 1993) in all other cases. The Firth test is known to provide good type I error control for single variant test with moderate and large MACs (Ma and others, 2013). For gene based tests, we used the ER hybrid approach, with ER-mid to compute $p$-values of Burden, SKAT and SKAT-O when $\text{MAC} \leq 40$, and MA otherwise. We estimated the effective number of tests and obtained MAP-adjusted QQ plots as described in Web Appendix C. The ANNOVAR (Wang and others, 2010) package was used for the annotation, and Polyphen2 (Adzhubei and others, 2010) was applied to predict the functional effects of non-synonymous variants.

**H. Comparison of $p$-values of dosage data obtained using ER or whole sample permutations**

We compared SKAT $p$-values calculated from ER and whole sample permutation methods using dosage data from the GoT2D project. The GoT2D project used low-pass sequencing (4x) to sequence non-exome regions and used a LD-aware variant caller, which adopts the imputation strategy to improve the quality of variant call in low-pass
sequencing data. We randomly selected 20,000 1kbp regions and compared SKAT p-values of ER and whole sample permutations. On average, the selected regions had MAC=21 and m=94. We used the original case-control phenotypes (1,326 cases and 1,331 controls) and p-values were calculated with and without covariates. Figure S16 shows that -log10 SKAT p-values were very highly concordant with Pearson-correlation > 0.99. We observed equally concordant p-values for Burden and SKAT-O tests. These results clearly conclude that the proposed ER method is well applicable to dosage data.

References


Supplementary Figures and Tables

Figure S1. Distribution of minor allele frequencies for simulated and GoT2D sequence data

MAFs for simulated data are based on the population allele frequencies under the faster than exponential growth model (A). GOT2D exome MAFs estimated for all variants (B), nonsynonymous variants (C) and potentially damaging variants (D).
Figure S2. False positive rates for Burden test using ER-based and existing methods to compute \( p \)-values for variant sets with \( \text{MAC} \leq 40 \)

From top to bottom the plots show variant sets with \( \text{MAC} \leq 10 \); 10 < \( \text{MAC} \) ≤ 20 and 20 < \( \text{MAC} \) ≤ 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 false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FPR =1 (gray dashed line).
Figure S3. False positive rates for SKAT-O using ER-based and existing methods to compute p-values for variant sets with MAC ≤ 40

From top to bottom the plots show variant sets with MAC ≤ 10; 10 < MAC ≤ 20 and 20 < MAC ≤ 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 false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FPR =1 (gray dashed line).
Figure S4. Estimated effective number of tests ($K_{eff}$) and false positive rates for Burden-ER-mid for variant sets with MAC ≤ 20

Variant sets with MAC ≤ 10 (top row) and 10 < MAC ≤ 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 $K$ (blue line) or $K_{eff}$ (red line). A well-calibrated test should have empirical/expected FPR =1 (black dashed line). The x-axis shows the number of variant sets ($K$) and their corresponding Bonferroni corrected level $\alpha$ ($=0.05/K$).
Figure S5. Estimated effective number of tests ($K_{eff}$) and false positive rates for SKAT-O-ER-mid for variant sets with MAC ≤ 20

Variant sets with MAC ≤ 10 (top row) and 10 < MAC ≤ 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 $K$ (blue line) or $K_{eff}$ (red line). A well-calibrated test should have empirical/expected FPR =1 (black dashed line). The x-axis shows the number of variant sets ($K$) and their corresponding Bonferroni corrected level $\alpha (=0.05/K)$. 
Figure S6. False positive rates for SKAT using ER-based and existing methods to compute p-values for variant sets with 40 < MAC ≤ 500

From top to bottom the plots show variant sets with 40 < MAC ≤ 100, 100 < MAC ≤ 200 and 200 < MAC ≤ 500. 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 false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FRP =1 (gray dashed line).
Figure S7. False positive rates for Burden test using ER-based and existing methods to compute p-values for variant sets with 40 < MAC ≤ 500

From top to bottom the plots show variant sets with 40 < MAC ≤ 100, 100 < MAC ≤ 200 and 200 < MAC ≤ 500. 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 α (=0.05/K), and the y-axis is the empirical false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FRP =1 (gray dashed line).
Figure S8. False positive rates for SKAT-O using ER-based and existing methods to compute p-values for variant sets with $40 < \text{MAC} \leq 500$.

From top to bottom the plots show variant sets with $40 < \text{MAC} \leq 100$, $100 < \text{MAC} \leq 200$ and $200 < \text{MAC} \leq 500$. 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 false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FPR =1 (gray dashed line).
Figure S9. False positive rates for SKAT using ER-based and existing methods to compute p-values for variant sets in the presence of population stratification when MAC ≤ 40

From top to bottom the plots show variant sets with MAC ≤ 10; 10 < MAC ≤ 20 and 20 < MAC ≤ 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 α (=0.05/K), and the y-axis is the empirical false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FPR =1 (gray dashed line).
Figure S10. False positive rates for Burden test using ER-based and existing methods to compute \( p \)-values for variant sets in the presence of population stratification when \( \text{MAC} \leq 40 \)

From top to bottom the plots show variant sets with \( \text{MAC} \leq 10; 10 < \text{MAC} \leq 20 \) and \( 20 < \text{MAC} \leq 40 \). 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 false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FPR =1 (gray dashed line).
Figure S11. False positive rates for SKAT-O using ER-based and existing methods to compute p-values for variant sets in the presence of population stratification when MAC ≤ 40

From top to bottom the plots show variant sets with MAC ≤ 10; 10 < MAC ≤ 20 and 20 < MAC ≤ 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 false positive rates (FPR) divided by the expected FPR. A well-calibrated test should have empirical/expected FPR = 1 (gray dashed line).
Figure S12. Empirical power of ER-hybrid, MA and UA methods for computing SKAT, Burden test and SKAT-O p-values for a causal variant set MAC= 20 with correction for 20,000 variant sets.

Power was estimated for variant sets in which all causal variants increased risk (top row) or 50% of variants increased risk and 50% decreased risk (bottom row). From left to the right, the plots consider case:control=1000:1000, 500:1500 and 200:1800. Empirical power was estimated for family wise error rate=0.05 using 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals. The experiment-wise permutation (ER-hybrid with EWP) and estimated number of variant sets (ER-hybrid with $K_{eff}$) were used to control family-wise error rate. For MA and UA, experiment-wise permutation was used to control family-wise error rate.
Figure S13. Empirical power of ER-hybrid, MA and UA methods for computing SKAT, Burden test and SKAT-O p-values for a causal variant set MAC= 40 with correction for 20,000 variant sets.

Power was estimated for variant sets in which all causal variants increased risk (top row) or 50% of variants increased risk and 50% decreased risk (bottom row). From left to the right, the plots consider case:control=1000:1000, 500:1500 and 200:1800. Empirical power was estimated for family-wise error rate=0.05 using 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals. The experiment-wise permutation (ER-hybrid with EWP) and estimated number of variant sets (ER-hybrid with $K_{eff}$) were used to control family-wise error rate. For MA and UA, experiment-wise permutation was used to control family-wise error rate.
Figure S14. QQ plots of ER-hybrid and unadjusted SKAT, Burden test and SKAT-O p-values of the analysis of GoT2D chromosome 2 exome data with disruptive+potentially damaging variants.

ER-hybrid (top panel) and unadjusted (bottom panel) methods were used to calculate SKAT, Burden test, and SKAT-O p-values. The x-axis in the top panel represents MAP-adjusted and unadjusted expected quantile of \(-\log_{10} p\)-values, and the x-axis in the bottom panel represents (unadjusted) expected quantiles of \(-\log_{10} p\)-values. The y-axis represents observed quantiles of \(-\log_{10} p\)-values. The dashed line represents a 95% confidence band based on 500 random draws from the MAP-based mixture distribution.
Figure S15. QQ plots of ER-hybrid and unadjusted SKAT, Burden test and SKAT-O p-values of the analysis of GoT2D chromosome 2 exome data with disruptive variants.

ER-hybrid (top panel) and unadjusted (bottom panel) methods were used to calculate SKAT, Burden test, and SKAT-O p-values. The x-axis in the top panel represents MAP-adjusted and unadjusted expected quantile of -log_{10} p-values, and the x-axis in the bottom panel represents (unadjusted) expected quantiles of -log_{10} p-values. The y-axis represents observed quantiles of -log_{10} p-values. The dashed line represents a 95% confidence band based on 500 random draws from the MAP-based mixture distribution.
Figure S16. Comparison of SKAT p-values obtained using ER or whole sample permutations for the GoT2D project dosage data
In the absence of covariates, SKAT p-values were obtained through ER or whole sample permutation (Perm) of disease status (left panel). In the presence of covariates SKAT p-value were obtained through ER or Fisher’s noncentral hypergeometric distribution based whole sample permutation (FNHPPerm) implemented in the BiasedUrn R-package (right panel). The x-axis represents –log₁₀ SKAT-ER p-values and y-axis represents –log₁₀ SKAT-Perm or SKAT-FNHPPerm p-values. 20000 variant sets were randomly selected, and 10⁵ resamples were generated to compute p-values for each method.
Table S1. Estimated computation time of PLINK/SEQ (ver 1.09) and SCORE-seq (ver 5.2) to carry out $10^7$ resampling to obtain a $p$-value for a single variant set. Each entry represents median of five replicates (cpu hours) to carry out whole sample permutation on a linux cluster node with a 2.9 GHz Xeon CPU. Since Fisher’s noncentral hypergeometric distribution based permutation is not implemented in PLINK/SEQ, PLINK/SEQ was only applied with no covariates. SCORE-Seq carried out Burden, SKAT, VT, EREC and WSS tests with covariates.

<table>
<thead>
<tr>
<th>Total Sample Size</th>
<th>PLINK/SEQ Calpha</th>
<th>No Covariates</th>
<th>PLINK/SEQ Burden</th>
<th>PLINK/SEQ SKAT</th>
<th>With Covariates</th>
<th>SCORE-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.1 h</td>
<td>15.6 h</td>
<td>23.9 h</td>
<td>4.2 h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>0.7 h</td>
<td>132.8 h</td>
<td>211.1 h</td>
<td>62.2 h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50000</td>
<td>3.6 h</td>
<td>&gt; 240 h</td>
<td>&gt; 240 h</td>
<td>&gt; 240 h</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S2. Estimated computation time of ER and QA. ER carried out $10^7$ resampling to obtain a $p$-value. Each entry in ER and QA columns represents a median of ten experiments. The number of variant loci was 30 when $m \geq 30$, otherwise it was the same as $m$. The last column (ER/QA) shows ER computation time divided by QA computation time.

<table>
<thead>
<tr>
<th>Test</th>
<th>m</th>
<th>ER</th>
<th>QA</th>
<th>ER/QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burden</td>
<td>10</td>
<td>&lt;.01</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>45</td>
<td>2.5</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>240</td>
<td>14</td>
<td>17.1</td>
</tr>
<tr>
<td>SKAT</td>
<td>10</td>
<td>.01</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>10</td>
<td>1.6</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>58</td>
<td>3.3</td>
<td>17.6</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>310</td>
<td>18</td>
<td>17.2</td>
</tr>
<tr>
<td>SKAT-O</td>
<td>10</td>
<td>.1</td>
<td>4.7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>56</td>
<td>4.7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>160</td>
<td>8.4</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>430</td>
<td>35.2</td>
<td>12</td>
</tr>
</tbody>
</table>