RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Evaluating the Calibration and Power of Three Gene-Based Association Tests of Rare Variants for the X Chromosome

Clement Ma, Michael Boehnke, Seunggeun Lee,* and the GoT2D Investigators

Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America

Received 13 April 2015; Revised 11 August 2015; accepted revised manuscript 2 September 2015.
Published online 10 October 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21935

ABSTRACT: Although genome-wide association studies (GWAS) have identified thousands of trait-associated genetic variants, there are relatively few findings on the X chromosome. For analysis of low-frequency variants (minor allele frequency <5%), investigators can use region- or gene-based tests where multiple variants are analyzed jointly to increase power. To date, there are no gene-based tests designed for association testing of low-frequency variants on the X chromosome. Here we propose three gene-based tests for the X chromosome: burden, sequence kernel association test (SKAT), and optimal unified SKAT (SKAT-O). Using simulated case-control and quantitative trait (QT) data, we evaluate the calibration and power of these tests as a function of (1) male:female sample size ratio; and (2) coding of haploid male genotypes for variants under X-inactivation. For case-control studies, all three tests are reasonably well-calibrated for all scenarios we evaluated. As expected, power for gene-based tests depends on the underlying genetic architecture of the genomic region analyzed. Studies with more (haploid) males are generally less powerful due to decreased number of chromosomes. Power generally is slightly greater when the coding scheme for male genotypes matches the true underlying model, but the power loss for misspecifying the (generally unknown) model is small. For QT studies, type I error and power results largely mirror those for binary traits. We demonstrate the use of these three gene-based tests for X-chromosome association analysis in simulated data and sequencing data from the Genetics of Type 2 Diabetes (GoT2D) study.
Genet Epidemiol 39:499–508, 2015.

KEY WORDS: rare variants; low-frequency variants; gene-based association tests; genome-wide association study

Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with hundreds of diseases and traits [Hindorff et al., 2012]. However, the proportion of associated variants on the X chromosome, relative to its chromosomal length, lags far behind those on the autosomes [Wise et al., 2013]. Analysis of X-chromosome association requires proper treatment of diploid female and haploid male participants. Although we can code female genotypes under an additive model as the number of minor alleles for a specific variant: $g = \{0,1,2\}$, just as we do for autosomal variants, for male genotypes, there are two obvious coding schemes. For a variant under X-inactivation [Lyon, 1961], where one copy of the female X chromosome is inactivated, one copy of the male allele is equivalent to two copies of the female allele, and hence we code haploid male genotypes as $g = \{0,2\}$. For a variant at a locus that does not undergo X-inactivation, we code male genotypes as $g = \{0,1\}$. For analysis of a mixed sample of males and females, specialized analytical tools are needed for initial data processing (e.g., estimating allele frequencies and testing Hardy-Weinberg equilibrium) [Purcell et al., 2007], genotype imputation [Howie et al., 2012; Marchini et al., 2007], and association analysis [Clayton, 2008; Zheng et al., 2007]. Hence, in many GWAS, the analysis of the X chromosome has been omitted due to the additional analysis steps required and/or lack of available software tools [Wise et al., 2013]. With use of specialized analytical tools, additional trait-associated variants on the X chromosome are likely to be identified.

Existing X-chromosome analysis methods focus on single-marker association analysis. Zheng et al. [2007] proposed tests comparing differences in allele counts between cases and controls for males and females jointly, and assume no X-inactivation (coding male genotypes as $g = \{0,1\}$). Clayton [2008] proposed score tests for the additive and dominant genetic models assuming X-inactivation (coding male genotypes as $g = \{0,2\}$). His test assumes equal allele frequencies in males and females; if this assumption is violated, he recommended stratifying by sex and combining score statistics across strata. Loley et al. [2011] evaluated the calibration and power of these tests and showed that no single test is uniformly most powerful over all genetic models and simulation parameters. Loley demonstrated that Clayton's nonsex-stratified tests can be anti-conservative when allele frequencies differ between the sexes. Hickey and Bahlo

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Seunggeun Lee, Department of Biostatistics and Center for Statistical Genetics, University of Michigan, M4148 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: leeshawn@umich.edu

[2011] conducted a similar evaluation, and showed that tests that made use of both male and female data were uniformly more powerful than tests that only use female data.

Many recent genetic studies use genome or exome sequencing [Steinthorsdottir et al., 2014] or specialized genotyping arrays [Huyghe et al., 2013] to better assay low-frequency genetic variants (minor allele frequency [MAF] < 5%). Single-marker tests have low power to test for association with low-frequency variants unless the sample and/or effect size is very large [Asimit and Zeggini, 2010]. In contrast, region- or gene-based tests in which multiple markers are analyzed jointly can be more powerful for analyzing low-frequency variants [Lee et al., 2014]. The calibration and power of gene-based tests have not been evaluated in the context of analyzing low-frequency variants on the X chromosome. In this paper, we describe, apply, and evaluate three gene-based tests for the X chromosome: burden, sequence kernel association test (SKAT), and optimal unified SKAT (SKAT-O) [Lee et al., 2012]. Specifically, using simulated binary and quantitative trait (QT) datasets, we evaluate the calibration and power of these tests with (1) different male:female ratios in cases and controls, and (2) different coding of male genotypes.

We find that for case-control studies, all tests are well-calibrated or very slightly anti-conservative for different male:female ratios in cases and controls, and different coding of male genotypes. As expected, power depends on the underlying genetic architecture of the genomic region analyzed. Studies with more males than females are typically less powerful. In most scenarios, power is slightly greater when we analyze data assuming the true model to code male genotypes. For QT studies, burden and SKAT are well-calibrated, while SKAT-O can be slightly anti-conservative, and power results are similar to those for binary traits. We conclude that these gene-based tests can be directly applied to the association analysis of low-frequency variants for both binary and QTs. We implemented these tests in the SKAT R package [Lee, 2014].

Methods

Notation

Consider n individuals sequenced at m variants in a genomic region of interest. For individual i, let $\mathbf{X_i} = (x_{i1}, \ldots, x_{is})'$ be the vector of s covariates (including a covariate for sex) and $\mathbf{G_i} = (g_{i1}, \ldots, g_{im})'$ be the vector of genotypes. For (diploid) females, let $g_{ij} = \{0,1,2\}$ be the number of minor alleles for variant j. For (haploid) males, we consider two coding schemes: (1) $g_{ij} = \{0,2\}$ when assuming X-inactivation in the females and (2) $g_{ij} = \{0,1\}$ when assuming no X-inactivation. For binary traits, $y_i = 1$ or $y_i = 0$ denotes a case or control, respectively; for QTs, y_i denotes the QT value. In a combined sample of n_m males and n_f females (all unrelated), the maximum-likelihood estimate of the MAF p of a biallelic variant with alleles A and a is

$$\hat{p} = \frac{2n_{aa,f} + n_{Aa,f} + n_{a,m}}{2n_f + n_m},$$

where $n_{aa,f}$ and $n_{Aa,f}$ are the number of females with genotypes aa and Aa, and $n_{a,m}$ is the number of males with the a allele.

Gene-Based Tests

For binary traits, we consider the logistic regression model:

$$\text{logit}(\Pr(y_i = 1)) = \text{logit}(\pi_i) = \gamma_0 + \mathbf{X}'_\mathbf{i}\gamma_\mathbf{1} + \mathbf{G}'_\mathbf{i}\beta, \qquad (1)$$

where γ_0 is the intercept, γ_1 is the $s \times 1$ vector of regression coefficients for the covariates, and $\beta = (\beta_1, \ldots, \beta_m)'$ is the $m \times 1$ vector of regression coefficients for the genetic variants. For QTs, the linear regression model is

$$y_i = \gamma_0 + \mathbf{X}'_\mathbf{i}\gamma_\mathbf{1} + \mathbf{G}'_\mathbf{i}\beta + \varepsilon_i, \qquad (2)$$

where ε_i is the normally distributed error term with mean zero and variance σ^2.

Because there is limited power to test the null hypothesis that the vector $\boldsymbol{\beta} = 0$ for large m, the burden test combines the genetic effects over the genomic region by assuming $\beta_j = w_j\beta_c$, given weights w_j. Thus, Equations (1) and (2) become

$$\text{logit}(\pi_i) = \gamma_0 + \mathbf{X}'_\mathbf{i}\gamma_\mathbf{1} + \beta_c\left(\sum_{j=1}^{m} w_j g_{ij}\right) \qquad (3)$$

$$y_i = \gamma_0 + \mathbf{X}'_\mathbf{i}\gamma_\mathbf{1} + \beta_c\left(\sum_{j=1}^{m} w_j g_{ij}\right) + \varepsilon_i \qquad (4)$$

We recommend using MAF-based weights to up-weight low-frequency variants: $w_j = Beta(\hat{p}_j,\ \alpha = 1,\ \beta = 25)$, where weights have beta density function with prespecified parameters α and β, and $\hat{p}_j$ is the variant MAF [Wu et al., 2011]. In simulations and real data analysis, we use $\alpha = 1$ and $\beta = 25$ as suggested by Wu et al. [2011]. To test the gene-based null hypothesis H_o: $\beta_c = 0$, the burden score statistic is

$$Q_B = \left(\sum_{j=1}^{m} w_j S_j\right)^2, \qquad (5)$$

where $S_j = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)g_{ij}$ is the score statistic for testing H_o: $\beta_j = 0$ with only variant j in the regression model, and $\hat{\mu}_i$ is the estimated mean of y_i under H_o. The burden score statistic is evaluated relative to a scaled χ^2_1 distribution [Wu et al., 2011].

SKAT assumes the β_j's follow an arbitrary distribution with mean zero and variance $w_j^2\tau$. Testing the null hypothesis H_o: $\boldsymbol{\beta} = 0$ is equivalent to testing H_o: $\tau = 0$. The SKAT score statistic is

$$Q_S = \sum_{j=1}^{m} w_j^2 S_j^2 \qquad (6)$$

and follows a mixture of chi-square distributions [Lee et al., 2012].

Wu et al. [2011] showed that the power of the burden test and SKAT depends on the underlying genetic architecture of

the analyzed genomic region. For example, the burden test is more powerful when most variants in the region are causal and have the same direction of effect; in contrast, SKAT is more powerful when fewer variants are causal and/or have opposite directions of effect. The combined test of burden test and SKAT, SKAT-O [Lee et al., 2012] combines the strength of both tests and is powerful in both scenarios. The SKAT-O statistic is a weighted average of Q_B and Q_S:

$$Q_\rho = \rho Q_B + (1-\rho)Q_S, \quad 0 \le \rho \le 1 \tag{7}$$

with weight parameter ρ. In practice, ρ is unknown. To estimate the optimal ρ, we perform a grid search on $0 = \rho_1 < \rho_2 < \cdots < \rho_b = 1$ and select ρ such that the Q_ρ is maximized (or the corresponding P-value is minimized). We choose to perform the search on $\rho = \{0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1\}$ as suggested by Lee et al. [2012]. Significance is evaluated analytically by numerical integration [Lee et al., 2012].

For analysis of case-control studies, we examine the effect of applying the small-sample adjustment described in Lee et al. [2012].

Numerical Simulations

To generate simulated genomic regions, we used 10,000 haplotypes simulated using the COSI coalescent simulator [Schaffner et al., 2005], as provided in the SKAT R package. For each simulated male individual, we randomly selected a 3 Kb region from a single random haplotype. For each simulated female individual, we selected a 3 Kb region from two random haplotypes and paired them together. For a simulated sample of 1,000 cases and 1,000 controls under the null hypothesis, the 3 Kb region has average number of variants = 36.8 (SD = 6.0), with a median total minor allele count [MAC] = 2,812 (interquartile range [IQR] = 1,766–4,270). When considering only variants with MAF < 0.01, the average number of variants = 28.4 (SD = 5.2), with median total MAC = 99 (IQR = 79–125).

Type I Error Simulations

For binary traits, we simulated case-control datasets with N_{cases} = 1,000 and N_{ctrls} = 1,000 under the logistic regression model:

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \beta_1 g_{i1} + \cdots + \beta_t g_{it} \tag{8}$$

with one continuous covariate X_{1i} normally distributed with mean zero and variance one, one binary covariate X_{2i} distributed Bernoulli with success probability $f = 0.5$, sex covariate X_{3i}, and selected causal variants $g_{i1}, \ldots, g_{it}$; under the null model, we set genetic effects $\beta_1 = \cdots = \beta_t = 0$. The sex covariate accounts for differences in genotype frequency between males and females, so that we can avoid inflated type I error rates as for Clayton's score test [2008] when allele frequencies differ between sexes [Loley et al., 2011]. We set the intercept γ_0 so that the disease prevalence is 10%, the covariate regression coefficients $\gamma_1 = \gamma_2 = 0.5$, and the effect for sex $\gamma_3 = 0$. We explored a broad range of male:female ratios in cases and controls by sampling the exact number of males and females (Table 1) from the simulated cases and controls. Note that there is an implicit sex-phenotype effect when we sample unbalanced numbers of males and females in cases and controls in simulation scenarios B and C (Table 1).

Table 1. Sample sizes for simulated case-control datasets

Simulation	No. of cases males:females	No. of controls males:females
A	500:500	500:500
B	900:100	500:500
C	100:900	500:500

Table 2. Sample sizes for simulated quantitative trait datasets

Simulation	No. of individuals males:females
D	1,000:1,000
E	200:1,800
F	1,800:200

For QTs, we took a similar approach to simulate datasets of N = 2,000 individuals under the null linear regression model:

$$y_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \beta_1 g_{i1} + \cdots + \beta_t g_{it} + \varepsilon_i \tag{9}$$

with X_{1i}, X_{2i}, X_{3i}, and $g_{i1}, \ldots, g_{it}$ as for Equation (8). We set covariate effect sizes $\gamma_1 = \gamma_2 = 0.5$, or equivalently, the proportion of trait variance explained $\sigma^2_{X1} = (\gamma_1)^2$ and $\sigma^2_{X2} = (\gamma_2)^2 f(1-f)$; we set the effect of sex $\gamma_3 = 0$ or 0.5. The normally distributed residual error had mean zero and variance = $1 - \sigma^2_{X1} - \sigma^2_{X2}$. We sampled the desired number of males and females (Table 2) from the simulated individuals.

We analyzed each simulated dataset using the six combinations of three gene-based tests (Eqs. (5)–(7)) and two coding schemes for male genotypes. To increase computational efficiency, we simulated 100,000 independent datasets per simulation scenario, and resampled the phenotype 1,000 times per independent dataset, resulting in a total of 100 million simulation replicates. We evaluated the robustness of the resampling approach by comparing results with 1 million independent simulated datasets without resampling for a subset of the simulation scenarios. We estimated type I error as the proportion of simulation replicates with a P-value $< \alpha = 2.5 \times 10^{-6}$, corresponding to Bonferroni correction for association testing of the approximately 20,000 genes in the human genome.

Power Simulations

Within the 3 Kb region, we selected 10% or 50% of variants with MAF < 0.03 as causal. Using the same settings as for the type I error simulations, we simulated case-control datasets using the logistic regression model (Eq. (8)). We simulated QT datasets using the linear regression model (Eq. (9)), assuming the normally distributed residual error ε_i had mean zero and variance = $1 - \sigma^2_{X1} - \sigma^2_{X2} - \sum_{j=1}^{t} \sigma^2_j$, where

Table 3. Type I error rates for burden, SKAT, and SKAT-O tests for simulated binary and quantitative trait studies

Binary traits						
				Type I error rate ($\times 10^{-6}$)		
Simulation	No. of cases males:females	No. of controls males:females	Coding for male genotypes	Burden	SKAT	SKAT-O
A	500:500	500:500	No X-inactivation $g = \{0,1\}$	2.4	2.6	2.4
			X-inactivation $g = \{0,2\}$	1.8	1.9	1.6
B	900:100	500:500	No X-inactivation $g = \{0,1\}$	3.1	3.9	2.9
			X-inactivation $g = \{0,2\}$	1.8	2.9	2.1
C	100:900	500:500	No X-inactivation $g = \{0,1\}$	3.1	3.4	3.0
			X-inactivation $g = \{0,2\}$	2.8	5.2	3.7

Quantitative traits					
			Type I error rate ($\times 10^{-6}$)		
Simulation	No. of individuals males:females	Coding for male genotypes	Burden	SKAT	SKAT-O
D	1,000:1,000	No X-inactivation $g = \{0,1\}$	2.5	2.1	2.8
		X-inactivation $g = \{0,2\}$	2.5	2.4	2.8
E	1,800:200	No X-inactivation $g = \{0,1\}$	2.7	2.5	3.3
		X-inactivation $g = \{0,2\}$	2.5	2.4	3.1
F	200:1,800	No X-inactivation $g = \{0,1\}$	2.6	2.8	3.4
		X-inactivation $g = \{0,2\}$	2.7	2.8	3.7

Type I error estimates are based on 10^8 simulation replicates so that the nominal significance threshold of $\alpha = 2.5 \times 10^{-6}$ corresponds to 250 rejections. Empirical type I error rates between 2.2×10^{-6} and 2.8×10^{-6} have 95% confidence intervals that include the nominal value.

$\sigma_j^2 = (\beta_j)^2 2p_j(1-p_j)$ is the proportion of trait variance explained by variant j.

We simulated datasets under the alternative hypothesis assuming X-inactivation and non-X-inactivation coding for male genotypes. We assumed genetic effect sizes proportional to the variant MAF $|\beta_j| = c|\log_{10} p_j|/2$, and adjusted the tuning parameter c so that power estimates were not too close to 1 or 0. For binary traits, when 10% of variants were causal, we set $c = \log(15)$, to give an odds ratio of 15 when MAF = 0.01; when 50% variants were causal, we set $c = \log(3)$ or $\log(5)$. For QTs, when 10% of variants were causal, we set $c = \log(7)$, which gave a linear regression coefficient of approximately 1.95 when MAF = 0.01; when 50% of variants were causal, we set $c = \log(1.8)$. We assumed that either all causal variants were deleterious, or that 50% were deleterious and 50% were protective. We simulated 1,000 independent replicates per simulation scenario, and evaluated power as the proportion of replicates with P-value $< 2.5 \times 10^{-6}$.

Genetics of Type 2 Diabetes (GoT2D) Study

To assess these methods in the context of real data, we analyzed integrated sequencing and genotyping data from the GoT2D study, which aims to investigate the impact of low-frequency genetic variation on type 2 diabetes (T2D) risk. The GoT2D study sample is composed of 1,326 T2D cases and 1,331 normal glucose tolerant controls from the United Kingdom (322 cases / 322 controls), Finland (486/517), the Botnia region of Finland (199/159), Germany (104/101), and Sweden (222/227). There are 716 males and 610 females in cases and 592 males and 739 females in controls. For each sample, we performed low-pass (~5×) whole genome sequencing, augmented by deep (~100×) whole-exome sequencing, and genotyping of 2.5M single nucleotide polymorphisms (SNPs) using the Illumina HumanOmni2.5 array. To mimic the type I error and power simulations, we analyzed X-chromosome data using the six combinations of three gene-based tests and two coding schemes for male genotypes. For each test, we adjusted for the effects of sex, the first two genotype-based principal components (PCs) to control for population stratification [Price et al., 2006], and indicator functions for observed temporal stratification based on sequencing date and center. To filter out likely neutral variants from the analysis, we restricted the gene-based analysis to protein-truncating and missense variants. We used the *seqminer* [Zhan and Liu, 2013] and *SKAT* [Lee, 2014] R packages to extract the genotypes and perform the gene-based analysis, respectively.

Results

Type I Error Rates

For binary traits, the burden, SKAT, and SKAT-O tests are well-calibrated or slightly anti-conservative (at $\alpha = 2.5 \times 10^{-6}$) for all scenarios considered (Table 3). For each gene-based test, we examined type I error rates for two male genotype coding schemes: (1) $g_{ij} = \{0,1\}$ and (2) $g_{ij} = \{0,2\}$; and datasets with three male:female ratios in cases: (Simulation A) 500:500, (B) 900:100, and (C) 100:900. For datasets with male:female ratio = 500:500 in cases (Simulation A), non-X-inactivation coding is less conservative than X-inactivation coding for all three tests: burden (non-X-inactivation type I error rate = 2.4×10^{-6} vs. X-inactivation = 1.8×10^{-6}), SKAT (2.6×10^{-6} vs. 1.9×10^{-6}), and SKAT-O (2.4×10^{-6} vs. 1.6×10^{-6}). These patterns also hold true for datasets with male:female ratio = 900:100 in cases (Simulation B). In comparison, for datasets with male:female ratio = 100:900 in cases (Simulation C), non-X-inactivation coding is less

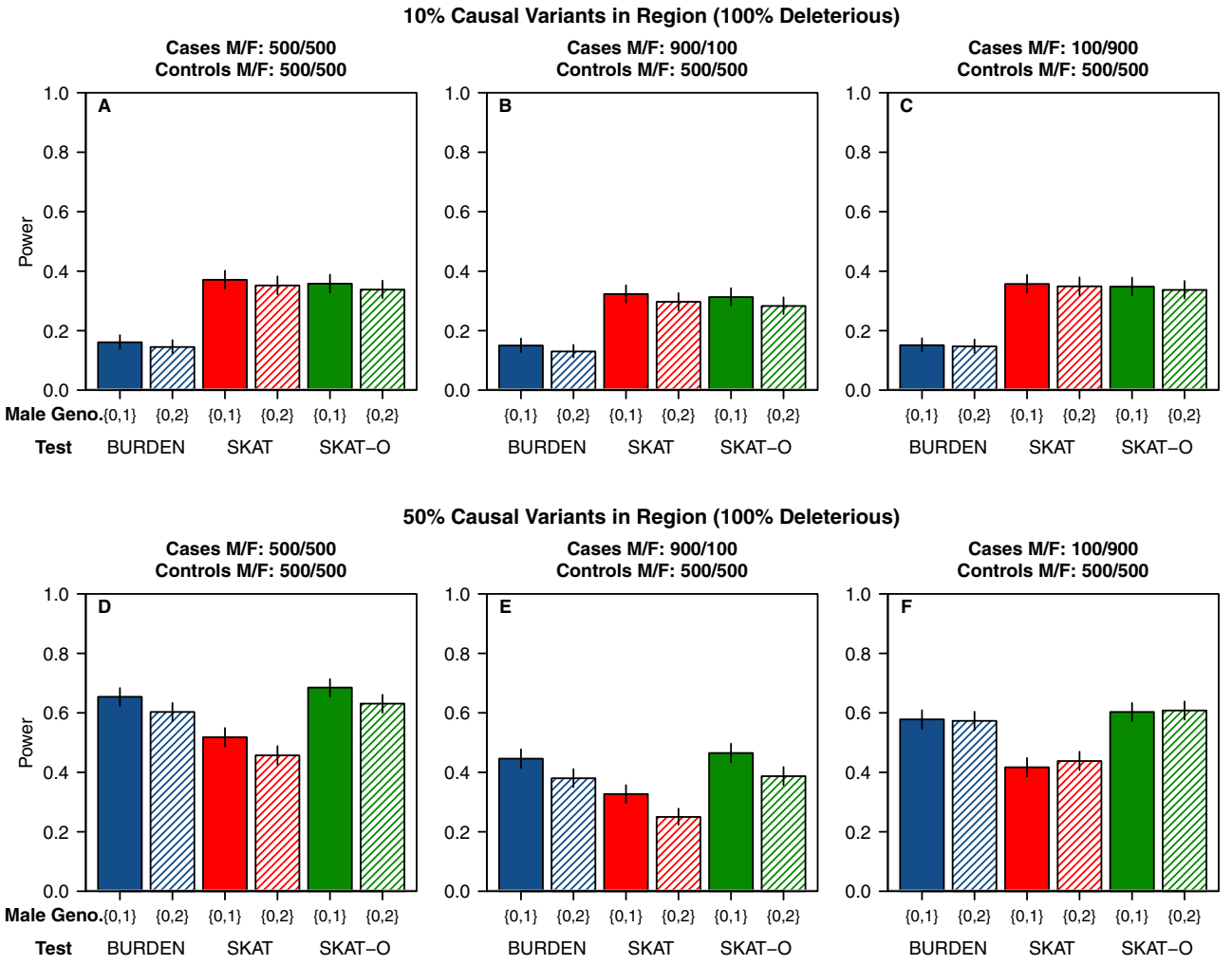

Figure 1. Power for gene-based tests in case-control studies assuming all causal variants are deleterious. Causal variants are simulated with non-X-inactivation coding of male genotypes $g_{ij} = \{0,1\}$. Within each simulated 3 Kb genomic region, (A–C) 10% or (D–F) 50% of variants with MAF < 0.03 are selected as causal. The effect size for causal variants is given by $|\beta_j| = c\,|\log_{10} p_j|/2$, and is proportional to MAF (p_j) and scaled by tuning parameter (A–C) $c = \log(15)$ and (D–F) $c = \log(3)$; all causal variants are simulated as deleterious. Simulated cases have (A, D) 500:500, (B, E) 900:100, and (C, F) 100:900 males and females, respectively. All simulated controls have 500 males and 500 females. Power estimates (at $\alpha = 2.5 \times 10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

conservative for burden, but more conservative for SKAT and SKAT-O. These results are generalizable to other male:female sample sizes and depend only on the male:female ratio in cases and controls (supplementary Fig. S1).

We examined the impact of the implicit nonzero sex-phenotype effect on test calibration by sampling unbalanced numbers of males and females in cases and controls (Simulations B and C; Table 1). Tests are similarly calibrated with or without the sex-phenotype effect (Table 3). We also examined the effect of applying the small-sample adjustment [Lee et al., 2012] to the three gene-based tests; type I error rates are generally slightly anti-conservative after applying the small-sample adjustment (supplementary Fig. S1), but the patterns of type I error rates between male genotype coding schemes and male:female ratios are identical to those without small-sample adjustment. Finally, we demonstrated the accuracy of our computationally efficient resampling approach by comparing type I error rates with resampling to those without resampling (10^6 independent replicates; $\alpha = 5 \times 10^{-4}$); type I error rates are comparable with (supplementary Fig. S2A–C) and without resampling (supplementary Fig. S2D–F).

For QTs, the burden and SKAT tests are well-calibrated and SKAT-O can be very slightly anti-conservative across the three simulated datasets with male:female ratios of 1,000:1,000; 1,800:200; and 200:1,800 (Table 3). Type I error rates are nearly identical between the two male coding schemes. Tests are similarly well-calibrated with or without the inclusion of a nonzero sex-phenotype effect (supplementary Table S1).

Power

We examined power for four combinations of proportion of causal variants in a region (10% or 50%) and causal variant direction of effect (all deleterious, or 50% deleterious and

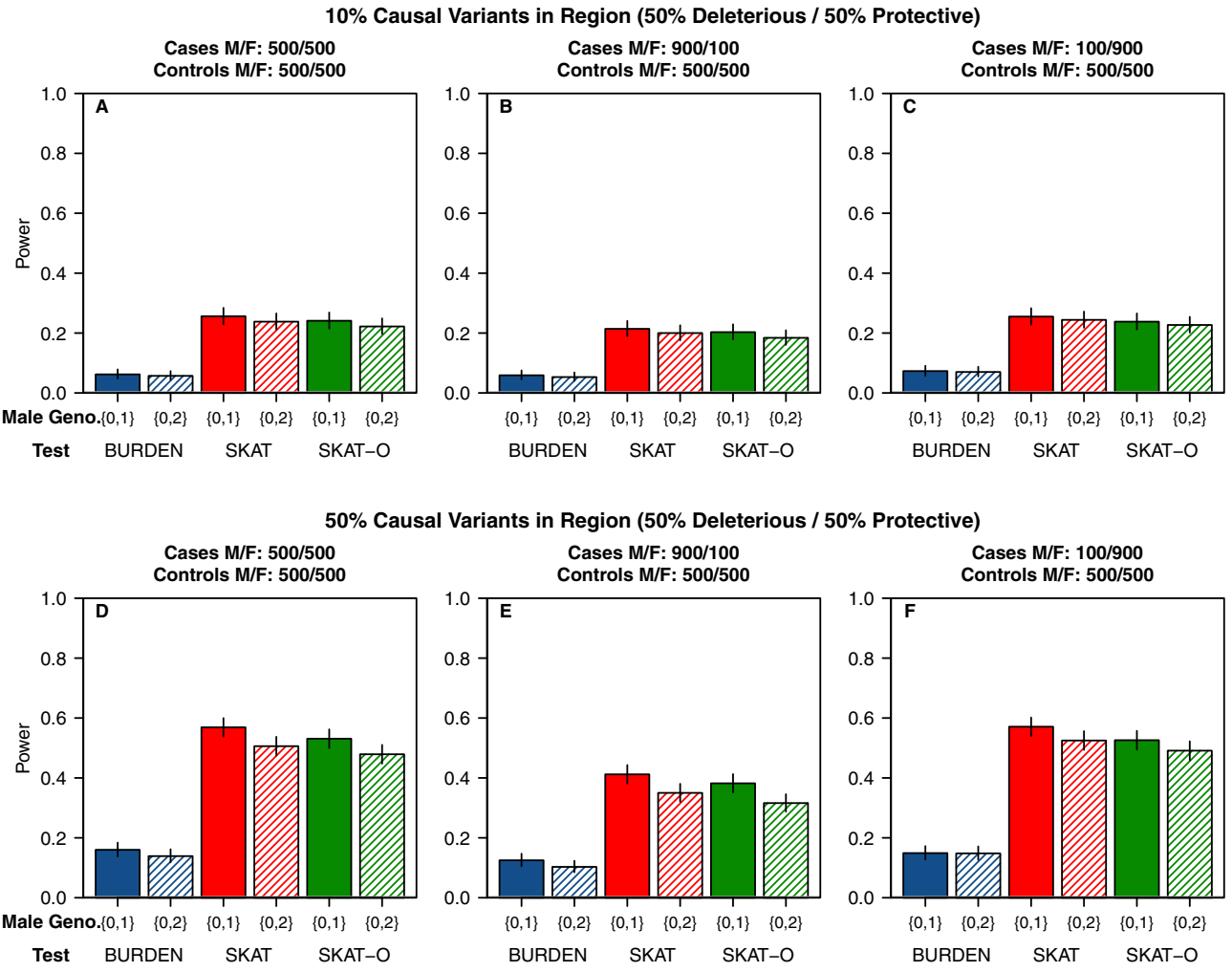

Figure 2. Power for gene-based tests in case-control studies assuming causal variants are 50% deleterious and 50% protective. Causal variants are simulated with non-X-inactivation coding of male genotypes $g_{ij} = \{0,1\}$. Within each simulated 3 Kb genomic region, (A–C) 10% or (D–F) 50% of variants with MAF < 0.03 are selected as causal. The effect size for causal variants is given by $|\beta_j| = c\,|\log_{10} p_j|/2$, and is proportional to MAF (p_j) and scaled by tuning parameter (A–C) $c = \log(15)$ and (D–F) $c = \log(5)$; causal variants are simulated as 50% deleterious and 50% protective. Simulated cases have (A, D) 500:500, (B, E) 900:100, and (C, F) 100:900 males and females, respectively. All simulated controls have 500 males and 500 females. Power estimates (at $\alpha = 2.5 \times 10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

50% protective). For binary traits, power results (Figs. 1 and 2) reflect the previously described relative power of gene-based tests for different underlying genetic architectures [Lee et al., 2012]. For example, the burden test is more powerful when 50% of low-frequency variants are causal and have the same direction of effect (Fig. 1D–F). SKAT is more powerful when 10% of low-frequency variants are causal with the same or opposite direction of effect (Fig. 1A–C), or when 50% of causal variants have opposite direction of effect (Fig. 2). SKAT-O is generally robust and powerful across all scenarios tested. Despite the slightly anti-conservative type I error rates, the small-sample adjusted and nonadjusted power estimates are comparable (data not shown).

Next, we investigated the effect of simulating causal variants with (male genotype coding $g_{ij} = \{0,2\}$) and without ($g_{ij} = \{0,1\}$) X-inactivation. The two coding schemes for male genotypes have only a small effect on power. When simulating variants assuming no X-inactivation, non-X-inactivation coding ($g_{ij} = \{0,1\}$) is slightly more powerful in all scenarios (Figs. 1 and 2). However, when simulating variants assuming X-inactivation, X-inactivation coding ($g_{ij} = \{0,2\}$) is slightly more powerful in nearly all scenarios (supplementary Figs. S3 and S4). However, the power loss for misspecifying the unknown model is small. For example, in simulations assuming non-X-inactivation coding, the largest power loss for misspecifying the coding scheme is 7.7% (32.7% vs. 25.0%) for SKAT (Fig. 1E), and the average absolute difference is 2.8%. Power is generally lower for studies with more males than females, due to decreased effective sample size (e.g., number of chromosomes). For example, for the burden test with non-X-inactivation coding ($g_{ij} = \{0,1\}$), studies with 900 males:100 females in cases is less powerful than with 500 males:

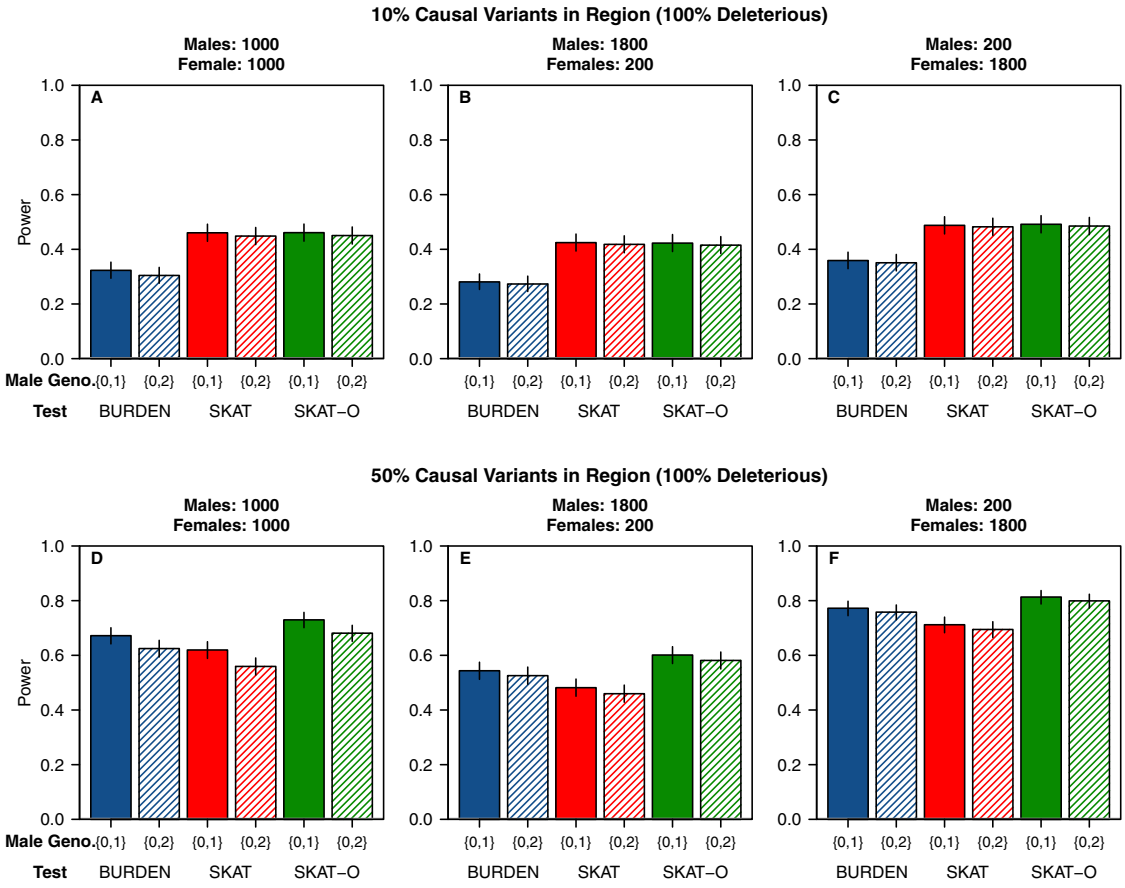

Figure 3. Power for gene-based tests in QT studies assuming all causal variants are deleterious. Causal variants are simulated with non-X-inactivation coding of male genotypes $g_{ij} = \{0,1\}$. Within each simulated 3 Kb genomic region, (A–C) 10% or (D–F) 50% of variants with MAF < 0.03 are selected as causal. The effect size for causal variants is given by $|\beta_j| = c\,|\log_{10} p_j|/2$, and is proportional to MAF (p_j) and scaled by tuning parameter (A–C) $c = \log(7)$ and (D–F) $c = \log(1.8)$; all causal variants are simulated as deleterious. Simulated datasets have (A, D) 1,000:1,000, (B, E) 1,800:200, and (C, F) 200:1,800 males and females, respectively. Power estimates (at $\alpha = 2.5 \times 10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

500 females (44.6% vs. 65.4%; Fig. 1D and 1E). Finally, for QTs, power comparisons are very similar to those for binary traits (Figs. 3 and 4).

Analysis of GoT2D X-Chromosome Data

To examine the effect of different coding schemes for male genotypes, we performed gene-based association testing for 505 X-chromosome genes with T2D risk using low-pass sequencing data from 1,326 T2D cases and 1,331 healthy controls from the GoT2D study. Within each gene region, we restricted the variants to those predicted to be protein truncating or missense mutations. Gene regions had an average number of variants = 6.1 (SD = 5.5), and a median total MAC = 9 (IQR = 4–20). We analyzed the data using six combinations of three gene-based tests and two male genotype coding schemes, adjusting for the effects of sex and the first two PCs to account for population stratification.

Scatterplots and quantile-quantile plots show that the association analysis *P*-values are concordant between X-inactivation and non-X-inactivation coding for all three tests (Fig. 5A–F). The differences in *P*-value between coding schemes are generally small (Fig. 5G–I). For example, for the burden test, 68.5% (346/505 genes) of the analysis *P*-values have an absolute difference of 0.1 $\log_{10}$ units or less; 99.4% (502/505) of *P*-values are within 0.5 $\log_{10}$ units.

Discussion

We examined the calibration and power of the burden, SKAT, and SKAT-O gene-based association tests for analyzing the X chromosome in simulated binary and QT data. For binary traits, all tests are well-calibrated or slightly anticonservative for all simulation scenarios. Power differences reflected the previously described strengths and weaknesses of each test for analyzing regions with differing underlying genetic architectures. Power is usually slightly increased when we code male genotypes with the coding scheme that matches the underlying genetic model (e.g., with or without X-inactivation), but power loss is modest when we misspecify the (unknown) coding scheme. Studies with more male samples typically have lower power. For QTs, the burden and

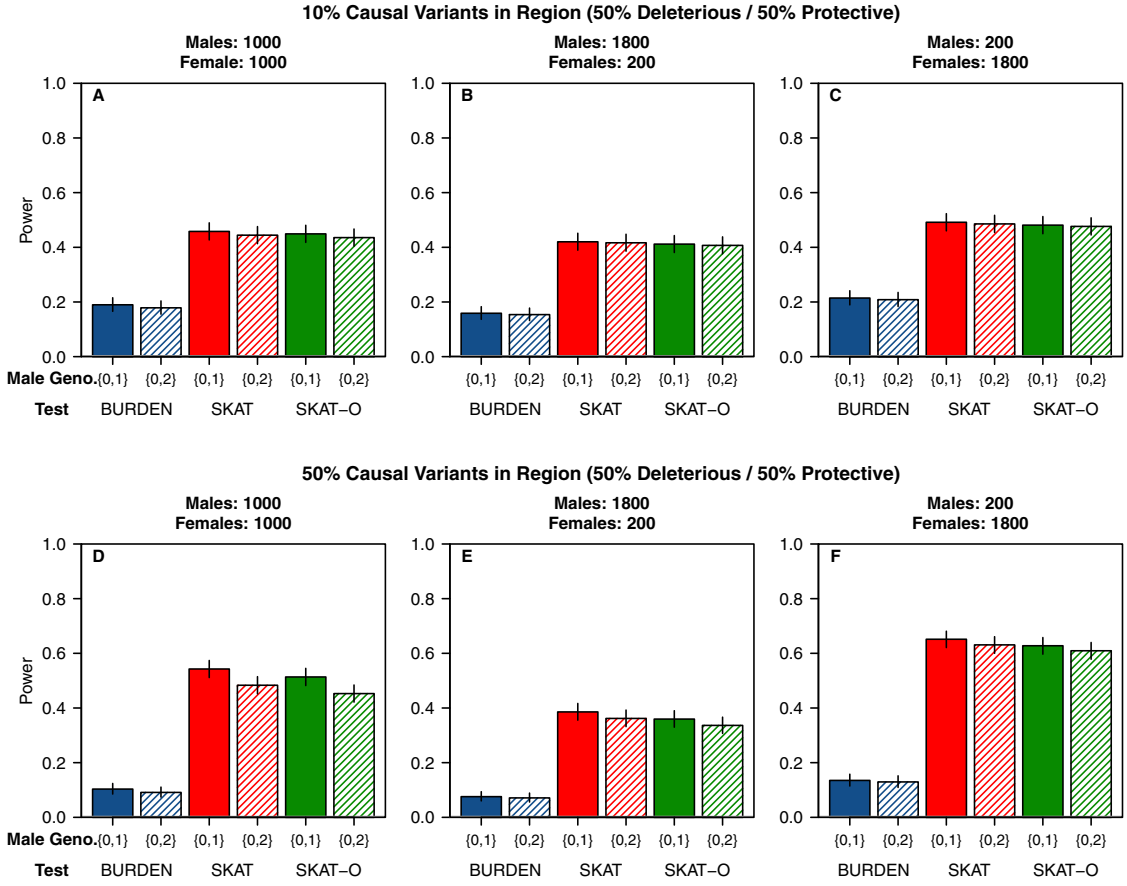

Figure 4. Power for gene-based tests in QT studies assuming causal variants are 50% deleterious and 50% protective. Causal variants are simulated with non-X-inactivation coding of male genotypes $g_{ij} = \{0,1\}$. Within each simulated 3 Kb genomic region, (A–C) 10% or (D–F) 50% of variants with MAF < 0.03 are selected as causal. The effect size for causal variants is given by $|\beta_j| = c\,|\log_{10} p_j|/2$, and is proportional to MAF (p_j) and scaled by tuning parameter (A–C) $c = \log(7)$ and (D–F) $c = \log(1.8)$; causal variants are simulated as 50% deleterious and 50% protective. Simulated datasets have (A, D) 1,000:1,000, (B, E) 1,800:200, and (C, F) 200:1,800 males and females, respectively. Power estimates (at $\alpha = 2.5 \times 10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

SKAT tests are well-calibrated, and SKAT-O is very slightly anti-conservative; power results mirror those for binary traits.

In practice, investigators will need to use external biological information to select the preferred male coding scheme. For example, we recommend using the non-X-inactivation coding scheme for known pseudoautosomal regions. In the absence of external information, we suggest using the X-inactivation coding because a higher proportion of X-linked genes do undergo X-inactivation. It is known that only 15% of X-linked genes escape inactivation to some degree [Carrel and Willard, 2005]. We have demonstrated that misspecifying the coding scheme leads to only a small loss in power. In the GoT2D study, we also demonstrated that using either coding scheme produces similar results.

Although we only presented calibration and power results for a specific set of simulation settings, we performed a variety of simulations with other covariate settings, case:control ratios, and prevalence rates to demonstrate that our results are generalizable (data not shown). We also considered the effect on calibration of allele frequency differences between males and females. To introduce large differences in allele frequency between sexes, we sampled male genotypes from a set of simulated haplotypes mimicking European American ancestry, and female genotypes from haplotypes mimicking African American ancestry. The gene-based tests remain well-calibrated despite the allele frequency differences (supplementary Table S2).

We estimated nominal power as the proportion of simulation replicates (under the alternative hypothesis) with P-values $< 2.5 \times 10^{-6}$. We estimated empirical power as the proportion of simulated datasets with P-values less than the estimated empirical threshold: the αth quantile of the 10^8 P-values for the simulated samples under the null hypothesis. Overall, power using nominal or empirical thresholds yield near-identical power because tests are relatively well-calibrated across all scenarios, and the empirical thresholds (range $= 1 \times 10^{-6}$–3.5×10^{-6}) are very similar to the nominal significance threshold. The greatest difference between nominal and empirical power is for a study with 100 males and 900 females in cases, and 500 males and 500 females in controls, where 50% of low-frequency variants are causal and have the same direction of effect. For the SKAT test using

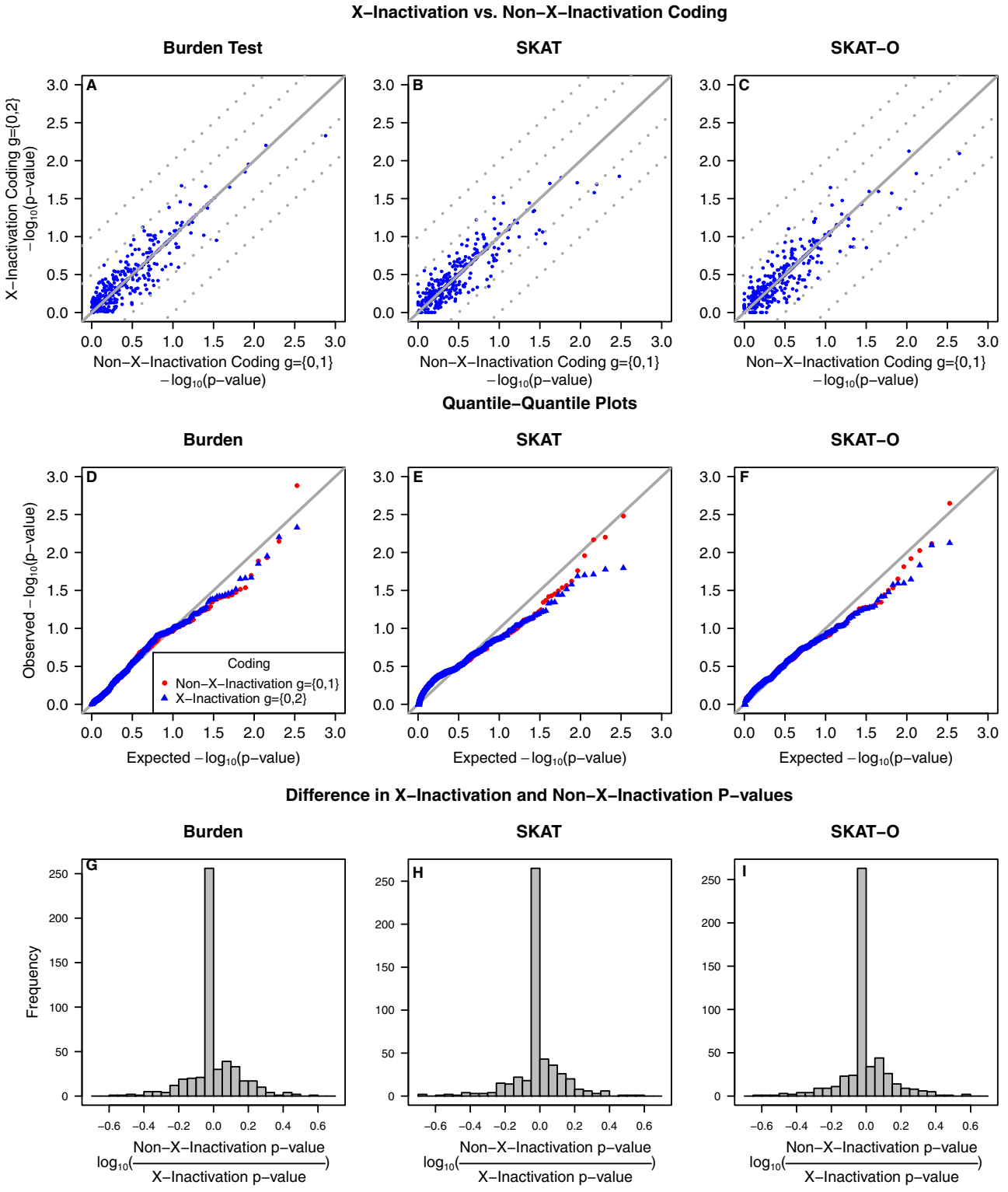

Figure 5. Comparison of association analysis *P*-values in the GoT2D study. (A–C) Scatterplots compare analysis *P*-values using non-X-inactivation $g = \{0,1\}$ and X-inactivation $g = \{0,2\}$ coding schemes for male genotypes. (D–F) Quantile-quantile plots compare expected and observed *P*-value distributions. (G–I) Histograms display the difference in *P*-values between the coding schemes.

X-inactivation coding ($g_{ij} = \{0,2\}$), nominal power = 41.7% vs. empirical power = 40.1%. We decided to present our power results assuming nominal thresholds, because this is how analysis actually will be done.

We did not evaluate the calibration and power for other gene-based association methods, such as the WST [Madsen and Browning, 2009], C-alpha [Neale et al., 2011], and SSU [Pan, 2009] tests. However, our burden test is equivalent to the WST, and SKAT includes the C-alpha and SSU tests as a special case, indicating that our results would likely extend to these gene-based tests. We only evaluate test calibration and power assuming an additive genetic model. For gene-based tests, we are primarily interested in testing the (joint) effect of rare variants. Because minor allele homozygotes are rare, we expect the dominant genetic model to have near identical calibration and power to the additive model, and the recessive model to have very low power. We only explored the effect of random X-inactivation, where ~50% of the cells have one female allele inactivated and the remaining ~50% of the other. We did not examine the possibility of nonrandom or skewed X-inactivation [Amos-Landgraf et al., 2006], where >75% of cells have one allele inactivated. Although a unified approach to account for both random and nonrandom X-inactivation may be more robust and powerful, as demonstrated by Wang et al. [2014] for single-marker association testing, we speculate that misspecifying the model for the burden, SKAT, and SKAT-O gene-based tests will only result in a small power loss.

In conclusion, we generalized the burden, SKAT, and SKAT-O tests to analyze X-chromosome variants, and demonstrated that these tests are generally well-calibrated and powerful for a wide range of simulation scenarios. These tests can be directly applied to the association analysis of less common variants on the X chromosome.

Acknowledgments

We thank Laura J. Scott, Gonçalo Abecasis, and Cristen Willer for their helpful discussions and suggestions for the manuscript, and our GoT2D colleagues for allowing us to use sequencing data. This research was supported by the National Institutes of Health grants HG000376 and DK062370 to M.B. and HL113164 to S.L.

References

Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, Willard HF. 2006. X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *Am J Hum Genet* 79:493–499.

Asimit J, Zeggini E. 2010. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 44:293–308.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434:400–404.

Clayton D. 2008. Testing for association on the X chromosome. *Biostatistics* 9:593–600.

Hickey PF, Bahlo M. 2011. X chromosome association testing in genome wide association studies. *Genet Epidemiol* 35:664–670.

Hindorff LA, MacArthur J, Wise A, Junkins HA, Hall PN, Klemm AK, Manolio TA. 2012. A catalog of published genome-wide association studies. *NHGRI*. Available at: www.genome.gov/gwastudies

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 44:955–959.

Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A, Stringham HM, Sim X, Yang L, Fuchsberger C, Cederberg H, and others. 2013. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* 45:197–201.

Lee S. 2014. SKAT: SNP-set (Sequence) Kernel Association Test. R package version 1.0.5. Available at: http://CRAN.R-project.org/package=SKAT.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team NGESP-ELP, Christiani DC, Wurfel MM, Lin X. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91:224–237.

Lee S, Abecasis GR, Boehnke M, Lin X. 2014. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 95:5–23.

Loley C, Ziegler A, Konig IR. 2011. Association tests for X-chromosomal markers—a comparison of different test statistics. *Hum Hered* 71:23–36.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* 190:372–373.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5:e1000384.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39:906–913.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7:e1001322.

Pan W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 33:497–507.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and others. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15:1576–1583.

Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadottir HT, Johannsdottir H, Magnusson OT, Gudjonsson SA, and others. 2014. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* 46:294–298.

Wang J, Yu R, Shete S. 2014. X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genet Epidemiol* 38:483–493.

Wise AL, Gyi L, Manolio TA. 2013. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* 92:643–647.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82–93.

Zhan X, Liu DJ. 2013. TaSer (TabAnno and SeqMiner): a toolset for annotating and querying next-generation sequence data. Available at: http://zhanxw.com/seqminer.

Zheng G, Joo J, Zhang C, Geller NL. 2007. Testing association for markers on the X chromosome. *Genet Epidemiol* 31:834–843.