A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS

Rounak Dey,[1,2] Ellen M. Schmidt,[1,2] Goncalo R. Abecasis,[1,2] and Seunggeun Lee[1,2,*]

The availability of electronic health record (EHR)-based phenotypes allows for genome-wide association analyses in thousands of traits and has great potential to enable identification of genetic variants associated with clinical phenotypes. We can interpret the phenome-wide association study (PheWAS) result for a single genetic variant by observing its association across a landscape of phenotypes. Because a PheWAS can test thousands of binary phenotypes, and most of them have unbalanced or often extremely unbalanced case-control ratios (1:10 or 1:600, respectively), existing methods cannot provide an accurate and scalable way to test for associations. Here, we propose a computationally fast score-test-based method that estimates the distribution of the test statistic by using the saddlepoint approximation. Our method is much (~100 times) faster than the state-of-the-art Firth's test. It can also adjust for covariates and control type I error rates even when the case-control ratio is extremely unbalanced. Through application to PheWAS data from the Michigan Genomics Initiative, we show that the proposed method can control type I error rates while replicating previously known association signals even for traits with a very small number of cases and a large number of controls.

Introduction

Over the last decade, genome-wide association studies (GWASs) have proved instrumental to unravelling the genetic complexities of hundreds of diseases and traits and their associations with common genomic variations. To date, thousands of GWASs have identified more than 4,000 significant loci to be associated with human diseases and traits.[1] However, because most GWASs investigate a single disease or trait, they cannot exploit the cross-phenotype associations or pleiotropy,[2] where a single genetic variant can be associated with multiple phenotypes. The phenome-wide association study (PheWAS) has been proposed as an alternative approach to take advantage of the pleiotropy phenomenon by studying the impact of genetic variations across a broad spectrum of human phenotypes or "phenome." It is a complementary approach to the GWAS in the sense that whereas a GWAS attempts to identify phenotype-to-genotype associations, a PheWAS uses a genotype-to-phenotype approach. The first PheWAS[3] was published as a proof-of-principle study that demonstrated that the PheWAS strategy could be applied to successfully identify the expected gene-disease associations. Additional studies[4–8] have shown that the PheWAS approach can further identify previously unreported disease-SNP associations.[9]

The PheWAS approach depends on the availability of detailed phenotypic information. Currently, most PheWASs are applied to clinical cohorts linked to electronic health records (EHRs) and utilize the International Classification of Disease (ICD) billing codes to define clinical phenotypes. The ICD codes provide intuitive phenotype ordering based on clinical disease and trait classifications. Given that the current genotyping and imputation technologies allow for genotyping of tens of millions of variants at a very low cost,[10] an extensive PheWAS can attempt to investigate the genotype-phenotype associations by performing genome-wide association analyses in thousands of traits. We can interpret the PheWAS result of a single genetic variant by observing its associations across the phenome. Such a PheWAS is exhaustive in nature and has great potential to identify variants associated with clinical diseases.

One of the main challenges of the PheWAS approach is that most of the phenotypes are binary phenotypes with unbalanced (1:5) or often extremely unbalanced (1:600) case-control ratios (see Figure S1), given that the data are collected in cohorts. Although standard asymptotic tests (such as the Wald, score, and likelihood-ratio tests) are relatively well calibrated and asymptotically equivalent[11] for common (minor allele frequency [MAF] > 0.05) variants in balanced case-control studies, they can inflate type I error for low-frequency ($0.01 < \text{MAF} \leq 0.05$) and rare (MAF ≤ 0.01) variants in unbalanced case-control studies.[12] Moreover, because the Wald and likelihood-ratio tests need to calculate the likelihood or the maximum-likelihood estimator under the full model, which is computationally expensive, they are not scalable for the amount of tests that PheWASs attempt. On the other hand, the score test is computationally efficient because it does not need to calculate the maximum likelihood under the full model. However, as mentioned before, it suffers from having highly inflated type I error rates in unbalanced studies. Ma et al. proposed Firth's penalized likelihood-ratio test[13] as a solution to control the type I error rates in such situations. Firth's test, despite being well calibrated and robust for testing low-frequency and rare variants in unbalanced studies, lacks computational efficiency because it also involves calculating the maximum likelihood under the

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA
*Correspondence: leeshawn@umich.edu
http://dx.doi.org/10.1016/j.ajhg.2017.05.014.

full model. For instance, the projected computation time for testing 1,500 phenotypes across 10 million SNPs is ~117 CPU years (2,000 cases and 18,000 controls). Thus, it is impractical to apply Firth's test for analyzing large PheWAS datasets.

In this paper, we propose a score-based single-variant test for binary phenotypes that is well calibrated for controlling type I error and can adjust for covariates even in extremely unbalanced case-control studies. Moreover, our test is computationally efficient and scalable to testing thousands of phenotypes across millions of SNPs in large PheWAS datasets. Our proposed test (SPA) is based on score statistics and estimates the null distribution by using the saddlepoint approximation[14–16] instead of the normal approximation[17] traditionally used in score tests. We further develop an improvement of our test (fastSPA) that renders the most computationally challenging steps dependent only on the number of carriers (subjects with at least one minor allele) rather than the sample size. This improved test can substantially reduce the computation time, especially for low-frequency and rare variants, where the number of carriers is much lower than the sample size. Our method's projected computation time for testing 1,500 phenotypes across 10 million SNPs is ~400 CPU days (2,000 cases and 18,000 controls), which is more than a 100 times better than that of Firth's test. In addition, through extensive simulation studies and analysis of the Michigan Genomics Initiative (MGI) data, we demonstrate that the proposed approach can control type I error and is powerful enough to replicate known association signals.

Material and Methods

Logistic Regression Model and Saddlepoint Approximation Method

We consider a case-control study with sample size n. For the i^{th} subject, let $Y_i = 1$ or 0 denote the case-control status, X_i denote the $k \times 1$ vector of non-genetic covariates (including the intercept), and G_i denote the number of minor alleles ($G_i = 0, 1, 2$) of the variant to be tested. To relate genotypes to phenotypes, we use the following logistic regression model:

$$\text{logit}[\Pr(Y_i = 1 \mid X_i, G_i)] = X_i^T\beta + G_i\gamma \quad \text{for } i = 1, 2, \ldots, n, \qquad \text{(Equation 1)}$$

where β is a $k \times 1$ vector of coefficients of the covariates, and γ is the genotype log odds ratio. Under this model, we are interested in testing for the genetic association by testing the null hypothesis H_0: $\gamma = 0$. Let $\widehat{\mu}_i$ be the estimate of $\mu_i = \Pr(Y_i = 1|X_i)$, which is the probability of being a case under H_0. A score statistic for γ from the model (Equation 1) is given by $S = \sum_{i=1}^{n} G_i(Y_i - \widehat{\mu}_i)$. Suppose $X = (X_1^T, \ldots, X_n^T)$ is the $n \times k$ matrix of covariates, $G = (G_1, \ldots, G_n)^T$ is the genotype vector, W is a diagonal matrix with $\widehat{\mu}_i(1 - \widehat{\mu}_i)$ as the i^{th} diagonal element, and $\tilde{G} = G - X(X^TWX)^{-1}X^TWG$ is a covariate-adjusted genotype vector in which covariate effects are projected out from the genotypes (details are given in Appendix A). Then, S can be written as

$$S = \sum_{i=1}^{n} \tilde{G}_i(Y_i - \widehat{\mu}_i), \qquad \text{(Equation 2)}$$

and the mean and variance of S under H_0 are $E_{H_0}(S) = 0$ and $V_{H_0}(S) = \sum_{i=1}^{n} \tilde{G}_i^2 \widehat{\mu}_i(1 - \widehat{\mu}_i)$, respectively, where $\tilde{G}_i$ is the i^{th} element of $\tilde{G}$.

The traditional score test approximates the null distribution by using a normal distribution, which depends only on the mean and the variance of the score statistic. We can obtain the p value by comparing the observed test statistic (s) and $N(0, V_{H_0}(S))$. Normal approximation works well near the mean of the distribution but performs very poorly at the tails. The performance is especially poor when the underlying distribution is highly skewed, such as in unbalanced case-control outcomes,[12] because normal approximation cannot incorporate higher moments such as skewness. In addition, the convergence rate of normal approximation[18–20] is $O(n^{-1/2})$, which is not fast enough for rare variants.

Saddlepoint approximation was introduced by Daniels[14] as an improvement over the normal approximation. Contrary to normal approximation, where only the first two cumulants (mean and variance) are used for approximating the underlying distribution, saddlepoint approximation uses the entire cumulant-generating function (CGF). Jensen[21] further showed that saddlepoint approximation has a relative error bound of $O(n^{-3/2})$, making it a considerable improvement over the normal approximation.

To use the saddlepoint approximation, we first derive the CGF of S from the fact that $Y_i \sim \text{Bernoulli}(\mu_i)$ under H_0. Let $\widehat{\mu}$ be an $n \times 1$ vector with $\widehat{\mu}_i$ as the i^{th} element. From Equation 2, the estimate of the CGF of the score statistic S is

$$K(t) = \log(E_{H_0}(e^{tS})) = \sum_{i=1}^{n} \log\left(1 - \widehat{\mu}_i + \widehat{\mu}_i e^{\tilde{G}_i t}\right) - t\sum_{i=1}^{n} \tilde{G}_i\widehat{\mu}_i,$$

and the estimates of the first- and second-order derivatives of K are

$$K'(t) = \sum_{i=1}^{n} \frac{\widehat{\mu}_i\tilde{G}_i}{(1 - \widehat{\mu}_i)e^{-\tilde{G}_i t} + \widehat{\mu}_i} - \sum_{i=1}^{n} \tilde{G}_i\widehat{\mu}_i$$

and

$$K''(t) = \sum_{i=1}^{n} \frac{(1 - \widehat{\mu}_i)\widehat{\mu}_i\tilde{G}_i^2 e^{-\tilde{G}_i t}}{\left[(1 - \widehat{\mu}_i)e^{-\tilde{G}_i t} + \widehat{\mu}_i\right]^2},$$

respectively. We note that K, K', and K'' are plug-in estimates in which we plug in $\widehat{\mu}_i$ instead of μ_i. Then, according to the saddlepoint method (Barndorff-Nielson[15,16]), the distribution of S at s can be approximated by

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

where $w = \text{sgn}(\hat{t})\sqrt{2(\hat{t}s - K(\hat{t}))}$, $v = \hat{t}\sqrt{K''(\hat{t})}$, $\hat{t}$ is the solution to the equation $K'(\hat{t}) = s$, and Φ is the distribution function of a standard normal distribution.

Implementation Details and Approaches to Reducing Computation Time

The saddlepoint approximation method involves finding the root of the saddlepoint equation $K'(t) = s$. It is easy to verify that K' strictly increases as $K''(t) > 0$ for all $-\infty < t < \infty$, and $s = \sum_{i=1}^{n} \tilde{G}_i(Y_i - \widehat{\mu}_i)$ lies between $\lim_{t\to\infty} K'(t) = \sum_{i:\tilde{G}_i>0} \tilde{G}_i - \sum_{i=1}^{n} \tilde{G}_i\widehat{\mu}_i$ and $\lim_{t\to-\infty} K'(t) = \sum_{i:\tilde{G}_i<0} \tilde{G}_i - \sum_{i=1}^{n} \tilde{G}_i\widehat{\mu}_i$. Therefore, a unique root exists, and we can use popular root-finding algorithms

(Newton-Raphson,[22,23] bisection,[23] secant,[23] and Brent's method[24]) to efficiently solve this equation. For our simulation studies and real-data applications, we applied a combination of the Newton-Raphson and bisection method to solve the saddlepoint equations.

The most computationally demanding step in this saddlepoint approximation method is calculating the CGF and its derivatives. Here, we propose several approaches to reducing the computational complexities associated with these calculations.

Faster Calculation of the CGF by a Partially Normal Approximation Approach

The most computationally intensive step in the saddlepoint method is the calculation of the CGF K and its derivatives. In each step of the root-finding algorithm, we need to calculate K, K', and K'', each of which needs $O(n)$ computations. Using the fact that many elements of G are zeroes (i.e., homozygous major genotypes), we propose a fast computation method that speeds up the computation to $O(m)$, where m is the number of non-zero elements in G. Without loss of generality, we assume that the first m subjects have at least one minor allele each and the rest have homozygous major genotypes. We can then express S as $S = S_1 + S_2$, where $S_1 = \sum_{i=1}^{m} \tilde{G}_i(Y_i - \hat{\mu}_i)$ and $S_2 = \sum_{i=m+1}^{n} \tilde{G}_i(Y_i - \hat{\mu}_i)$. Let $Z = (X^T W X)^{-1} X^T W G$, and let Z_l be the l^{th} element of Z. Then, we can further express S_2 as

$$\begin{aligned} S_2 &= \sum_{i=m+1}^{n} \tilde{G}_i (Y_i - \hat{\mu}_i) = \sum_{i=m+1}^{n} (0 - X_i Z)(Y_i - \hat{\mu}_i) \\ &= -\sum_{i=m+1}^{n} \sum_{l=1}^{k} X_{il} Z_l (Y_i - \hat{\mu}_i) = -\sum_{l=1}^{k} Z_l \sum_{i=m+1}^{n} X_{il} (Y_i - \hat{\mu}_i) \\ &= -\sum_{l=1}^{k} Z_l S_{2l}, \end{aligned}$$

where $S_{2l} = \sum_{i=m+1}^{n} X_{il}(Y_i - \hat{\mu}_i)$. Now, if we assume that the non-genetic covariates are relatively balanced in the sample, then the normal distribution should be a good approximation of the null distribution of each S_{2l}. Because S_2 is a weighted sum of the S_{2l} variables, we can also approximate the null distribution of S_2 by using a normal distribution where the mean and variance under H_0 are given by $E_{H_0}(S_2) = 0$ and $V_{H_0}(S_2) = \sum_{i=m+1}^{n} \tilde{G}_i^2 \hat{\mu}_i(1 - \hat{\mu}_i)$, respectively. Then, the CGF of S_2 can be approximated by

$$K_2(t) = \frac{1}{2} t^2 V_{H_0}(S_2),$$

and the CGF of $S = S_1 + S_2$ can be approximated by

$$K(t) = \sum_{i=1}^{m} \log\left(1 - \hat{\mu}_i + \hat{\mu}_i e^{\tilde{G}_i t}\right) - t \sum_{i=1}^{m} \tilde{G}_i \hat{\mu}_i + \frac{1}{2} t^2 V_{H_0}(S). \qquad \text{(Equation 3)}$$

In order to calculate the first two terms on the right side of Equation 3, we will need $\tilde{G}_i$ values for $i = l, \ldots, m$, which can be calculated in $O(m)$ computations given that G has only m non-zero elements and the quantity $X(X^T W X)^{-1} X^T W$ can be pre-calculated. Then, the first two terms will require only $O(m)$ computations because both of them sum over m elements. Next, the variance $V_{H_0}(S)$ can be further broken down into

$$\begin{aligned} V_{H_0}(S) &= \sum_{i=m+1}^{n} \tilde{G}_i^2 \hat{\mu}_i(1 - \hat{\mu}_i) = \sum_{i=m+1}^{n} (X_i Z)^2 \hat{\mu}_i (1 - \hat{\mu}_i) \\ &= \sum_{i=1}^{n} (X_i Z)^2 \hat{\mu}_i(1 - \hat{\mu}_i) - \sum_{i=1}^{m} (X_i Z)^2 \hat{\mu}_i(1 - \hat{\mu}_i) \\ &= Z^T (X^T W X) Z - \sum_{i=1}^{m} (X_i Z)^2 \hat{\mu}_i(1 - \hat{\mu}_i). \end{aligned}$$

Because $X^T W X$ can be pre-calculated and Z is a $k \times 1$ vector, the first term requires $O(k)$ computations, and the second term requires $O(m)$ computations, which implies that the calculation of $V_{H_0}(S_2)$ requires $O(m)$ calculations under the assumption that $k < m$, i.e., the number of non-genetic covariates is smaller than the number of subjects with at least one minor allele each. Hence, the CGF $K(t)$ can be calculated in $O(m)$ computations. Using similar arguments, we can further show that the derivatives $K'(t)$ and $K''(t)$ can also be calculated in $O(m)$ computations. Therefore, this partially normal approximation reduces the computational complexity of our test from $O(n)$ to $O(m)$, which is especially useful for rare variants, where m is much smaller than n.

Using Normal Approximation near the Mean for Faster Computation

Because the normal approximation behaves well near the mean of the distribution, we can use it to obtain the p value when the observed score statistic (s) lies close to the mean (0). Moreover, saddlepoint approximation can be numerically unstable very close to the mean of the distribution. We can also avoid such situations by using normal approximation near the mean. One possible approach is to use a fixed threshold in which we apply normal approximation to obtain the p value if the absolute value of the observed score statistic, $|s| < r\sigma$, where $\sigma = \sqrt{V_{H_0}(S)}$ and r is a pre-specified value. For example, we used $r = 2$ in our simulation studies and real-data analyses. For a given level α, this approach does not inflate type I error rates if $r < \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the inverse function of the standard normal distribution function, $\Phi(x)$.

Alternatively, we can adaptively select the threshold by using the error bound of the normal approximation given by the Berry-Esseen theorem. Suppose we are interested in controlling the type I error rate at level α. Let $F_n(x)$ be the true distribution function of the standardized score test statistic $S/\sqrt{V_{H_0}(S)}$. Then, according to Berry-Esseen theorem,[18–20] the maximum error bound in approximating $F_n(x)$ by $\Phi(x)$ is

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \leq B_n = C(\sigma^2)^{-3/2} \left(\sum_{i=1}^{n} \rho_i \right), \qquad \text{(Equation 4)}$$

where $\rho_i = E_{H_0}[|\tilde{G}_i(Y_i - \hat{\mu}_i)|^3] = \tilde{G}_i^3 \hat{\mu}_i(1 - \hat{\mu}_i)[\hat{\mu}_i^2 + (1 - \hat{\mu}_i^2)]$ and C is a constant. As of now, the best-known estimate for C is 0.56, given by Shevtsova.[25] Suppose p_F and p_N are $F_n(x)$- and $\Phi(x)$-based p values, respectively. From the Berry-Esseen theorem, we can show $p_N \leq p_F + B_n$. Suppose $q = B_n + \alpha/2$ and $r_\alpha = \Phi^{-1}(1 - q)$. Then, $p_N \geq q$ indicates $p_F \geq \alpha/2$. Therefore, we use $r_\alpha \sigma$ as a threshold at level α in which we will apply normal approximation if $|s| < r_\alpha \sigma$.

Numerical Simulations

To evaluate the computation times, type I error rates, and power of the proposed method, we carried out extensive simulation studies. We considered three different case-control ratios: balanced with 10,000 cases and 10,000 controls, moderately unbalanced with 2,000 cases and 18,000 controls, and extremely unbalanced with 40 cases and 19,960 controls. For each choice of case-control ratio, the phenotypes were simulated on the basis of the following logistic model:

$$\text{logit}[\Pr(Y_i = 1)] = \beta_0 + X_{1i} + X_{2i} + \gamma G_i,$$

where the two non-genetic covariates X_{1i} and X_{2i} were simulated from $X_{1i} \sim$ Bernoulli (0.5) and $X_{2i} \sim N(0, 1)$, respectively. The intercept β_0 was chosen to correspond to a prevalence of 0.01. The

genotype G_i values were generated from a binomial(2, p) distribution where p was the MAF. The parameter γ represents the genotype log odds ratio.

To estimate computation times and type I error rates in realistic scenarios, we randomly sampled the MAF (p) from the MAF distribution in the MGI data. To compare computation times, we simulated 10^4 variants with $\gamma = 0$. To compare type I error rates, we simulated 10^9 variants with $\gamma = 0$ and recorded the number of rejections at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}. We also used fixed MAFs to evaluate the effect of MAF on computation time and type I error rates. For the power calculations, we considered two different choices for MAF (p = 0.01 and 0.05) and wide ranges of γ (Figure 4). For each choice of p and γ, we generated 5,000 variants.

We compared the computation times of seven different tests: a traditional score test using normal approximation (Score), the saddlepoint-approximation-based test with a standard-deviation threshold at 0.1 and 2 (SPA-0.1 and SPA-2, respectively), the fast saddlepoint-approximation-based test with the partially normal approximation improvement and a standard-deviation threshold at 0.1 and 2 (fastSPA-0.1 and fastSPA-2, respectively), the fastSPA test with the Berry-Esseen bound threshold at $\alpha = 5 \times 10^{-8}$ (fastSPA-BE), and Firth's penalized likelihood-ratio test. Next, we compared the empirical type I errors and power curves for fastSPA-2, Score, and Firth's test at 5×10^{-8}. Because performing Firth's test 10^9 times, which is required for estimating type I error rates at 5×10^{-8}, is practically impossible given the heavy computational burden, we performed a hybrid approach in which we used Firth's test only when the fastSPA-2 p values were smaller than 5×10^{-3}. For the power comparison, because Score has extremely inflated type I errors in the unbalanced and extremely unbalanced case-control scenarios (as shown in the Results), it might not be appropriate to directly compare the power of Score with that of the other two tests at the same nominal α level. In order to provide a more meaningful comparison, we compared their powers at the empirical α levels where their empirical type I errors became 5×10^{-8}. The empirical α levels were selected on the basis of the type I error simulations, whereby variants were simulated with MAF randomly sampled from the MAF distribution of the MGI data. This approach is similar to performing resampling (e.g., permutation) to control family-wise error rates. We also estimated the powers at the nominal fixed $\alpha = 5 \times 10^{-8}$. In order to compare the p values resulting from different tests, we also simulated 5×10^{-6} variants with MAFs randomly sampled from the MAF distribution of the MGI data. We further compared the inflation factors of the genomic controls at different p value quantiles for fastSPA-2, fastSPA-BE, and fastSPA-0.1 in order to explore the effect of the standard-deviation threshold on the inflation factor.

Application to MGI Data

To illustrate the performance of the proposed methods in real-data application, we analyzed four selected phenotypes in the MGI data. The main goal of MGI is to create an institutional repository of genetic data together with rich clinical phenotypes for a broad portfolio of future medical research. DNA from blood samples of >20,000 individuals who underwent surgical procedures at the University of Michigan Health System was genotyped (with their informed consent) on the Illumina HumanCoreExome v.12.1 array, which is a combined GWAS plus exome array composed of >500,000 SNPs. Genotypes of the Haplotype Reference Consortium[26] (chromosomes 1–22: HRC release 1; chromosome X: HRC release 1.1) were imputed into the phased MGI genotypes (SHAPEIT2[27] on autosomal chromosomes and Eagle2[28] on chromosome X) with Minimac3.[29] Excluding variants with low imputation quality ($R^2 < 0.3$) resulted in dense mapping at over 39 million quality-imputed genetic markers.

Phenotypes derived from 8,940 ICD-9 billing codes were classified into 1,815 PheWAS disease states of shared disease etiology, of which 1,448 had at least 20 cases. Standard code translations were used for converting the taxonomy of diagnostic ICD-9 codes into PheWAS code groups (PheWAS code translation table v.1.2[30]). Cases were derived from EHRs of individuals with at least two encounters with an ICD-9 billing code. This is a typical example of many recent large-scale PheWASs. To compare our proposed fastSPA-2 with Score and the current gold-standard Firth's test in analyzing such PheWAS data, we performed genome-wide association analyses for four selected traits—skin cancer (PheWAS code: 172), type 2 diabetes (PheWAS code: 250.2; MIM: 125853), primary hypercoagulable state (PheWAS code: 286.81; MIM: 188055), and cystic fibrosis (PheWAS code: 499; MIM: 219700)—in 18,267 unrelated individuals of European ancestry while adjusting for age, sex, and four principal components. Genotyped samples with any missing covariate information were excluded from the analysis. Given that imputation quality is low for very rare variants,[26] we excluded the imputed variants with MAF < 0.001 in our main analysis, which resulted in 13 million variants. For Firth's test, we used the hybrid approach used in the type I error simulation, where Firth's test was performed only when the fastSPA-2 p value was smaller than 5×10^{-3}.

Results

Numerical Simulations

We examine the computation time, type I error control, and power of the proposed fastSPA and two existing approaches, Score and Firth's test, across ranges of case-control imbalance and MAFs.

Comparison of Computation Times

The projected computation times for testing 1,500 phenotypes across 10 million variants by different testing methods are presented in Figure 1. To obtain computation time under realistic scenarios of the MAF distribution, we randomly sampled the MAFs of the simulated SNPs from the MAF spectrum of the MGI data (Figure S2). fastSPA-2 performs 100–300 times faster than Firth's test. In the unbalanced case-control setup of 2,000 cases and 18,000 controls, for example, Firth's test takes 117 CPU years to analyze 10 million SNPs across 1,500 phenotypes, whereas fastSPA-2 takes only 1.09 CPU years. This indicates that on a cluster with 100 CPU cores, the proposed test would require 4 days (without data reading), but Firth's test would need more than a year. When we compare fastSPA and SPA, fastSPA-0.1 performs 4–6 times faster than SPA-0.1 (e.g., 2.90 versus 12.32 CPU years when the case-control ratio = 2,000:18,000), and fastSPA-2 performs 1.5–2 times faster than SPA-2 (e.g., 1.09 versus 1.62 CPU years when the case-control

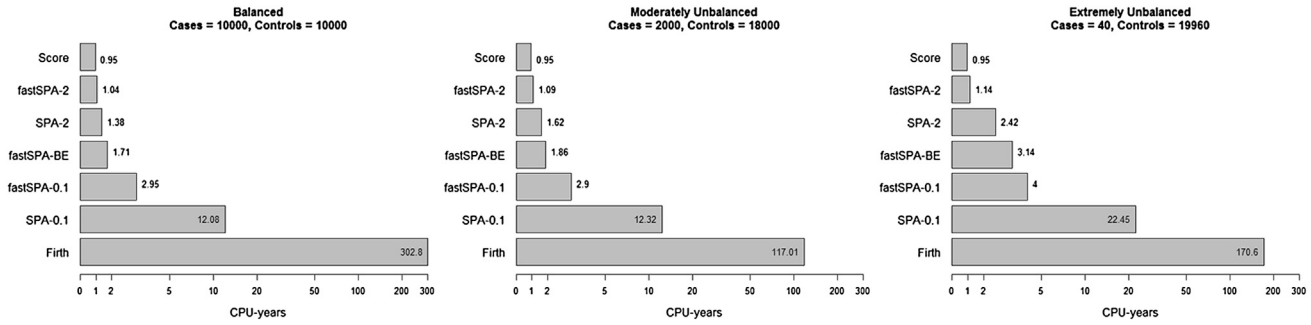

Figure 1. The Projected Computation Times for Testing 10 Million Variants across 1,500 Phenotypes by Various Tests with MAFs Sampled from the MAF Distribution of the MGI Data

The computation times are based on testing 10,000 simulated variants on an Intel i7 2.70GHz processor and then projecting them onto a PheWAS with 10 million variants and 1,500 phenotypes.

ratio = 2,000:18,000). Expectedly, the computation time for fastSPA-BE is in between the computation times for fastSPA-2 and fastSPA-0.1. fastSPA-BE performs 1.3–1.8 times faster than fastSPA-0.1 and 1.6–2.8 times slower than fastSPA-2 (e.g., 1.09, 1.86, and 2.9 CPU years for fastSPA-2, fastSPA-BE, and fastSPA-0.1, respectively, when the case-control ratio = 2,000:18,000).

We also recorded the computation times for variants with three different fixed MAFs (0.1, 0.01, and 0.001) in order to assess the effect of MAF on the performance of the tests. Similar to Figure 1, Table 1 shows the superior performance of fastSPA-2 over all other tests. Moreover, whereas the computation time of SPA increases with decreasing MAFs, which could be due to the slow convergence caused by the discrete nature of the underlying distribution, fastSPA requires less computation time for rarer variants (smaller MAFs) than for more common variants (larger MAFs). This demonstrates the potential of the partially normal approximation improvement in terms of faster computation of the p values, especially for low-frequency and rare variants.

Type I Error Comparison

The type I error rates from 10^9 simulated datasets are presented in Figure 2. Because of the heavy computation burden for testing these extremely large numbers of datasets, in this comparison, we considered only Score, fastSPA-2, and the hybrid version of Firth's test, in which we used Firth's test only when the fastSPA-2 p values were smaller than 5×10^{-3}. We note that both fastSPA-2 and Firth's test had well-calibrated quantile-quantile (Q-Q) plots up to 10^{-6} p values (Figure 5), and whenever fastSPA-2 p values were greater than 5×10^{-3}, Firth's test p values were greater than 4.8×10^{-4} (see p Value and Inflation Factor Comparison), indicating that the hybrid approach can provide very accurate estimation of the type I error rates of Firth's test at very stringent α levels.

Score had greatly inflated type I error rates for moderately unbalanced and extremely unbalanced case-control ratios, whereas fastSPA-2 could control the type I error in such situations. At the genome-wide significance level of $\alpha = 5 \times 10^{-8}$, for example, the empirical type I error rates of Score were 32 (1.63×10^{-6} when the case-control ratio = 2,000:18,000) and 26,600 (1.33×10^{-3} when the case-control ratio = 40:19,960) times higher than the nominal $\alpha = 5 \times 10^{-8}$. In contrast, fastSPA-2 had empirical type I error rates nearly identical to (4.9×10^{-8} when the case-control ratio = 2,000:18,000) or slightly lower than (3.5×10^{-8} when the case-control ratio = 40:19,960) the nominal $\alpha = 5 \times 10^{-8}$. Firth's test also had well-controlled

Table 1. Computation Times for Various Tests of 10,000 Simulated Variants with Different MAFs

Case-Control Ratio	MAF	Score	SPA-0.1	fastSPA-0.1	fastSPA-BE	SPA-2	fastSPA-2	Firth's Test
10,000:10,000	0.1	20	214	75	37	28	23	7,251
	0.01	19	225	38	35	27	20	6,918
	0.001	19	242	33	36	30	20	5,304
2,000:18,000	0.1	21	256	84	37	36	24	3,940
	0.01	20	284	39	36	35	21	4,312
	0.001	19	326	34	41	40	20	3,804
40:19,960	0.1	21	376	98	70	38	24	3,615
	0.01	20	477	42	58	44	21	3,598
	0.001	20	647	38	51	79	21	3,525

All computation times are in CPU seconds on an Intel i7 2.70GHz processor.

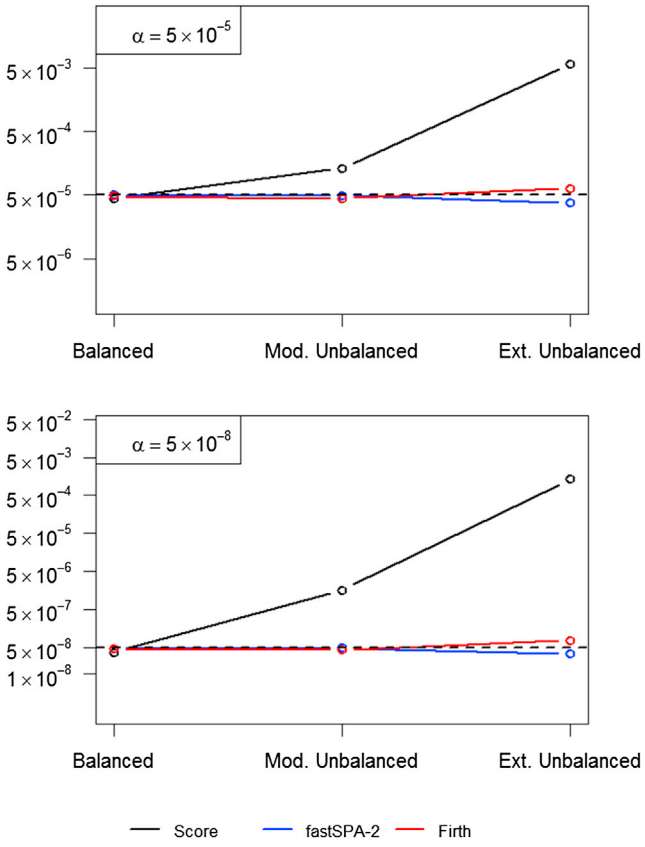

Figure 2. Type I Error Comparison between Score, fastSPA-2, and Firth's Test for Variants Simulated with MAFs Sampled from the MAF Distribution of the MGI Data

Type I error rates were estimated on the basis of 10^9 simulated datasets. From left to right on the x axis, the plots consider case-control ratios 10,000:10,000 (balanced), 2,000:18,000 (moderately unbalanced), and 40:19,960 (extremely unbalanced). The top and bottom panels show empirical type I error rates at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively.

type I error rates in the balanced and moderately unbalanced case-control scenarios (4.7×10^{-8} and 4.9×10^{-8}, respectively, at $\alpha = 5 \times 10^{-8}$). Interestingly, it showed slight inflation (7.8×10^{-8} at $\alpha = 5 \times 10^{-8}$) in the extremely unbalanced scenario. We also estimated empirical type I error rates at six different MAFs (Figure 3). Score had deflated type I error rates for low-frequency and rare variants for the balanced case-control ratio and inflated and extremely inflated type I error rates for moderately and severely unbalanced case-control ratios. fastSPA-2 had overall well-controlled type I error rates regardless of MAF and case-control ratio. Firth's test had either well-controlled or slightly conservative type I error rates when the case-control ratio was balanced or moderately unbalanced. However, when the case-control ratio was extremely unbalanced, Firth's test had inflated type I error rates, especially when the minor allele count was small (e.g., 1.33×10^{-7} and 1.47×10^{-7} for MAF = 0.0005 and 0.001, respectively, at $\alpha = 5 \times 10^{-8}$ when the case-control ratio = 40:19,960).

Power Comparison

Next, we compared the power curves of fastSPA-2, Score, and Firth's test. Note that Firth's test is a current gold-standard method.[13] Because Score had greatly inflated type I error rates, we compared the empirical powers of different tests at their test-specific empirical α levels. Figure 4 shows power by odds ratios when the MAF of the variant was 0.05 (top panel) and 0.01 (bottom panel). As expected, the power was higher when the case-control ratio was balanced. The empirical powers of fastSPA-2 and Firth's test were nearly identical for all case-control ratios and MAFs, which suggests that our proposed test does not suffer from any loss in power in comparison with Firth's test. The empirical powers of Score were almost identical to those of fastSPA-2 and Firth's test for the balanced case-control ratio. However, Score showed substantially lower power than the other two tests for the unbalanced case-control ratios as a result of the very small empirical α levels, and the power gap was especially large when the case-control ratio was extremely unbalanced. The simulation results clearly show that the proposed approach improves power over Score when type I error rates are properly controlled. When we used the nominal $\alpha = 5 \times 10^{-8}$ level instead of the empirical α levels, Score had higher power than the other two approaches (Figure S3) as expected, given that its type I error rates were not controlled.

p Value and Inflation Factor Comparison

To compare p value distributions of various tests, we generated Q-Q plots and calculated the inflation factor (λ) of the genomic control. Figure 5 suggests strong deflation (smaller than expected) in the p values from Score in the moderately unbalanced and extremely unbalanced case-control setups, whereas fastSPA-2, SPA-2, and Firth's test resulted in well-calibrated Q-Q plots, which suggests that these methods can control for type I errors. Moreover, the minimum Firth's test p value was 4.8×10^{-4} for the variants with a fastSPA-2 p value $> 5 \times 10^{-3}$ among all case-control setups, which justifies our hybrid approach of performing Firth's test only when the fastSPA-2 p value is less than 5×10^{-3} in the type I error simulation studies.

None of fastSPA-2, fastSPA-BE, and fastSPA-0.1 showed any inflation or deflation in genomic controls (λ) in the balanced and moderately unbalanced case-control setups (Table S1). In the extremely unbalanced case-control setup, fastSPA-2 resulted in a greatly deflated λ (0.48) at the median p value (q = 0.5). Interestingly, fastSPA-BE and fastSPA-0.1 resulted in an inflated λ (both 1.83) at q = 0.5, which could be due to the discrete nature of p values. However, when λ was measured at p value quantiles q = 0.01 and 0.001, all three tests provided λ very close to unity.

Analysis of MGI Data

We applied Score, Firth's test, and fastSPA-2 to the MGI data with four phenotypes: skin cancer, type 2 diabetes,

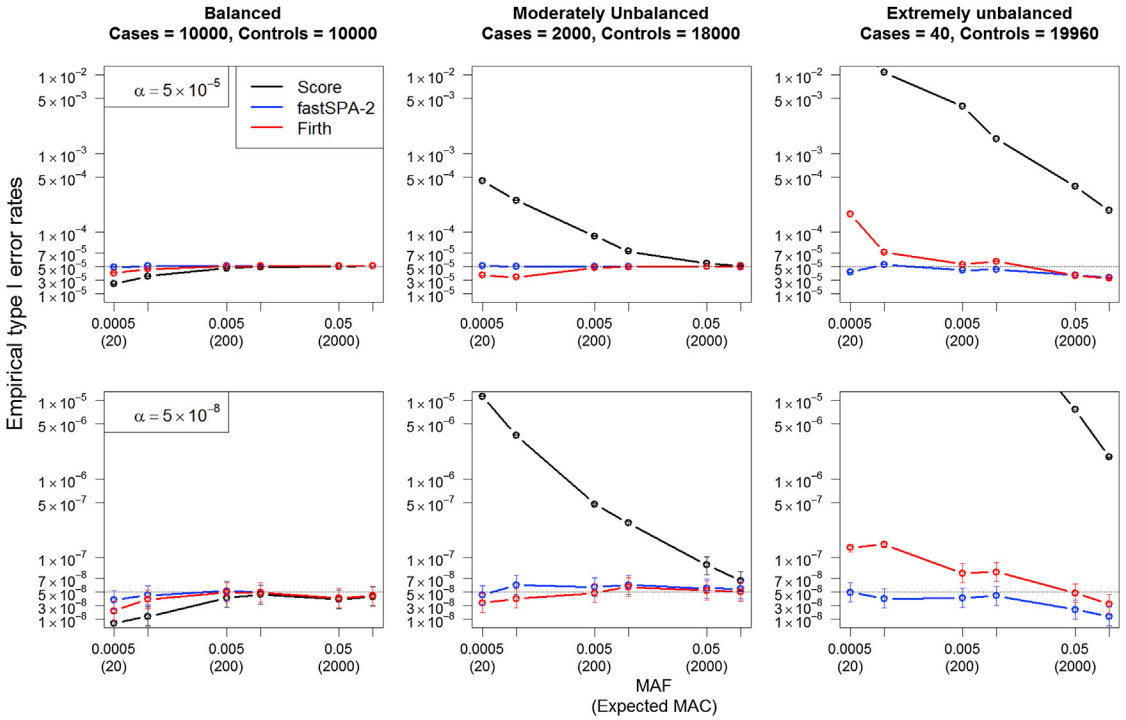

Figure 3. Type I Error Comparison at Different MAFs between Score, fastSPA-2, and Firth's Test
The top and bottom panels show empirical type I error rates at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. From left to right, the plots consider case-control ratios 10,000:10,000 (balanced), 2,000:18,000 (moderately unbalanced), and 40:19,960 (extremely unbalanced). In each plot, the x axis represents MAF with the expected minor allele count (MAC) in parentheses, and the y axis represents empirical type I error rates. Empirical type I error rates were estimated on the basis of 10^9 simulated datasets. 95% confidence intervals at different MAFs are also presented.

primary hypercoagulable state, and cystic fibrosis, which were selected on the basis of case-control ratios. Skin cancer (2,359 cases and 15,265 controls) and type 2 diabetes (1,987 cases and 14,906 controls) were moderately unbalanced, whereas primary hypercoagulable state (168 cases and 16,401 controls) and cystic fibrosis (28 cases and 18,212 controls) were extremely unbalanced phenotypes.

The Manhattan plots (Figure 6) show that Score produced a large number of potentially spurious associations for all of these phenotypes, whereas all of the significant variants from our proposed test at the genome-wide significance level of $\alpha = 5 \times 10^{-8}$ can be verified as truly associated with the phenotypes on the basis of previous findings (Table 2). In the analysis of skin cancer, variants in or near *IRF4* (MIM: 601900), *MC1R* (MIM: 155555), *RALY* (MIM: 614663), and *SLC45A2* (MIM: 606202) were significant at $\alpha = 5 \times 10^{-8}$, and all four of these genes were previously identified as associated with pigmentation traits and skin cancers.[31–36] In the other traits, variants in *TCF7L2* (MIM: 602228), *F5* (MIM: 612309), and *CFTR* (MIM: 602421) were significantly associated with type 2 diabetes,[37] primary hypercoagulable state,[38] and cystic fibrosis,[39] respectively, and all of these genes are well known to be associated with the risk of their respective diseases. The Q-Q plots (Figure 7) also suggest that the p values based on Score are much smaller than expected, especially for low-frequency and rare variants, whereas the p values based on fastSPA-2 closely follow the uniform distribution. We also observed the Manhattan plots (Figure S4) including the imputed variants with MAF $<$ 0.001 in the analysis. The inclusion of rarer variants resulted in extreme inflation in the number of potentially spurious associations for Score. However, our proposed test still produced none to very few new associations. The Manhattan plots and Q-Q plots for Firth's test were almost identical to those of our proposed test.

Further, on the basis of the p values from our proposed test, we obtained the inflation factor (λ) of the genomic control at different p value quantiles (q) and different MAF cutoffs (Table S2). Only the imputed variants were removed when we used different MAF cutoffs. The SNPs present on the Illumina HumanCoreExome v.12.1 array were preserved. To evaluate whether using a smaller standard-deviation threshold (r) improves the estimation of λ, we also applied fastSPA with r = 0.1 (fastSPA-0.1) and fastSPA with the Berry-Esseen bound threshold at $\alpha = 5 \times 10^{-8}$ (fastSPA-BE) on these four phenotypes. When all variants were included in the analysis, there was slight inflation ($\lambda = 1.11$, type 2 diabetes) or great deflation ($\lambda = 0.12$, cystic fibrosis) at the median level for fastSPA-2.

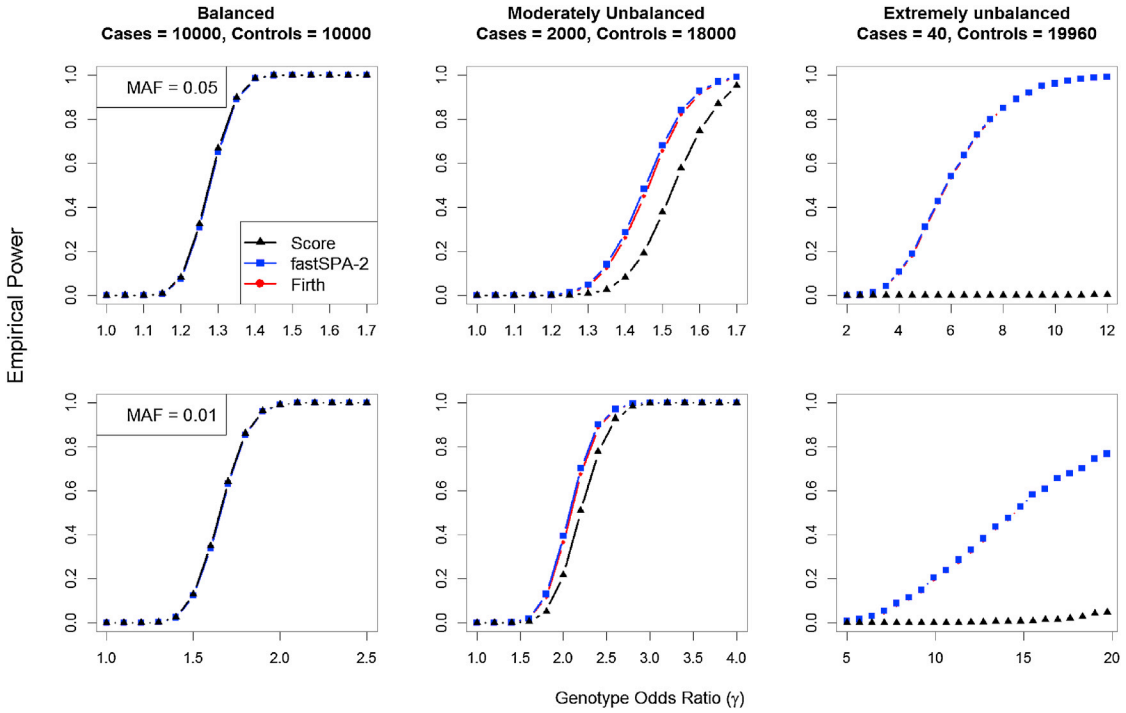

Figure 4. Empirical Power Curves for Score, fastSPA-2, and Firth's Test
The top and bottom panels consider MAF = 0.05 and 0.01, respectively. From left to right, the plots consider case-control ratios 10,000:10,000 (balanced), 2,000:18,000 (moderately unbalanced), and 40:19,960 (extremely unbalanced). In each plot, the x axis represents genotype odds ratios, and the y axis represents the empirical power. Empirical power was estimated from 5,000 simulated datasets at the test-specific α levels where their empirical type I errors were equal to 5×10^{-8}.

However, the genomic controls were very close to unity at q = 0.01 and 0.001. When we considered only the variants with MAF > 0.001, fastSPA-2 did not show any significant inflation in λ at the median for skin cancer, type 2 diabetes, or primary hypercoagulable state. However, it showed deflated genomic control for cystic fibrosis (λ = 0.63) as a result of the discrete nature of the underlying distribution. However, when we excluded the rare variants and considered only the variants with MAF > 0.01, all four of the phenotypes showed λ very close to unity. Neither fastSPA-0.1 nor fastSPA-BE showed a significant inflation or deflation in λ at any quantiles or MAF cutoffs, except for cystic fibrosis (both with λ = 1.27) when all variants were considered and genomic control was measured at the median level.

Discussion

In this paper, we propose a fast and scalable test for analyzing large PheWAS datasets that is well calibrated even in extremely unbalanced case-control settings. The method uses computationally efficient saddlepoint approximation to accurately calculate p values of score test statistics. We further propose an improved version of our test that substantially reduces the computation time, especially for low-frequency and rare variants. Our proposed test can also adjust for additional covariates. Through extensive numerical studies, we have demonstrated that our test can perform 100–300 times faster than the currently used Firth's test while retaining similar power and well-controlled type I error rates. Analysis of MGI data illustrates that by applying the proposed method to PheWAS datasets, we can identify true association signals while controlling for type I error, even for traits with a very small number of cases and a large number of controls.

Our test calculates p values on the basis of Score if the score statistics lie sufficiently close to the mean. Even though normal approximation is accurate near the mean, those p values might not be well calibrated. In such cases, because the median p values might come from Score, we can encounter a slightly inflated or deflated inflation factor at the median. When the case-control ratio is extremely unbalanced, this phenomenon is more pronounced. One way to circumvent this issue is to measure the inflation factor at more extreme quantiles (0.01, 0.001, etc.) or to exclude rare variants when estimating the inflation factor. Another approach is to decrease the standard-deviation threshold so that the median p values come from the saddlepoint approximation. In the analysis of MGI data, fastSPA-0.1 produced substantially improved inflation-factor estimates than fastSPA-2. However, the use of threshold 0.1 instead of 2 would increase the computation time

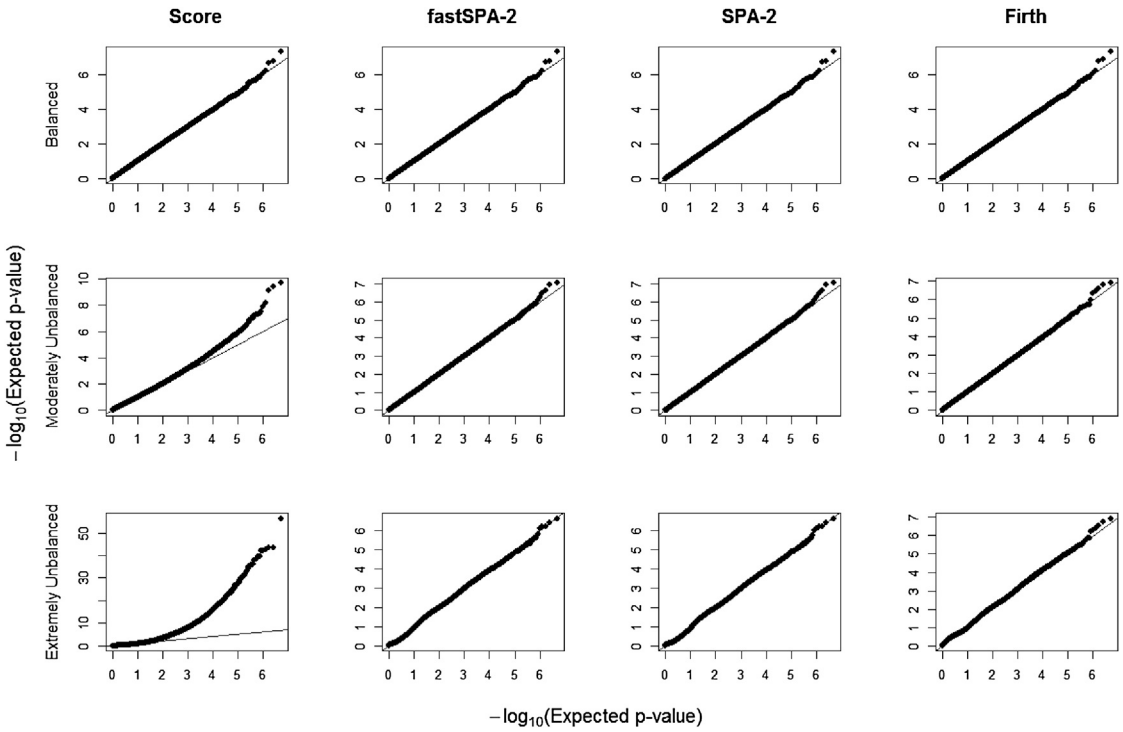

Figure 5. Q-Q Plots for Score, fastSPA-2, SPA-2, and Firth's Test on 5×10^6 Simulated Variants with MAF Randomly Sampled from the MAF Distribution of the MGI Data

The top, middle, and bottom panels show Q-Q plots in the balanced (case-control ratio = 10,000:10,000), moderately unbalanced (case-control ratio = 2,000:18,000), and extremely unbalanced (case-control ratio = 40:19,960) case-control scenarios, respectively. In each plot, the x axis represents $-\log_{10}$ expected p values, and the y axis represents $-\log_{10}$ observed p values.

from ~3 to 4 times. The Berry-Esseen threshold can be viewed as a compromise between these two thresholds. If there is no restriction in computational resource, we recommend using fastSPA-0.1 so that most of the p values are calculated by the saddlepoint approximation. If computational resource is limited, or researchers want to obtain results quickly, either a larger threshold (i.e., fastSPA-2) or Berry-Esseen bound can be a better choice.

As sequencing costs continue to drop, whole-exome or whole-genome sequencing will be used for PheWASs to identify rare variants associated with clinical phenotypes.[40] In rare-variant association analysis, gene- or region-based multiple-variant tests are commonly used to improve power.[41] When case-control ratios are unbalanced, popular rare-variant tests, including burden tests, SKAT, and SKAT-O, can also have substantially inflated type I error rates. Although resampling-based approaches have been developed to address this problem,[42] the existing methods are not fast enough to be used in PheWASs. One possible approach is to first adjust single-variant score statistics by SPA and then use the adjusted score statistics to control for the type I error. We have left this for future research.

In summary, we have proposed an accurate and scalable method for PheWAS data analysis. With the growing effort to build large research cohorts for precision medicine,[40] future PheWASs will have hundreds of thousands of samples and hundreds of millions of variants. Our method will provide a scalable solution for this large-scale problem and contribute to finding genetic components of complex traits. All of our tests are implemented in the R package SPAtest.

Appendix A: Explanation behind Using $\tilde{G}$ instead of G

We first note that $S = \tilde{G}^T(Y - \hat{\mu}) = G^T(Y - \hat{\mu})$ given that $\hat{\mu}$ is the maximum-likelihood estimator of μ under the null model and $X^T(Y - \hat{\mu}) = 0$. Now, the score function and the observed information matrix under the null model are given by

$$U_0 = \begin{bmatrix} X^T(Y - \hat{\mu}) \\ G^T(Y - \hat{\mu}) \end{bmatrix} = \begin{bmatrix} 0 \\ S \end{bmatrix}$$

and

$$I_0 = \begin{bmatrix} X^T W X & X^T W G \\ G^T W X & G^T W G \end{bmatrix},$$

respectively.

Therefore, the variance of S under H_0 is given by

$$V_{H_0}(S) = G^T W G - G^T W X (X^T W X)^{-1} X^T W G = G^T W \tilde{G} = \tilde{G}^T W \tilde{G}.$$

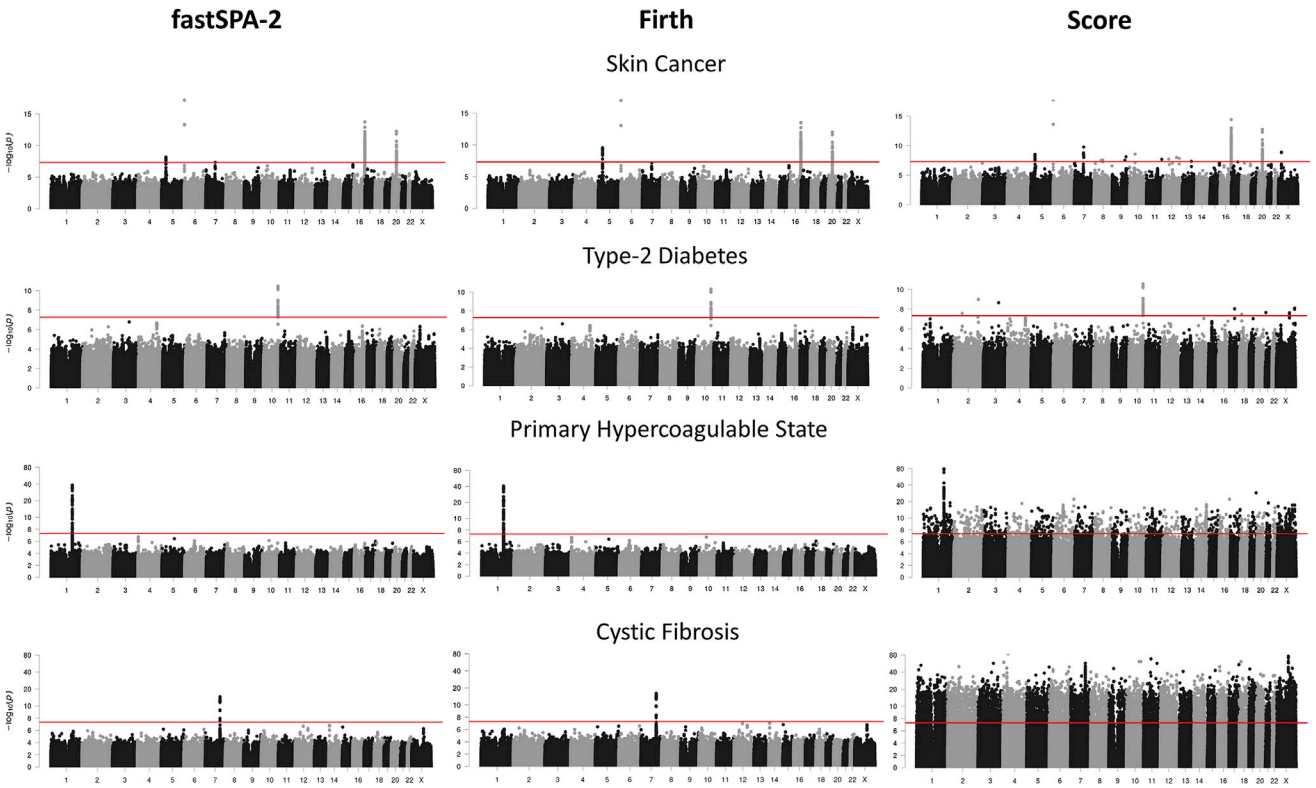

Figure 6. Manhattan Plots for Four Different Phenotypes from MGI Data
All imputed variants with MAF > 0.001 and all directly genotyped variants were included in this analysis. From left to right, the three panels show associations based on fastSPA-2, Firth's test, and Score. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

So, even though the two expressions of S are algebraically the same, the variance can be expressed as a weighted sum of $\hat{\mu}_i(1-\hat{\mu}_i)$ values, where the weights are given by $\tilde{G}_i$ values. Therefore, we used $\tilde{G}$ instead of G to express the score statistic.

Supplemental Data

Supplemental Data include four figures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.05.014.

Acknowledgments

This work was supported by NIH grant R01 HG008773 (R.D. and S.L.). We would like to thank the investigators of the Michigan Genomics Initiative project for access to the PheWAS dataset and Dr. Hyun Min Kang for implementing the methods in the Epacts package.

Received: February 16, 2017
Accepted: May 17, 2017
Published: June 8, 2017

Table 2. Significant SNP-Phenotype Associations Based on fastSPA-2 on MGI Data and Previous Findings Confirming Such Associations

Phenotype	Location	dbSNP ID	Nearest Gene	Alleles	MAF	p Value	Previous Findings
Skin cancer	6:396321	rs12203592	*IRF4*	C>T	0.16	6.71×10^{-18}	Zhang et al.[31] Sulem et al.,[32] Jacobs et al.,[33] and Liu et al.[34]
	16:89986117	rs1805007	*MC1R*	C>T	0.077	1.86×10^{-14}	Zhang et al.[31] Sulem et al.,[32] Jacobs et al.,[33] and Liu et al.[34]
	20:32538391	rs62211989	*RALY*	G>C	0.075	5.59×10^{-13}	Zhang et al.[31] Sulem et al.,[32] Jacobs et al.,[33] and Liu et al.[34]
	5:33951693	rs16891982	*SLC45A2*	C>G	0.038	7×10^{-9}	Liu et al.,[34] Barrett et al.,[35] and Nan et al.[36]
Type 2 diabetes	10:114754071	rs34872471	*TCF7L2*	T>C	0.29	3.4×10^{-11}	Scott et al.[37]
Primary hypercoagulable state	1:169519049	rs6025	*F5*	T>C	0.029	4.9×10^{-39}	Bertina et al.[38]
Cystic fibrosis	7:117299434	rs113827944	*CFTR*	G>A	0.018	3.11×10^{-15}	Kerem et al.[39]

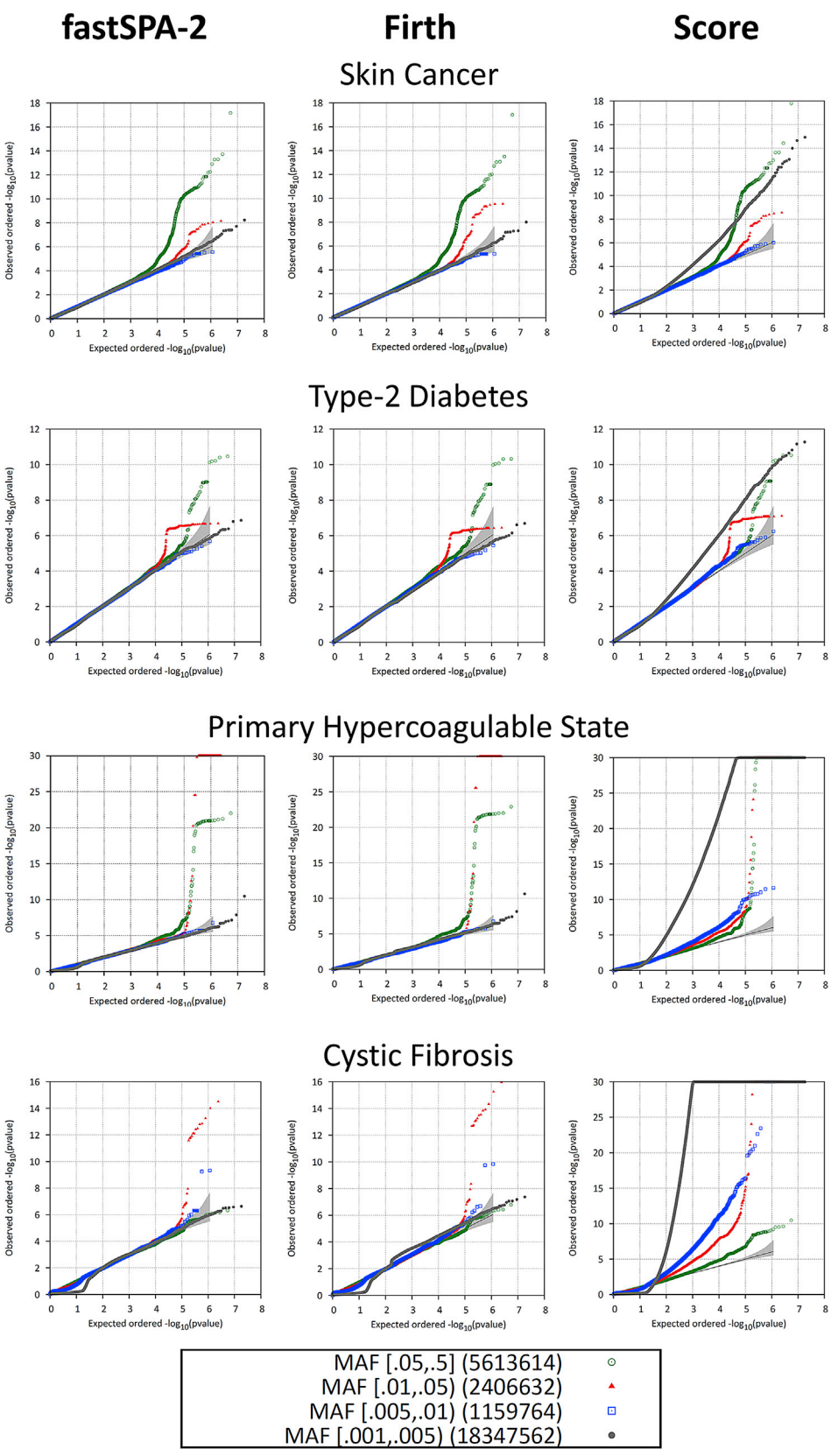

Figure 7. Q-Q Plots for Four Different Phenotypes from MGI Data
From left to right, the three panels show the Q-Q plots based on fastSPA-2, Firth's test, and Score. The plots are color coded according to different MAF categories. 95% confidence bands are presented in gray to signify the deviance from the uniform distribution.

Web Resources

Michigan Genomics Initiative, https://www.michigangenomics.org/
OMIM, http://www.omim.org
SPAtest R-package, https://sites.google.com/a/umich.edu/leeshawn/software

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
2. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. *14*, 483–495.
3. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210.
4. Denny, J.C., Crawford, D.C., Ritchie, M.D., Bielinski, S.J., Basford, M.A., Bradford, Y., Chai, H.S., Bastarache, L., Zuvich, R., Peissig, P., et al. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am. J. Hum. Genet. *89*, 529–542.
5. Hebbring, S.J., Schrodi, S.J., Ye, Z., Zhou, Z., Page, D., and Brilliant, M.H. (2013). A PheWAS approach in studying HLA-DRB1*1501. Genes Immun. *14*, 187–191.
6. Ritchie, M.D., Denny, J.C., Zuvich, R.L., Crawford, D.C., Schildcrout, J.S., Bastarache, L., Ramirez, A.H., Mosley, J.D., Pulley, J.M., Basford, M.A., et al.; Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) QRS Group (2013). Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. Circulation *127*, 1377–1385.
7. Pendergrass, S.A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E.S., Goodloe, R., Ambite, J.L., Avery, C.L., Buyske, S., Bůžková, P., et al. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet. *9*, e1003087.
8. Shameer, K., Denny, J.C., Ding, K., Jouni, H., Crosslin, D.R., de Andrade, M., Chute, C.G., Peissig, P., Pacheco, J.A., Li, R., et al. (2014). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Hum. Genet. *133*, 95–109.
9. Hebbring, S.J. (2014). The challenges, advantages and future of phenome-wide association studies. Immunology *141*, 157–165.
10. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. *11*, 499–511.
11. Cox, D., and Hinkley, D. (1974). Theoretical Statistics (Chapman and Hall).
12. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J.; and GoT2D investigators (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet. Epidemiol. *37*, 539–550.
13. Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika *80*, 27–38.
14. Daniels, H.E. (1954). Saddlepoint Approximations in Statistics. Ann. Math. Stat. *25*, 631–650.
15. Barndorff-Nielsen, O.E. (1990). Approximate Interval Probabilities. J. R. Stat. Soc. B *52*, 485–496.
16. Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. Biometrika *86*, 929–935.
17. Feller, W. (1945). The fundamental limit theorems in probability. Bull. Amer. Math. Soc. *51*, 800–832.
18. Berry, A.C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. Trans. Am. Math. Soc. *49*, 122–136.
19. Esseen, C.G. (1942). On the Liapounoff Limit of Error in the Theory of Probability. Ark. Mat. Astr. Fys. *28A*, 1–19.
20. Esseen, C.G. (1956). A Moment Inequality with an Application to the Central Limit Theorem. Skand Aktuarietidskr *39*, 160–170.
21. Jensen, J.L. (1995). Saddlepoint Approximations (Oxford University Press).
22. Whittaker, E.T., and Robinson, G. (1967). The Newton-Raphson Method. In The Calculus of Observations: A Treatise on Numerical Mathematics, Fourth Edition (Dover), pp. 84–87.
23. Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1992). Numerical Recipes in Fortran 77: The Art of Scientific Computing, Second Edition (Cambridge University Press).
24. Brent, R.P. (1973). Algorithms for Minimization without Derivatives (Prentice-Hall).
25. Shevtsova, I.G. (2010). An improvement of convergence rate estimates in the Lyapunov theorem. Dokl. Math. *82*, 862–864.
26. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
27. Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.
28. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.
29. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
30. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics *30*, 2375–2376.
31. Zhang, M., Song, F., Liang, L., Nan, H., Zhang, J., Liu, H., Wang, L.-E., Wei, Q., Lee, J.E., Amos, C.I., et al. (2013). Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. Hum. Mol. Genet. *22*, 2948–2959.
32. Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Magnusson, K.P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., et al. (2007). Genetic determinants of

hair, eye and skin pigmentation in Europeans. Nat. Genet. *39*, 1443–1452.

33. Jacobs, L.C., Hamer, M.A., Gunn, D.A., Deelen, J., Lall, J.S., van Heemst, D., Uh, H.-W., Hofman, A., Uitterlinden, A.G., Griffiths, C.E.M., et al. (2015). A Genome-Wide Association Study Identifies the Skin Color Genes IRF4, MC1R, ASIP, and BNC2 Influencing Facial Pigmented Spots. J. Invest. Dermatol. *135*, 1735–1742.
34. Liu, F., Visser, M., Duffy, D.L., Hysi, P.G., Jacobs, L.C., Lao, O., Zhong, K., Walsh, S., Chaitanya, L., Wollstein, A., et al. (2015). Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. Hum. Genet. *134*, 823–835.
35. Barrett, J.H., Iles, M.M., Harland, M., Taylor, J.C., Aitken, J.F., Andresen, P.A., Akslen, L.A., Armstrong, B.K., Avril, M.-F., Azizi, E., et al.; GenoMEL Consortium (2011). Genome-wide association study identifies three new melanoma susceptibility loci. Nat. Genet. *43*, 1108–1113.
36. Nan, H., Kraft, P., Qureshi, A.A., Guo, Q., Chen, C., Hankinson, S.E., Hu, F.B., Thomas, G., Hoover, R.N., Chanock, S., et al. (2009). Genome-wide association study of tanning phenotype in a population of European ancestry. J. Invest. Dermatol. *129*, 2250–2257.
37. Scott, L.J., Bonnycastle, L.L., Willer, C.J., Sprau, A.G., Jackson, A.U., Narisu, N., Duren, W.L., Chines, P.S., Stringham, H.M., Erdos, M.R., et al. (2006). Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample. Diabetes *55*, 2649–2653.
38. Bertina, R.M., Koeleman, B.P., Koster, T., Rosendaal, F.R., Dirven, R.J., de Ronde, H., van der Velden, P.A., and Reitsma, P.H. (1994). Mutation in blood coagulation factor V associated with resistance to activated protein C. Nature *369*, 64–67.
39. Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., and Tsui, L.C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science *245*, 1073–1080.
40. Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. N. Engl. J. Med. *372*, 793–795.
41. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. *95*, 5–23.
42. Lee, S., Fuchsberger, C., Kim, S., and Scott, L. (2016). An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. Biostatistics *17*, 1–15.