STATISTICAL METHODS FOR LOW-FREQUENCY AND RARE GENETIC VARIANTS

by

Clement Ma

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2014

Doctoral Committee:

Professor Michael L. Boehnke, Co-Chair
Research Associate Professor Laura J. Scott, Co-Chair
Professor Gonçalo Abecasis
Assistant Professor Hyun M. Kang
Assistant Professor Seunggeun Lee
Professor Peter X. Song
Assistant Professor Cristen J. Willer

Dedication

To my wife, Joyce.

Acknowledgements

I would like to express my deepest thanks and gratitude to my advisors Mike Boehnke and Laura Scott. Both of you were outstanding mentors, and encouraged me to achieve what I once thought was impossible. Without your careful guidance and mentorship, I would not be the statistical geneticist that I am today. I also want to thank my committee members: Gonçalo Abecasis, Hyun Min Kang, Seunggeun Lee, Peter Song, and Cristen Willer for their constructive feedback and support for my dissertation research.

I would like to thank all my colleagues from the Center for Statistical Genetics. I learned a great deal from Tom Blackwell, who was an active collaborator on my first two dissertation topics. Thanks to Sean Caron and Paul Anderson for helping me run my simulations smoothly and efficiently on the computing cluster. I want to thank Dawn Keene and Laura Baker for helping me on my faculty applications and other administrative issues.

I want to thank my many colleagues and collaborators outside the University of Michigan. Thank you to all the GoT2D study collaborators for allowing me to use an early data freeze of the sequencing data for my dissertation research. I gained many useful insights from the regular participants of the Single Variant Group conference call. Thanks to Georg Heinze for helpful discussions regarding the Firth bias-corrected logistic regression test.

I want to thank all my Ann Arbor area friends who have made my five years here fun and memorable. Thanks to Mark Reppell and Adrian Tan who were always there, and served as groomsmen at my wedding ceremony. I would like to thank Ryan Welch, Rebecca Rothwell, Yancy Lo, Giorgio Pistis, Eleonora Porcu, Zhenzhen Zhang, Caroline Cheng, Katie Huang, Lisa Henn, Min A Jhun, Yeji Lee, Tanya Teslovich, Xueling Sim, Adam Locke, Christopher Moraes, and Stefanie Moraes for the many board game nights, dinners, and happy hours.

I want to thank my family, Danny, Joyce, and Winnie Ma for supporting me throughout my graduate studies. I was very fortunate that Ann Arbor is within driving distance to Toronto, so I was able to visit home frequently. Thanks to my new family, King, Susan, and Coral Wong, who have been very supportive and encouraging during my time in Michigan.

Most of all, I want to thank my wife, Joyce Wong, who supported me every step of the way. Five years ago, she encouraged me to pursue my dream of doctoral studies, even though it would mean we would spend over four years living apart from each other. You were always there to cheer me up, listen to my fears, and share a laugh. I am so happy that you were able to join me in Ann Arbor, and watch me complete this long yet rewarding journey. I truly could not have done this without you.

Table of Contents

List of Tables

List of Figures

List of Supplemental Figures

Abstract

Genetic association studies using sequencing, dense-array genotyping, or sequencing-based imputation provide the means to identify low-frequency and rare variants associated with diseases and traits, but analysis of these variants presents new statistical challenges. Single marker tests (e.g. logistic and linear regression), and methods to combine information across studies (e.g. joint and meta-analysis) may be poorly calibrated and/or of low power. The calibration and power of aggregation tests, where multiple rare variants are analyzed jointly, have not been evaluated for variants on the X chromosome. In my dissertation, I address three topics:

First, for case-control studies, I evaluate the calibration and power of four logistic regression tests in joint and meta-analysis for low-frequency and rare variants and demonstrate that: (a) for joint analysis, the Firth bias-corrected test is best (e.g. most powerful among well-calibrated tests); (b) for meta-analysis of balanced studies (equal numbers of cases and controls), the score test is best, but is less powerful than Firth test-based joint analysis; and (c) for meta-analysis of sufficiently unbalanced studies, all four tests can be anti-conservative, particularly the score test.

Second, for quantitative trait (QT) studies, I evaluate the calibration and power of linear regression in joint and meta-analysis and demonstrate for normally distributed QTs that: joint and sample-size weighted meta-analysis are equally well-calibrated and powerful for variants with expected minor allele count E[MAC]≥10; inverse-variance weighted meta-analysis is slightly anti-conservative for small-sized studies. For non-normally distributed QTs, joint and meta-analysis is equally anti-conservative for low-frequency and rare variants. Inverse-normal transformation of the QT remedies this problem, but transforming QTs of any distribution reduces power.

Third, for case-control and QT studies, I evaluate the calibration and power of three aggregation tests for the X chromosome: burden, SKAT, and SKAT-O. For case-control studies, tests are relatively well-calibrated across all simulation scenarios. Power is usually slightly increased when the coding scheme for male genotypes matches the underlying model, but power loss is small when the model is misspecified. Differences in male:female ratio in cases and controls have little effect on power. For QTs, calibration and power results are very similar to those for binary traits.

Chapter 1: Introduction

Many human diseases and biological traits can be hereditary in nature [Gottlieb and Root, 1968; Kaprio et al., 1992; Silventoinen et al., 2003], but their genetic mechanisms are not fully understood. In genome-wide association studies (GWAS), we aim to identify genetic variants that cause differences in biological traits or disease risk. While many associated variants identified by GWAS are not causal, associated variants help localize genes or genomic regions that may harbor the true causal variants. Through fine-mapping and functional studies, we hope to identify the true causal variants, and better understand the biological mechanisms underlying human diseases and traits [Shea et al., 2011; Kulzer et al., 2014].

Genotype array-based common-variant GWAS have identified thousands of genetic variants associated with hundreds of different traits [Hindorff et al., 2012]. Investigators typically use case-control studies to detect disease-associated and cohort studies to detect quantitative trait (QT)-associated variants. We also often analyze QTs collected from case-control studies to identify variants associated with these QTs. To increase power to detect novel variants with small effect sizes in GWAS, investigators often combine samples across multiple association studies, typically using meta-analysis of summary-level association results [Scott et al., 2007], and less frequently, joint analysis of the combined individual-level data [Schizophrenia Psychiatric Genome-Wide Association Study Consortium, 2011].

Although early genotyping arrays can only assay hundreds of thousands of common variants per individual, these variants are sufficient to tag a large proportion of the common variation in the population [International HapMap Consortium, 2005]. Since studies use different genotype arrays, only the small subset of overlapping variants can be meta-analyzed together directly. Genotype imputation using early reference panels (such as HapMap haplotypes [International HapMap Consortium, 2005]) fills in missing common genotypes with high accuracy, and allows the meta-analysis of the same dense set of genetic markers across all available samples [Marchini et al., 2007; Li et al., 2010].

Nearly all associated variants identified by GWAS are common [Hindorff et al., 2012]. Tagging single nucleotide polymorphisms (SNPs) on early genotype arrays, combined with genotype imputation, accurately captures a large proportion of common variation, but could not reliably tag low-frequency (0.005 < minor allele frequency [MAF] ≤ 0.05) and rare (MAF ≤ 0.005) variants [Zeggini et al., 2005]. Most low-frequency and rare variants are inaccurately imputed, because they are poorly tagged, and difficult to phase [Asimit and Zeggini, 2012]. For many traits, the associated common variants have small effect sizes and only account for a small fraction of the total genetic variability [Maher, 2008; Manolio et al., 2009]. Low-frequency and rare variants may have larger effect sizes to explain a portion of this "missing heritability", leading to new insights on the genetic architecture of different traits [Cirulli and Goldstein, 2010].

Next generation sequencing has great promise to evaluate the impact of low-frequency and rare variants on human diseases and traits. As sequencing costs decrease and accuracy increases, many sequencing-based association study designs are now being undertaken. Sequencing of hundreds to a few thousand study samples can allow investigators to evaluate the impact of low-frequency variation on traits, although sequencing costs remain prohibitive for larger sample sizes. Investigators can also leverage existing GWAS cohorts by imputation using dense, sequencing-based reference panels, such as those from the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2010] or the Genetics of Type 2 Diabetes (GoT2D) study. Finally, specialized low-cost genotyping arrays, such as the Exome Chip [2013], and Metabochip [Voight et al., 2012], facilitate analysis of specific genomic regions in tens of thousands of individuals.

New statistical challenges emerge for the analysis of low-frequency and rare variants. For both binary and quantitative traits (QT), for common variants, Single marker tests (e.g. logistic and linear regression) are well-calibrated and powerful, and meta-analysis has comparable power to joint analysis [Lin and Zeng, 2010]. These methods can still be used to analyze low-frequency variants (0.005< MAF ≤0.05) given sufficiently large sample sizes, but specific tests in joint or meta-analysis may have poor calibration and/or low power for analysis of low-frequency variants. For example, the logistic regression Wald test can be extremely conservative when analyzing low-frequency variants in case-control studies [Xing et al., 2012]. For analysis of low-frequency variants, the calibration and power of logistic regression tests for case-control studies, and linear regression for quantitative traits (QT) studies in joint and meta-analysis, has not been fully investigated. Further, combining data across multiple studies to increase power is critical to detect both common and low-frequency variant associations. For both binary and quantitative traits, meta-analysis has comparable power to joint analysis for common variants [Lin and Zeng, 2010], but this is not necessarily true for low-frequency variants. Hence, we need a careful comparison of different association tests in both joint and meta-analysis for low-frequency variants.

For rare variants (MAF ≤ 0.005), single marker tests can have (very) low power unless sample sizes are very large. Furthermore, for case-control studies, variants with fewer than a minimum number of minor allele counts [MAC] can never attain the typical GWAS significance threshold ($\alpha = 5\times10^{-8}$), used to account for multiple testing of approximately one million independent markers in the human genome [Risch and Merikangas, 1996; Dudbridge and Gusnanto, 2008]. For example, in a balanced study with equal number of cases and controls ($N_{cases} = N_{controls} = 1{,}000$) and no additional covariates, a variant must have MAC ≥ 26 to have the possibility to attain a Fisher's Exact Test (FET) p-value < 5×10^{-8}; this remains true for all balanced studies with larger sample sizes ($N_{cases} = N_{controls} > 1{,}000$). For QTs, individual rare variants can be significantly associated at $\alpha = 5\times10^{-8}$ if individuals carrying the rare allele also have QT values in the tails of the phenotypic distribution. Nonetheless, the utility of single marker tests is diminished for both binary and quantitative trait studies when the variants of interest are rare in frequency.

In comparison, aggregation tests, where multiple markers in a region are jointly analyzed, can be more powerful for analyzing rare variants than single marker tests [Li and Leal, 2008]. In addition, we can evaluate the significance of aggregation tests at a less stringent threshold, typically $\alpha = 2.5\times10^{-6}$ to account for testing approximately 20,000 genes in the human genome, compared to $\alpha = 5\times10^{-8}$ for testing single markers genome-wide. While there are a multitude of published aggregation tests, most tests can be classified into two broad categories: (1) burden tests, and (2) dispersion tests. Burden tests assume that causal variants have similar effect sizes and measure the association between the (weighted) sum of the rare alleles in a region with the disease or QT; examples include the Combined Multivariate and Collapsing (CMC) test [Li and Leal, 2008] and Weighted Sum Statistic (WSS) [Madsen and Browning, 2009]. Dispersion tests assume a distribution of effect sizes and measure deviations from the expected distribution; examples include the Sequence Kernel Association Test (SKAT) [Wu et al., 2011] and C-alpha [Neale et al., 2011] test. Each class of tests is most powerful for genomic regions with different underlying genetic architectures [Wu et al., 2011; Lee et al., 2012]. For example, burden tests are most powerful when there are many causal variants with the same direction of effect in the genomic region, while dispersion tests are most powerful when causal variants have opposite directions of effect. Since the underlying genetic architecture is generally unknown, Lee et al. [2012] proposed the optimal unified association test (SKAT-O) to combine the strengths of burden and dispersion tests, and showed that SKAT-O is generally powerful across different genetic architectures.

These aggregation tests were developed to analyze rare variants on the autosomal chromosomes. Analysis of rare variants on the X chromosome requires proper treatment of diploid female and haploid male subjects. For example, while we typically code variant genotypes for females as the number of minor alleles $g = \{0,1,2\}$, there are at least two possible coding schemes for male genotypes. If the variant is in a locus with X-inactivation [Lyon, 1961], one copy of the female allele is inactivated so one copy of the male allele might be considered equivalent to two copies of the female allele; we might choose to code male genotypes as $g = \{0,2\}$. However, if the variant is in a locus without X-inactivation, we might choose to code male genotypes as $g = \{0,1\}$. Hence, we need specialized analysis methods to analyze X chromosomal data, but existing methods focus on single marker analysis, and are not powerful for rare variants [Zheng et al., 2007; Clayton, 2008]. To my knowledge, there are no aggregation association tests developed for analyzing rare variants on the X chromosome. Hence, we need to evaluate the applicability of existing aggregation tests for analyzing the X chromosome for two different coding schemes for male genotypes.

In this dissertation, I focus on three topics related to statistical methods for analyzing low-frequency and rare variants from genetic association studies. First, for case-control studies, I evaluate the calibration and power of logistic regression tests in joint and meta-analysis for low-frequency variants [Ma et al., 2013]. Second, for QT studies, I evaluate the calibration and power of linear regression in joint and meta-analysis for low-frequency variants. Third, for case-control and QT studies, I evaluate the calibration and power of the burden, SKAT, and SKAT-O aggregation tests for variants on the X chromosome.

In Chapter 2, for case-control studies, I seek to (1) identify the association test with maximal power among tests with well-controlled type I error rate and (2) compare the relative power of joint and meta-analysis tests. I use analytic calculation and simulation to compare the empirical type I error rate and power of four logistic regression-based tests: Wald, score, likelihood ratio, and Firth bias-corrected. I demonstrate for low count variants (roughly MAC < 400, empirically determined by the simulations) that: (1) for joint analysis, the Firth test has the best combination of type I error and power; (2) for meta-analysis of balanced studies (equal numbers of cases and controls), the score test is best, but is less powerful than Firth-test based joint analysis; and (3) for meta-analysis of sufficiently unbalanced studies, all four tests can be anti-conservative, particularly the score test. I also establish MAC, rather than MAF, as the key parameter determining test calibration for joint and meta-analysis.

In Chapter 3, for QT studies, I assess and compare the calibration and power of linear regression in joint and meta-analysis for QT association analysis across all a wide range of allele frequencies, but concentrate on low-frequency and rare variants, and examine the impact of sample size, additional covariates, and non-normally distributed QTs, and additional covariates. I show that for variants of any frequency that for normally-distributed QTs with inverse-normal transformation within study, joint analysis and sample-size weighted meta-analysis are equally well-calibrated and powered for variants with expected E[MAC] ≥ 10, but inverse-variance weighted meta-analysis is slightly anti-conservative for small-sized studies; joint and meta-analysis are increasingly conservative for E[MAC] < 10. These results continue to hold when modeling the effects of additional covariates. For non-normally distributed QTs, joint and meta-analysis can become equally anti-conservative for low-frequency and rare variants. Inverse-normal transformation of the QT remedies this problem, but the transformation reduces power for QTs of any distribution. These results continue to hold when modeling the effects of additional covariates. Hence, for association analysis of variants of any frequency, given normally distributed QTs, I recommend either joint or meta-analysis using linear regression.

In Chapter 4, for case-control and QT studies, I evaluate the calibration and power of three aggregation tests for the X chromosome: burden, SKAT, and SKAT-O. For case-control studies, I show that all tests are relatively well-calibrated across all simulation scenarios. As expected, the power of each test depends on the underlying genetic architecture of the analyzed region. Power is usually slightly increased when the coding scheme for male genotypes (e.g. with or without X-inactivation) matches the underlying genetic model, but misspecifying the coding scheme only results in small power loss. Power only varies slightly between datasets with different male:female ratios in cases and controls. For QTs, burden and SKAT are well-calibrated, and SKAT-O is very slightly anti-conservative; patterns of power are very similar to those for binary traits.

In Chapter 5, I summarize my results, discuss efficient simulation strategies for estimating type I error rates at stringent thresholds, and present my ideas for future work.

Chapter 2: Recommended joint and meta-analysis strategies for case-control association testing of single low count variants

Introduction

Genome-wide association studies (GWAS) have identified thousands of common variants associated with hundreds of diseases and traits [Hindorff et al., 2012]. The standard GWAS analysis framework using asymptotic-theory tests has proven to be well-calibrated and powerful, given sufficiently large sample sizes. In this context, for analysis of binary traits such as disease status, classical logistic regression-based Wald, score, and likelihood ratio tests have well-controlled type I error rates and are asymptotically equivalent [Cox and Hinkley, 1974]. Since individual studies often are not large enough to detect variants with modest genetic effects, information can be combined across multiple studies using either meta-analysis of study-level association results or joint analysis of the combined individual-level data. For common variants, meta-analysis is widely used since there are fewer logistical and ethical constraints in sharing association results than sharing individual-level data, and since meta-analysis has near equivalent power to joint analysis [Lin and Zeng, 2010].

Sequencing-based study designs including next-generation sequencing, imputation using dense reference panels, and specialized genotyping arrays provide new opportunities to test low frequency or low count variants for disease association. Here we operationally define as low count a variant with minor allele count (MAC) < 400, equivalent to minor allele frequency (MAF) < 0.05 for a study with N = 4000 individuals, or MAF < 0.01 for N = 20000. For a given study design with N > 2000, we demonstrate that MAC provides a more consistent and sample-size invariant measure of the genetic variant's inherent information, compared to MAF. We also show that a MAC of 400 is a rough threshold separating variants for which tests have relatively poor calibration (for MAC < 400) from relatively good calibration (for MAC > 400) for balanced and not too unbalanced studies.

For analysis of low count variants, collapsing [Li and Leal, 2008] and burden [Madsen and Browning, 2009; Wu et al., 2011] tests, in which multiple markers are analyzed together, are often performed. However, single marker tests remain important for variants that have sufficient counts. Analysis of individual low count variants poses new challenges and questions. The asymptotic assumptions for logistic regression may no longer be valid, resulting in either conservative or anti-conservative test behavior. For example, the Wald test is extremely conservative for low count variants [Hauck and Donner, 1977; Xing et al., 2012]. Since sequencing-based studies may discover tens of millions of mostly low count variants, we require even more stringent significance thresholds than for analysis of high count variants in GWAS, further straining asymptotic assumptions. Little is known about the relative efficiency of joint and meta-analysis for low count variants.

In this paper, we aim to identify the most powerful test(s) with well-controlled empirical type I error in joint and meta-analysis of binary traits for low count variants. In situations where all evaluated tests are either conservative or anti-conservative, we aim to identify the "best" test having type I error rates nearest to but not exceeding the nominal threshold, and with greatest power. To do so, we compare analytically calculated and simulation estimated type I error rates and power for four logistic regression tests in joint and meta-analysis. We evaluate these tests across a wide range of MACs at stringent significance thresholds in studies with varying sample size and case-control imbalance. For low count variants, our results show that joint analysis using the Firth bias-corrected logistic regression test [Firth, 1993] is consistently best for both balanced and unbalanced studies. For meta-analysis of balanced studies, the logistic regression score test is best. Comparing joint and meta-analysis for balanced studies, Firth test-based joint analysis is more powerful than score test-based meta-analysis. For meta-analysis of substantially unbalanced studies, all of the tests evaluated can be anti-conservative. We establish MAC as the key parameter determining test calibration.

Materials and Methods

Notation

We consider first a single case-control study with total sample size *N*. For individual *i*, let Y_i = 1 or Y_i = 0 denote a case or control respectively, and X_i = 0, 1, 2 the number of minor alleles for a specific genetic variant.

Logistic regression tests

We consider four asymptotic tests based on the logistic regression model

$$\text{logit}[\Pr(Y_i = 1)] = \alpha + \beta X_i \qquad \text{(Equation 2.1)}$$

where α is the study-specific intercept and β is the genotype log odds ratio (OR). We wish to test the null hypothesis of no association H$_0$: $\beta = 0$. The Wald test statistic is

$$W = \hat{\beta} \big/ SE(\hat{\beta}) \qquad \text{(Equation 2.2)}$$

where $\hat{\beta}$ is the maximum likelihood estimate (MLE) for β and $SE(\hat{\beta})$ is its standard error. Given the log-likelihood $l(\alpha, \beta)$, the likelihood ratio test statistic is

$$LR = -2[l(\tilde{\alpha}, 0) - l(\hat{\alpha}, \hat{\beta})] \qquad \text{(Equation 2.3)}$$

where $\tilde{\alpha}$ is the restricted MLE of α under the null model, and $(\hat{\alpha}, \hat{\beta})$ is the MLE of (α, β) under the full model. The score test statistic is

$$S = U_\beta \Big/ \sqrt{\text{var}(U_\beta)} \qquad \text{(Equation 2.4)}$$

where $U_\beta = \partial l / \partial \beta$ is the component of the score function corresponding to parameter β evaluated at $(\alpha, \beta) = (\tilde{\alpha}, 0)$. The variance of the score statistic [Cox and Hinkley, 1974] is

$$\text{var}(U_\beta) = I_{\beta\beta}(\tilde{\alpha}, 0) - I_{\beta\alpha}(\tilde{\alpha}, 0)\, I_{\alpha\alpha}^{-1}(\tilde{\alpha}, 0)\, I_{\alpha\beta}(\tilde{\alpha}, 0)$$

where $I_{AB} = -\partial l^2 \big/ \partial A \partial B$ is the *AB* component of the observed Fisher information matrix. The Wald and score test statistics are evaluated relative to a standard normal distribution, the likelihood ratio test statistic relative to a χ_1^2 distribution.

In logistic regression models, "separation" occurs when cases and controls can be perfectly explained by a non-trivial linear combination of the covariates [Albert and Anderson, 1984]. Separation occurs most often in small studies. It can also occur in larger studies with categorical covariates for which some categories are rare (for example, low count variants), since at least one covariate category may occur only in cases or only in controls. In separated datasets, logistic regression produces strongly biased parameter estimates diverging to ±∞. Firth [1993] proposed a penalized likelihood function to correct the first-order asymptotic bias of parameter estimates which is especially relevant for separated datasets. The Firth bias-corrected log-likelihood function is

$$l^*(\alpha,\beta) = l(\alpha,\beta) + 0.5\log\left|I(\alpha,\beta)\right|$$

where $I(\alpha,\beta)$ is the information matrix. The bias-corrected likelihood ratio statistic described by Heinze and Schemper [2002] is

$$F = -2[l^*(\tilde{\alpha}^*, 0) - l^*(\hat{\alpha}^*, \hat{\beta}^*)] \qquad \text{(Equation 2.5)}$$

where $\tilde{\alpha}^*$ and $(\hat{\alpha}^*, \hat{\beta}^*)$ are the corresponding bias-corrected MLEs for the null and full models (using the observed information matrix), respectively. The bias-corrected likelihood ratio statistic is evaluated relative to a χ_1^2 distribution. We modified Ploner's *R* implementation of the bias-corrected logistic regression test [Ploner, 2010] to increase computational efficiency, and included the modified implementation in the EPACTS software [Kang, 2012].

Combining data across studies: joint and meta-analysis

We next consider *K* case-control studies in which study *k* has sample size N_k. In joint analysis, we perform association testing on the individual-level genotype and phenotype data from all $N = \sum_k N_k$ individuals across the *K* studies. Thus, for each asymptotic test (Equations 2.2-2.5), we use the joint log-likelihood constructed based on all *N* individuals. To account for differences between studies in the logistic regression model (Equation 2.1), it is possible to include population or study-specific covariates such as study indicators or principal components and modify the asymptotic test statistics (Equations 2.2-2.5) accordingly.

In meta-analysis, we perform a separate association test within each study and combine the study-level association results (for example, using p-values and directions of effect, transformed into z-scores). For each asymptotic test (Equations 2.2-2.5) for study *k*, we use the study-specific log-likelihood constructed based on the relevant N_k individuals. We use sample-size weighted meta-analysis, since this requires only study-level p-values and direction of effect and so is applicable to all of the statistical tests we evaluated. We assume fixed underlying effects rather than random effects for each study since we wish to maximize power for hypothesis testing, rather than focus on effect estimation.

For study *k*, we determine the corresponding quantile q_k with χ_1^2 distribution, with upper tail probability equal to the association p-value, and calculate the equivalent z-score $Z_k = \pm\sqrt{q_k}$, with sign based on direction of effect. The sample-size weighted meta-analysis z-score is

$$Z_{SS} = \sum_{k=1}^{K} \sqrt{\overline{N}_k} Z_k \Big/ \sqrt{\sum_{k=1}^{K} \overline{N}_k}$$

where $\overline{N}_k = 4N_{1,k}N_{0,k}/(N_{1,k}+N_{0,k})$ is the effective sample size of study *k* with $N_{1,k}$ cases and $N_{0,k}$ controls [Mantel and Haenszel, 1959; Han and Eskin, 2011].

Analytical calculation of type I error rate for joint analysis

For joint analysis, we calculate type I error rates for significance levels $\alpha = 5\times10^{-5}$ and 5×10^{-8} by enumerating all possible MAC configurations, and summing the probabilities of configurations that reject H_0, similar to a method described by Upton [Upton, 1982]. For simplicity, we assume a dominant disease model, which is a good approximation to a multiplicative model (on the OR scale) for low count variants, since individuals homozygous for the minor allele are rare. For simulation-based estimation of type I error rates and power in the next section, we assume a multiplicative disease model (on the OR scale). In a single study with N_1 cases and N_0 controls, let T_1 and T_0 denote the number of cases and controls who carry at least one copy of the minor allele. Under the null hypothesis, given population MAF *p* and assuming Hardy-Weinberg equilibrium, T_1 and T_0 have binomial distributions:

$$T_1 \sim \text{Binomial}(N_1, 1-[1-p]^2)$$
$$T_0 \sim \text{Binomial}(N_0, 1-[1-p]^2)$$

There are $(N_1+1)\times(N_0+1)$ possible MAC configurations, and the joint probability of each configuration is the product of the corresponding marginal probabilities.

We calculate the Wald, score, likelihood ratio, and Firth bias-corrected p-values for every MAC configuration. The exact type I error rate for a given test is

$$\sum_{i=0}^{N_1}\sum_{j=0}^{N_0}\Pr[T_1=i,T_0=j]\cdot I[p-value_{ij}\le\alpha]$$

where $\Pr[T_1=i,T_0=j]$ is the probability for the *(i,j)th* configuration and $I[p-value_{ij}\le\alpha]$ is an indicator whether the configuration yields significant evidence for association at level α. Analytical calculation allows us to determine type I error rates efficiently at stringent significance thresholds (α = 5×10^{-8}) for a wide range of sample sizes and degrees of case-control imbalance.

Simulation-based estimation of type I error and power for joint and meta-analysis

For meta-analysis, analytic calculation of type I error is computationally infeasible since the number of possible configurations across multiple studies becomes extremely large. Instead, we simulate datasets using *R* [R Development Core Team, 2012] based on the logistic regression model (Equation 2.1) assuming disease prevalence 10%. Each dataset is simulated based on a causal variant with specified population-level MAF (and corresponding expected MAC) and genotype OR. In contrast to the dominant model assumed in the analytical calculations, we assume the more commonly used multiplicative genetic model (on the OR scale) in the simulated datasets. We verify that even for a variant with MAF = 0.05, type I error and power estimates for dominant (analytical) and multiplicative (simulated) models are nearly identical, and result in the same relative rankings among the tests (data not shown). For simplicity, we did not include additional covariates. We simulate full datasets with 10000/10000, 8000/12000, 5000/15000, and 1000/19000 cases and controls, respectively. We subdivide each full dataset into *K* = 10 equal-sized sub-studies with identical case-control ratios, analyze each sub-study separately, and meta-analyze the sub-study association results. We perform up to 10 million simulation replicates under the null model (OR = 1) to estimate type I error rates at $\alpha = 5\times10^{-4}$ or 5×10^{-5}, and 10000 replicates under alternative models (OR > 1) to estimate power at $\alpha = 5\times10^{-8}$.

Genetics of type 2 diabetes (GoT2D) study

To illustrate these methods, we analyze an early data-freeze subset of the whole-genome sequencing data from the Genetics of Type 2 Diabetes (GoT2D) study, which aims to assess the effect of low frequency variation on T2D risk in Northern Europeans. Our dataset contains 908 individuals (499 T2D cases and 409 controls) from three contributing studies: (1) 195 Swedish and Botnian Finnish individuals (116 cases / 79 controls) from the Diabetes Genetics Initiative, (2) 575 Finnish individuals (304/271) from the Finland-United States Investigation of NIDDM Genetics (FUSION) study, and (3) 138 British individuals (79/59) from the UK T2D Genetics Consortium. We perform joint analysis on the combined sample and sample-size weighted meta-analysis on association results from each of the three contributing studies using EPACTS [Kang, 2012] for association testing and METAL [Willer et al., 2010] for meta-analysis. To match simulation settings, we did not adjust for additional covariates in these analyses.

Results

Overview

We examine empirical type I error rates and power in joint and meta-analysis for the four logistic regression tests across a range of MACs, sample sizes, and degrees of case-control imbalance. For joint analysis, we analytically calculate empirical type I error rates for a nominal significance threshold of $\alpha = 5\times10^{-8}$. For sample-size weighted meta-analysis, we estimate type I error using simulation at a less stringent threshold ($\alpha = 5\times10^{-4}$ [Figure S2.2 only] or 5×10^{-5}) due to computational constraints. For both joint and meta-analysis, we estimate power using simulation at $\alpha = 5\times10^{-8}$ over a range of effect sizes (suited to the variant MAC). We seek to identify the "best" test with highest power while maintaining a well-controlled type I error rate. We confirm the consistency of type I error rates for a variant with fixed MAC.

Type I error rates of joint and meta-analysis tests

We first examine joint analysis type I error rates ($\alpha = 5\times10^{-8}$) for a single balanced study with 10000 cases and 10000 controls (Figure 2.1A). For high count variants (expected MAC > 400; MAF > 0.01 for $N = 20000$), we focus on type I error estimates for a variant with expected MAC = 2000 (MAF = 0.05); we observe that all tests are well-calibrated. For low count variants (E[MAC] < 400; MAF < 0.01), joint analysis using the Firth test (red solid line) consistently has type I error rates nearest to while not exceeding the nominal threshold. The score and Wald tests are very conservative, while the likelihood ratio test is slightly anti-conservative for some MACs.

Next, we consider type I error rates ($\alpha = 5\times10^{-5}$) for meta-analysis of 10 balanced sub-studies each with 1000 cases and 1000 controls (Figure 2.1G). For high count variants, all tests are again well-calibrated. For low count variants, score test-based meta-analysis (blue dashed line) has type I error rates nearest to but not exceeding the nominal threshold. Meta-analysis using Firth and particularly Wald test results are more conservative, while using likelihood ratio test results is again anti-conservative for some MACs. Comparing the joint and meta-analysis tests with type I error rates nearest to but not exceeding the nominal threshold, the Firth test-based joint analysis (red solid line; Figure 2.1D) is less conservative than the score test-based meta-analysis (blue dashed line; Figure 2.1G). For example, at E[MAC] = 40 (MAF = 0.001), the empirical type I error rate (at $\alpha = 5\times10^{-5}$) for Firth test-based joint analysis (4.2×10^{-5}) is less conservative than score test-based meta-analysis (2.3×10^{-5}).

We extend our investigation of joint analysis of unbalanced studies with 5000/15000 (1:3) and 1000/19000 (1:19) cases and controls, respectively (Figure 2.1B-C). For high count variants, the Firth (red) and likelihood ratio (black) tests are well-calibrated, but the score and Wald tests can be anti-conservative given substantial case-control imbalance. For low count variants, Firth test-based joint analysis has type I error rates consistently nearest to but not exceeding the nominal threshold. The Wald and particularly the score test become extremely anti-conservative for increasingly unbalanced studies, while the likelihood ratio test can be slightly anti-conservative for some MACs. We observe these trends for joint

Figure 2.1: Type I error rates by minor allele count (MAC) for logistic regression tests in joint and meta-analysis.

(A - C) Analytically calculated type I error rates ($\alpha = 5\times10^{-8}$) for joint analysis; (D - F) empirical type I error rates ($\alpha = 5\times10^{-5}$) for joint analysis; and (G - I) empirical type I error rates ($\alpha = 5\times10^{-5}$) for sample-size weighted meta-analysis. Type I error rates for joint analysis are estimated for studies with 10000/10000, 5000/15000, and 1000/19000 total cases and controls; meta-analysis is based on partitioning the full dataset into 10 equal-sized sub-studies. The horizontal dotted line denotes the corresponding nominal significance threshold. Points in panels D - I are based on 10^7 simulation replicates so that the nominal significance threshold of 5×10^{-5} corresponds to 500 rejections; empirical type I error rates between 4.6×10^{-5} and 5.4×10^{-5} have 95% confidence intervals which include the nominal value.

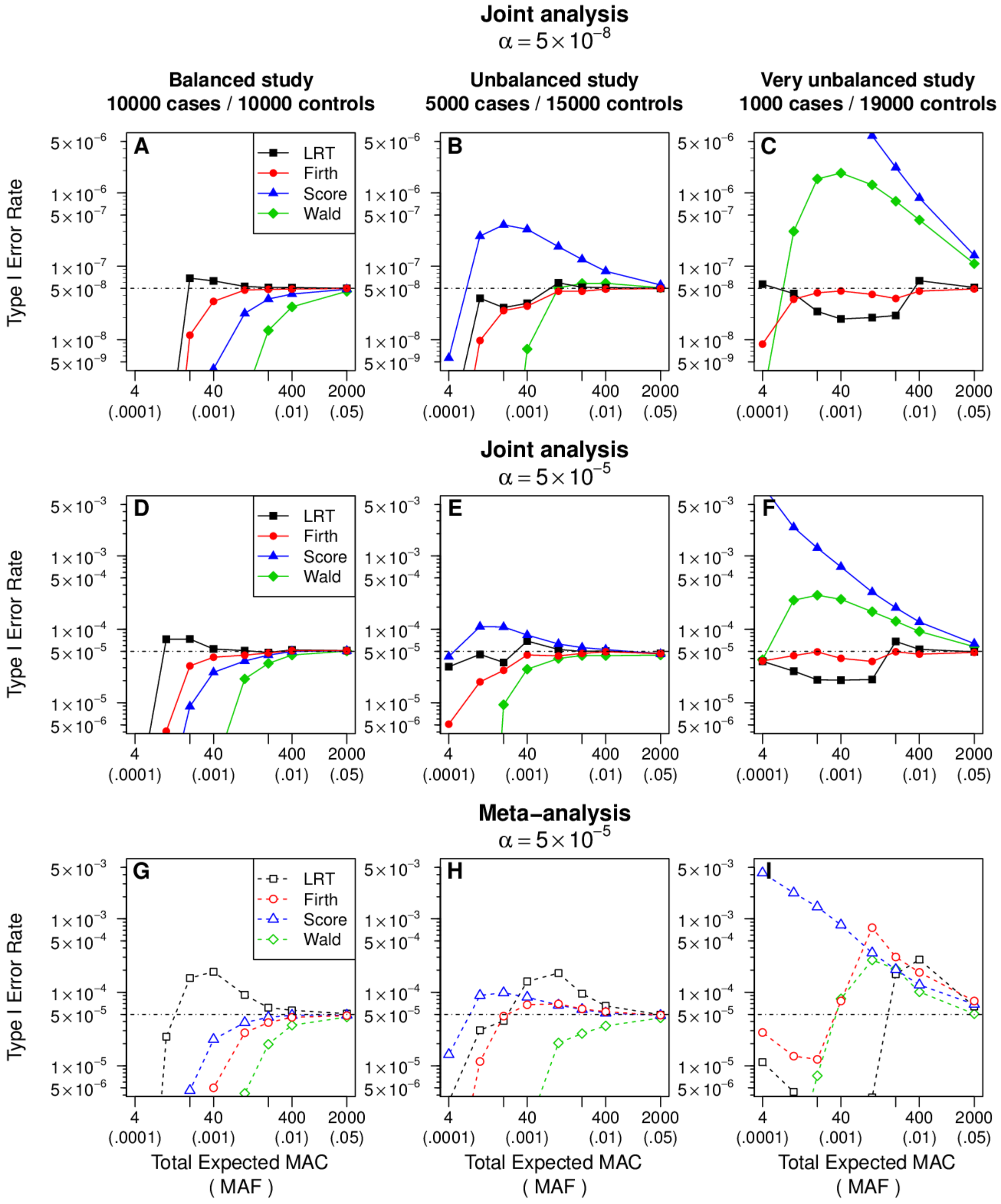
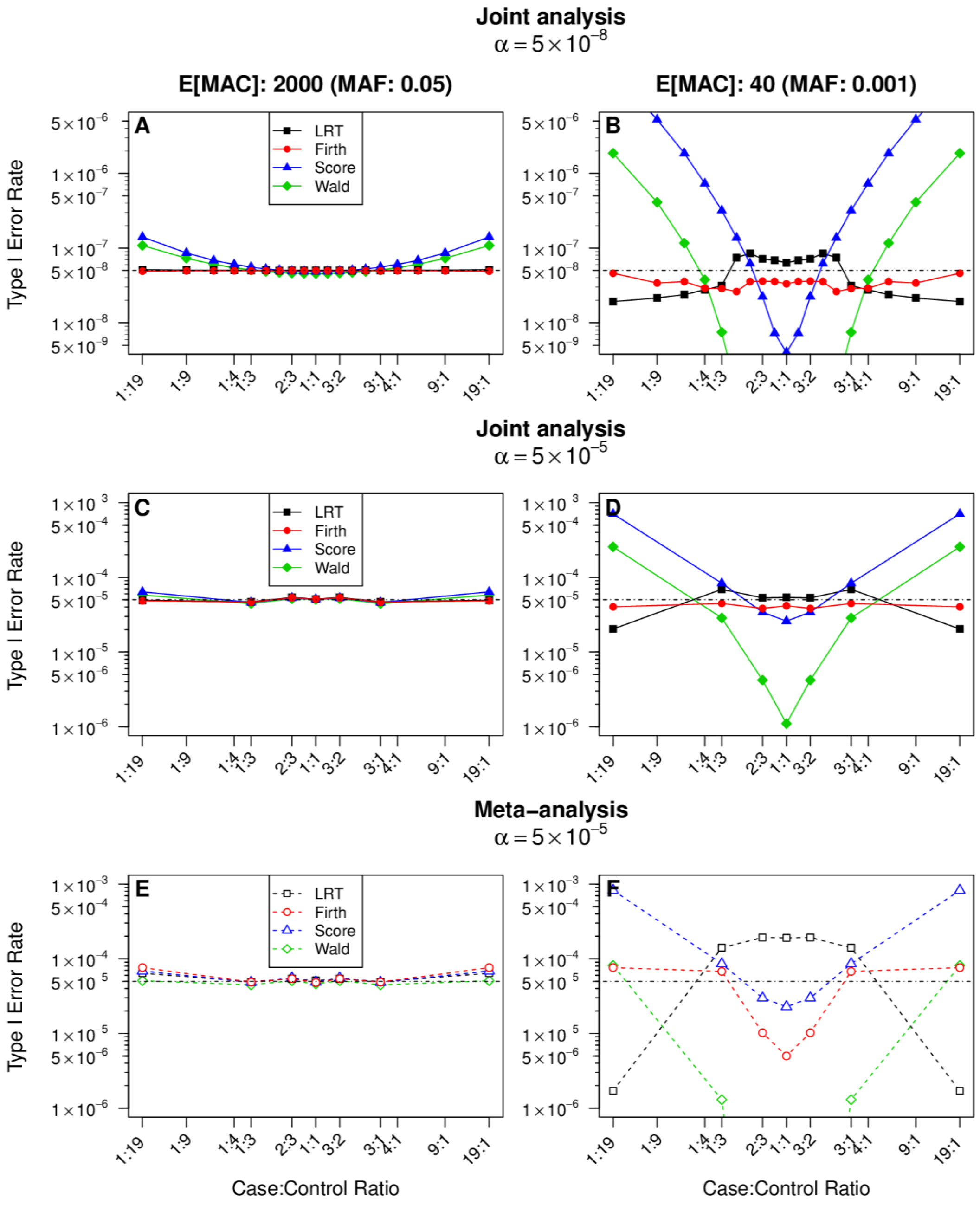


Figure 2.2: Type I error rates by case-control ratio for logistic regression tests in joint and meta-analysis.

(A, B) Analytically calculated type I error rates ($\alpha = 5\times10^{-8}$) for joint analysis; (C, D) empirical type I error rates ($\alpha = 5\times10^{-5}$) for joint analysis; and (E, F) empirical type I error rates ($\alpha = 5\times10^{-5}$) for sample-size weighted meta-analysis. Type I error rates are estimated for a high count (expected MAC = 2000; MAF = 0.05), and low count (E[MAC] = 40; MAF = 0.001) variant, in studies with N = 20000 individuals and varying case-control ratios. The horizontal dotted line denotes the corresponding nominal significance threshold.



analysis type I error rates at $\alpha = 5\times10^{-8}$ across a wide range of case-control ratios for high count (Figure 2.2A) and low count (Figure 2.2B) variants.

Finally, we examine type I error rates for meta-analysis of 10 unbalanced sub-studies each with 500/1500 (1:3) or 100/1900 (1:19) cases and controls. For high count variants, in a 1:3 study, all meta-analysis tests are well-calibrated (Figure 2.1H); in a 1:19 study, meta-analysis of Firth, score, and likelihood ratio test results can be slightly anti-conservative (Figure 2.1I). For low count variants, all four tests can be highly anti-conservative for specific combinations of allele counts and case-control ratios. For example, at E[MAC] = 40 (MAF = 0.001) in a 1:3 study, meta-analyses of every test except Wald are anti-conservative; in a 1:19 study, all except the likelihood ratio test are anti-conservative. For meta-analysis of studies with case-control ratios more extreme than approximately 2:3 (or 3:2), all tests can be anti-conservative (Figure 2.2F).

Power of joint and meta-analysis tests

We first examine the power ($\alpha = 5\times10^{-8}$) for joint and meta-analysis tests in balanced studies. For high count variants (E[MAC] = 2000; MAF = 0.05), all tests have near identical power for both joint and meta-analysis, as expected [Lin and Zeng, 2010] (Figure 2.3A). For low count variants (E[MAC] = 40; MAF = 0.001), we focus on tests with type I error rates not exceeding the nominal threshold (Figure 2.3D). Comparing joint and meta-analysis, Firth test-based joint analysis (red solid line) is more powerful than score test-based meta-analysis (blue dashed line). Meta-analysis of Wald test results has lowest power among all the tests. These results are consistent with the observation that statistical power often corresponds to relative conservativeness: more conservative tests usually have lower power.

Next we evaluate power for joint and meta-analysis tests in unbalanced studies. For high count variants, again all tests have near identical (1:3 study; Figure 2.3B) or similar (1:19 study; Figure 2.3C) power for both joint and meta-analysis. For low count variants, most power comparisons are not meaningful since all joint and meta-analysis tests except Firth test-based joint analysis can be anti-conservative for specific combinations of allele counts and case-control ratios (Figure 2.3E-F). Nonetheless, we again observe some correspondence between increased test conservativeness and reduced test power in unbalanced studies.

Figure 2.3: Simulation-based power curves for joint and meta-analysis.

Simulated power ($\alpha = 5\times10^{-8}$) in joint analysis (solid) and sample-size weighted meta-analysis (dashed) for (A - C) a high count variant (expected MAC = 2000; MAF = 0.05); and (D - F) a low count variant (E[MAC] = 40; MAF = 0.001). Power for joint analysis is estimated for studies with 10000/10000, 5000/15000, and 1000/19000 total cases and controls; meta-analysis is based on partitioning the full dataset into 10 equal-sized sub-studies.

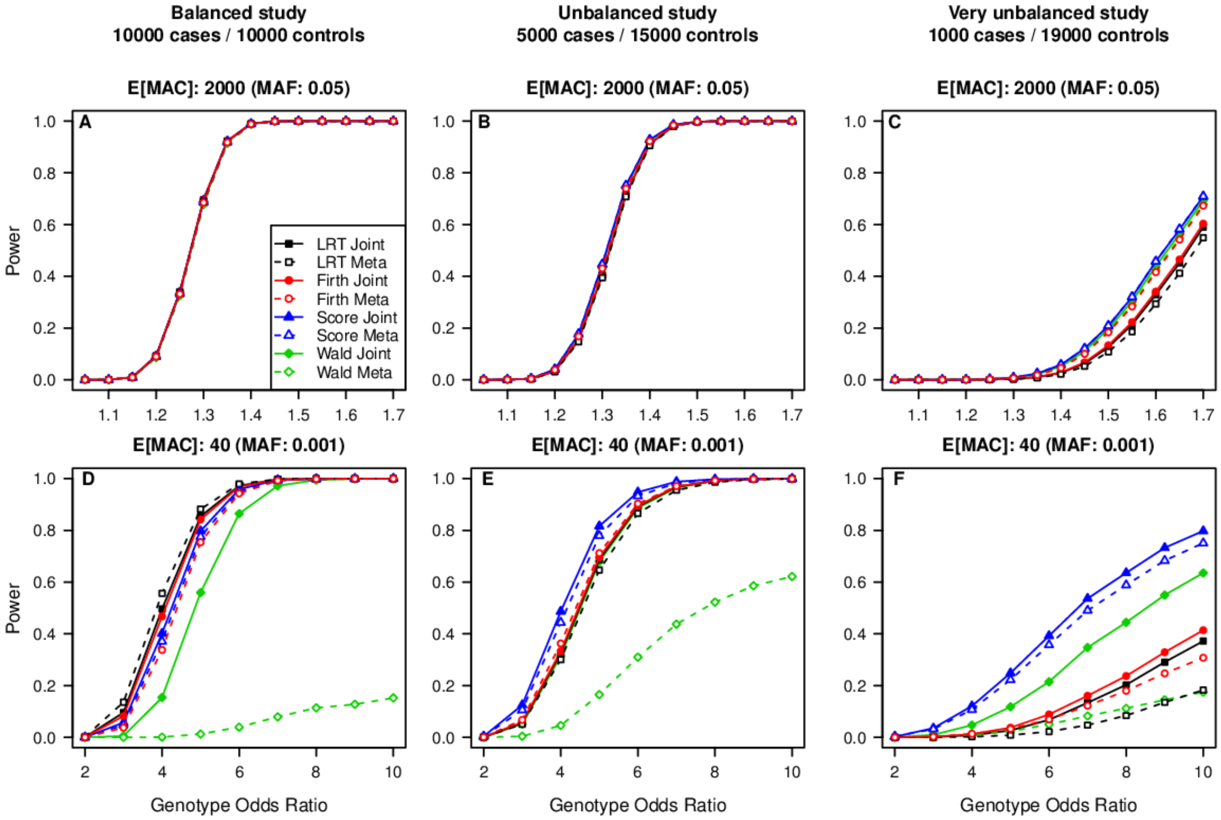
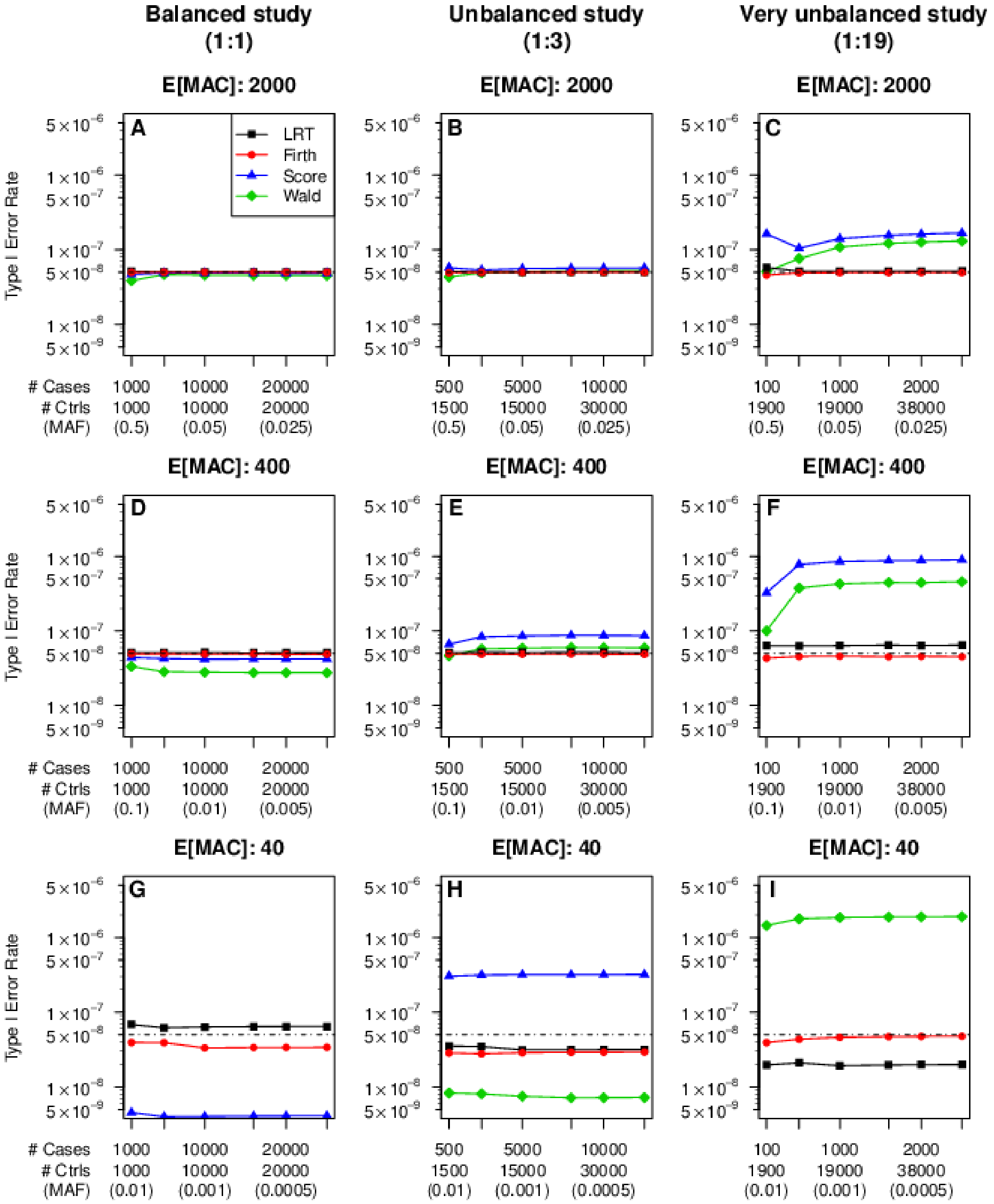


Figure 2.4: Joint analysis type I error rates by sample size for fixed expected minor allele count (MAC).

Analytically calculated joint analysis type I error rates for single balanced (case-control ratio 1:1), unbalanced (1:3), and very unbalanced studies (1:19) of various sample sizes. For each study, variant allele frequencies are selected so that variants have (A - C) expected MAC = 2000; (D -F) expected MAC = 400; or (G - I) expected MAC = 40. The horizontal dotted line denotes the corresponding nominal significance threshold (α = 5×10^{-8}). Very conservative or anti-conservative tests with type I error rates that exceed the vertical axis scale are not displayed.



Consistent test calibration with fixed total MAC

All of the results shown so far (Figure 2.1, 2.2, 2.3) refer to analyses with a total sample size of N = 20000 individuals. Here, we examine joint analysis (Figure 2.4, S2.1; $\alpha = 5\times10^{-8}$) and meta-analysis (Figure S2.2; $\alpha = 5\times10^{-4}$) type I error rates while varying N inversely to MAF, so that the expected MAC remains constant. For each case-control ratio, we observe a remarkable consistency of type I error rates across a broad range of sample sizes (N = 2000 to 50000) and MAF for all four tests in both joint and meta-analysis. The conservative or anti-conservative behavior of each test at a particular MAC, case-control ratio, and choice of joint or meta-analysis is almost invariant to N (given $N > 2000$). This demonstrates that MAC, rather than MAF, is the better index to describe the calibration of each test.

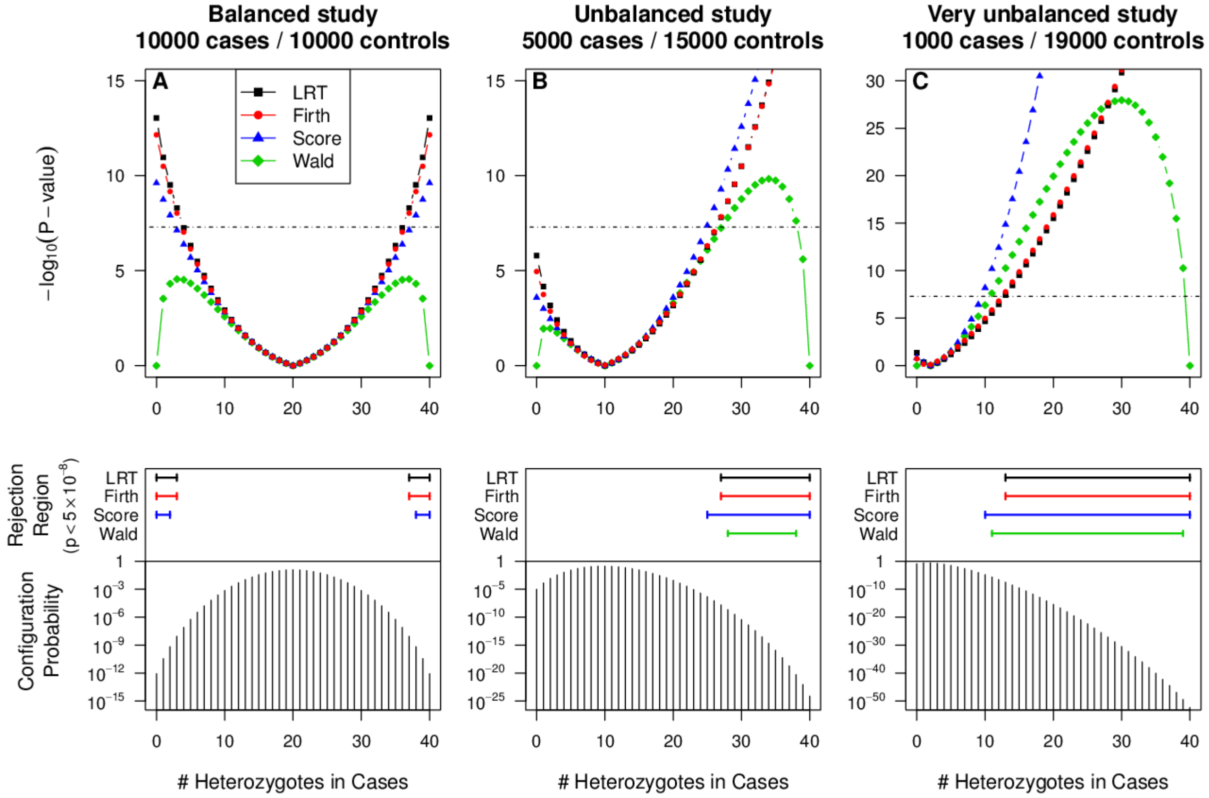
For the study designs we have considered, we find that MAC = 400 is a useful threshold separating high count and low count variants, based on our type I error results in balanced (1:1) and moderately unbalanced (1:3) studies. For variants with MAC < 400, we observe that all joint and meta-analysis tests can have different degrees of conservative or anti-conservative behavior (Figure 2.1). In contrast, for variants with MAC > 400, all tests are generally well-calibrated (for not too imbalanced studies). Hence, our threshold of MAC = 400 provides an approximate, sample size invariant threshold distinguishing high and low count variants, and a rule-of-thumb guideline for test selection. However, a more stringent MAC threshold may be needed for studies with more extreme case-control imbalance.

Detailed comparison of the four logistic regression tests

Our results show that the logistic regression tests, while asymptotically equivalent, are not equivalent when testing low count variants at stringent significance thresholds, even with large sample sizes. To understand the observed patterns of type I error rate and power for a low count variant (expected MAC = 40), we compare joint analysis test p-values for all possible case-control configurations for a variant with MAC = 40 in a study of N = 20000 individuals (Figure 2.5, upper panels). In Figure 2.5 (lower panels), horizontal bars denote the rejection region for each test at a nominal significance threshold of 5×10^{-8}, and the histogram displays hypergeometric probabilities for each MAC configuration. Tests with rejection regions containing configurations with greater total probability have higher type I error rates and power (averaged across all sampled MACs).

Figure 2.5: Logistic regression p-value distributions for fixed total minor allele count (MAC).

For a variant with MAC = 40, the upper panels display p-values for all 41 possible allele configurations for each test in a single study of (A) 10000/10000, (B) 5000/15000, and (C) 1000/19000 cases and controls, respectively. The horizontal dotted line denotes the corresponding nominal significance threshold (α = 5×10^{-8}). The lower panels display horizontal bars indicating the rejection region (p-value < 5×10^{-8}) for each test and hypergeometric probabilities of each allele configuration.



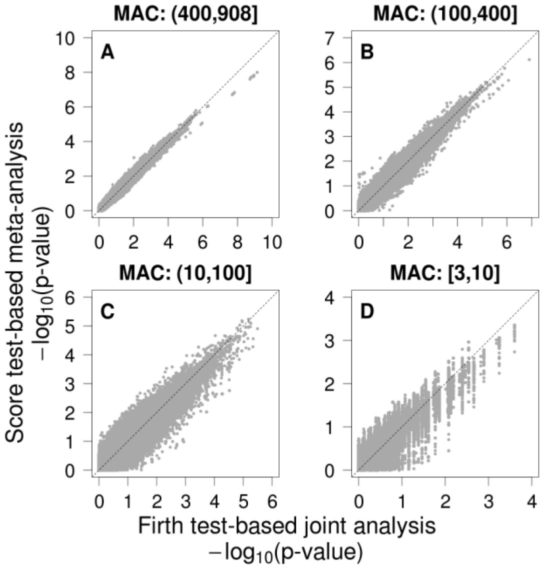
For a balanced study, at the low and high extremes of case MAC, the likelihood ratio test has the most significant p-values at each MAC, followed by the Firth, score, and Wald test p-values (Figure 2.5A, upper panel). The rejection regions contain the most probability for the likelihood ratio and Firth tests, less for the score test, and none for the Wald test (Figure 2.5A, lower panel). When other MACs consistent with an expected MAC of 40 are considered, the likelihood ratio test has the largest probability in the rejection region (data not shown). Tests with the highest to lowest type I error rates (likelihood ratio, Firth, score, Wald) (Figure 2.1A) mirror the observed trend for the rejection regions.

For an unbalanced (1:19) study, in configurations with 10-25 heterozygotes in cases, we observe the score, Wald, Firth, and likelihood ratio tests in order of decreasing significance (Figure 2.5C, upper panel). Again, this corresponds to the total configuration probability encompassed by the rejection regions (Figure 2.5C, lower panel), and the least to most conservative tests (Figure 2.1C), averaged across the sampled MACs.

In both balanced and unbalanced studies, the Wald test has substantially less significant p-values for configurations with zero or few alleles in either cases or controls (that is, (nearly) separated data), and thus has little or no power to detect the strongest associations. This unfortunate property of the Wald test is exacerbated in meta-analysis since each contributing study has a much smaller total MAC. As such, meta-analysis of Wald test results has extremely low power (green dashed line; Figure 2.3D-F) and should not be used.

Figure 2.6: Comparison of score test-based meta-analysis and Firth test-based joint analysis p-values in the GoT2D study.

For different minor allele count (MAC) categories, comparison of score test-based meta-analysis and Firth test-based joint analysis p-values.



Comparison of tests in joint and meta-analysis of GoT2D data

We analyzed preliminary low-pass sequencing data from an early data freeze of the GoT2D study to examine the differences between statistical tests in joint and meta-analysis. The dataset is comprised of three Northern European studies and is nearly balanced (N = 908; 499/409 cases/controls), with an overall case-control ratio of 1.22. We focus on the tests with the best combination of type I error and power in balanced studies: Firth test-based joint analysis and score test-based meta-analysis. We analyzed 8.58 million variants with MAC ≥ 3.

For high count variants (400 < MAC ≤ 908), score test-based meta-analysis and Firth test-based joint analysis produce similar p-values (Figure 2.6A). For low count variants (MAC < 400), Firth test-based joint analysis p-values are typically more significant than score test-based meta-analysis p-values, especially for the rarest variants (Figure 2.6B-D). These patterns are consistent with our analytic and simulation-based results. Additional comparisons between joint and meta-analysis test p-values can be found in supplemental materials (Figure S2.3, S2.4).

Discussion

Recommendations

For analysis of high count variants (MAC > 400), in balanced and moderately unbalanced (1:3) studies, joint and meta-analysis using any of the asymptotic tests have near nominal type I error rates and comparable power, so either joint or meta-analysis using any of the asymptotic tests can be recommended. For low count variants (MAC < 400), type I error rates and power can vary widely for different tests, MACs, and case-control ratios.

For low count variants, in balanced studies, joint analysis using the Firth test is best, and meta-analysis using the score test results is best, with (Firth test-based) joint analysis being more powerful than (score test-based) meta-analysis. In unbalanced studies, again joint analysis using the Firth test is best, but for meta-analysis, all tests can be (very) anti-conservative for many combinations of allele count and case-control ratio. If individual-level data are available for analysis, we recommend joint analysis using Firth bias-corrected logistic regression in both balanced and unbalanced studies. If not, we recommend meta-analysis of score test results for analysis of balanced and not-too-unbalanced studies. For meta-analysis of unbalanced studies with case-control ratio < 2:3

or > 3:2, none of the statistical tests considered can be recommended due to the inflated type I error rates, and the score test, in particular, is not recommended.

Use of MAC rather than MAF in describing test calibration

We present our recommendations using a rough MAC threshold, rather than a MAF threshold, since test calibration remains consistent as long as MAC is constant (given *N* > 2000, a consistent analytic strategy, and uniform scaling of *N* across studies in meta-analysis). We show that MAC = 400 is a threshold below which tests may begin to deviate substantially from the nominal significance threshold in balanced to moderately unbalanced studies. Investigators studying variants with MAC < 400 should take care in selecting an association test for analysis.

This MAC threshold is reminiscent of Yates' classic guideline for expected values in 2×2 contingency tables, which states that the χ^2 approximation is sufficiently accurate if each expected cell count ≥ 5 [Yates, 1934]. In the context of GWAS, we require a much larger minimum (marginal) cell count threshold since we are testing at considerably more stringent significance thresholds than envisioned by Yates.

Practical recommendations for meta-analysis

For meta-analysis, we recommend analyzing all variants with MAC ≥ 1 within each sub-study, since even variants with a single observed minor allele contribute to the overall meta-analysis. Imposing a more stringent study-level MAC filter leads to more conservative and less powerful meta-analysis results (Figure S2.5). When assessing the performance of a given meta-analysis using Quantile-Quantile (Q-Q) plots, it may be useful to apply a minimum total combined MAC threshold (say MAC ≥ 15 or 20), since the rarest variants are unlikely to attain genome-wide significance ($\alpha < 5\times10^{-8}$). For a given fixed total *N*, we observe that meta-analysis of many small sub-studies is more conservative and less powerful than meta-analysis of a few larger sub-studies (Figure S2.6). Smaller sub-studies are more likely to be monomorphic for low count variants, and so are effectively removed from the meta-analysis.

Study limitations and caveats

In this paper, we did not present meta-analysis of sets of studies with varying sample sizes and case-control ratios, although limited simulations in such settings suggested conclusions consistent with those presented (data not shown). Nor did we assess the effects of population stratification. Although joint analysis can be more powerful than meta-analysis for low frequency variants, for a dataset comprised of divergent samples, it may be difficult to control for specific within-sample confounding using the same covariates across all studies.

For simplicity, we did not include study covariates in the simulations described. Limited simulations including covariates independent of disease status or study indicators for joint analysis gave results consistent with those reported for both high count and low count variants (data not shown). We did explore the effect of covariate adjustment in the GoT2D data analysis, including age, sex, and three principal components for ancestry. The comparison between Firth test-based joint analysis and score test-based meta-analysis is similar to those shown in Figure 2.6, but covariate adjustment results in modestly increased differences between the p-values. However, for a very small number of low count variants, we observe large differences in p-values after adjustment for continuous covariates (i.e. age and principal components), especially for the score test.

While some simulation parameters may not reflect observed parameters in real datasets, our goal is to explore a wide range of parameters to illustrate the conclusions. For example, our very unbalanced (1:19) scenario is more imbalanced than expected under random sampling for the disease prevalence 10%. However, we wanted to explore the effect of extreme case-control imbalance, similar to those observed for population-based case-control studies of type 2 diabetes such as deCODE (1:16) [Steinthorsdottir et al., 2007]. Additional simulations demonstrate that type I error rates are consistent across prevalence rates of 1%, 10%, and 50% [data not shown].

For low count variants, we present results based on large ORs to illustrate the differences in power between the different joint and meta-analysis tests, and to emphasize the low power of Wald test-based meta-analysis even for very large ORs. However, finding variants with such large ORs is unlikely in complex diseases. Finally, we assess meta-analysis type I error rates at less stringent significance thresholds ($\alpha = 5 \times 10^{-4}$ and 5×10^{-5}) owing to computational limitations; we expect results to be similar at $\alpha = 5 \times 10^{-8}$.

Alternative analysis strategies

We explored several alternative analysis strategies for low count variants, with a particular focus on meta-analysis of unbalanced studies since standard methods are generally anti-conservative. First, we derived bias-corrected versions of the score and Wald tests; simulations show that these tests are also anti-conservative in meta-analysis of unbalanced studies (data not shown). Second, we considered exact logistic regression [Mehta and Patel, 1995], which evaluates significance based on the permutation distribution of sufficient statistics, but it is not useful in our context since it cannot adjust for continuous covariates and is computationally prohibitive for large sample sizes. Third, we evaluated Fisher's exact test (FET), which uses the hypergeometric distribution to test the significance of contingency tables (Figure S2.7, S2.8, S2.9), but since FET cannot adjust for covariates, it is not practical in actual data analysis. Fourth, we investigated using linear regression, treating the binary phenotype as a continuous outcome; linear regression produces nearly identical p-values as logistic regression score test, and thus is equally anti-conservative in unbalanced studies (data not shown).

Fifth, we examined meta-analysis with inverse variance weights (supplemental methods in Appendix 2.1); simulations show that inverse-variance weighted meta-analysis of Firth or Wald test results in unbalanced studies are also anti-conservative (Figure S2.7, S2.8, S2.9). Sixth, we explored fixed effects meta-analysis with sample size weights accounting for allele frequency ($\sqrt{\bar{N}_k p_k (1-p_k)}$) . These weights do not substantially affect simulated type I error rates or power since the expected MAF for each sub-study is identical in our simulations. If the underlying MAFs are different between studies, weights including allele frequency may result in higher power [Han and Eskin, 2011]. Seventh, we considered random effects meta-analysis [Dersimonian and Laird, 1986]. As expected, it is more conservative and less powerful than fixed effects meta-analysis (data not shown).

Eighth, we evaluated the strategy of randomly removing cases or controls from a highly unbalanced study to reduce the case-control imbalance. We find that this strategy can substantially decrease power. For example, in a study with 2000 cases and 18000 controls, randomly removing 12000 controls reduces score test-based joint analysis power for a variant with E[MAC] = 40 and OR = 5 from 49% in the full samples to 13% in the reduced sample.

Finally, we developed a "screen and permute" strategy in which we analyze all variants using a liberal test (for example, the likelihood ratio test), and perform case-control permutations of the strongest associated variants to compute empirical p-values. However, sample-size weighted meta-analysis of permuted p-values in unbalanced studies remains anti-conservative, even though study-level permuted p-values are conservative. In theory, permutation testing should always be well-calibrated, but this proposed strategy applies permutation only within individual studies. For each variant, the ideal permutation-based meta-analysis method is to compute millions of permutation p-values for each of the K studies, calculate the null distribution of meta-analysis p-values, and compare the observed meta-analysis p-value against this null distribution. While this strategy should work, it is practically infeasible since we would need to share millions of permuted p-values for each screened variant in every study.

Summary

In this study, we extend Lin and Zeng's [2010] evaluation of type I error and power in joint and meta-analysis for logistic regression tests to low count variants in balanced and unbalanced studies. When testing at a combination of three extremes: low MAC, stringent significance thresholds, and large case-control imbalance, asymptotic assumptions for standard tests and aggregation methods are not valid, leading to differences in type I error rate and power among the tests even for large sample sizes. For low count variants, we identify the Firth test as best for joint analysis in both balanced and unbalanced studies, and the score test as best for meta-analysis in balanced studies only. We show that Firth test-based joint analysis is more powerful than score test-based meta-analysis. We establish MAC as a sample-size invariant and consistent measure of test calibration and variant information. For balanced and moderately unbalanced studies, MAC = 400 is a practical threshold below which test calibration begins to diverge from the nominal significance threshold; a more stringent MAC threshold may be needed for very unbalanced studies. Further investigation is needed to identify a well-calibrated and powerful test for meta-analysis of unbalanced studies, since all tests evaluated can be anti-conservative.

Appendix

Appendix 2.1: Inverse variance weighted meta-analysis

Using study-level estimates of effect size and its variance, inverse variance weighted meta-analysis estimates a pooled effect size, its standard error, and the corresponding z-score:

$$\overline{\beta}_{IV} = \sum_{k=1}^{K} V_k^{-1}\beta_k \Big/ \sum_{k=1}^{K} V_k^{-1}\ ; \quad SE\left(\overline{\beta}_{IV}\right) = \left[\sqrt{\sum_{k=1}^{K} V_k^{-1}}\right]^{-1}\ ; \quad Z_{IV} = \overline{\beta}_{IV} / SE\left(\overline{\beta}_{IV}\right)$$

This method is only applicable for statistical tests that estimate those parameters, and so cannot be used for the score test or Fisher’s exact test.

Supplementary Figures

Figure S2.1: Type I error rates by fixed expected minor allele count (MAC) for different sample sizes.

Analytically calculated type I error rates (α = 5×10-8) for joint analysis in: balanced studies (A - C), unbalanced studies (D - F), and very unbalanced studies (G - I). Variant allele frequencies are selected so that the expected MAC remains constant across studies with total sample size N = 2000, 20000 and 50000 individuals respectively. The horizontal dotted line denotes the corresponding nominal significance threshold ($\alpha = 5x10^{-8}$).

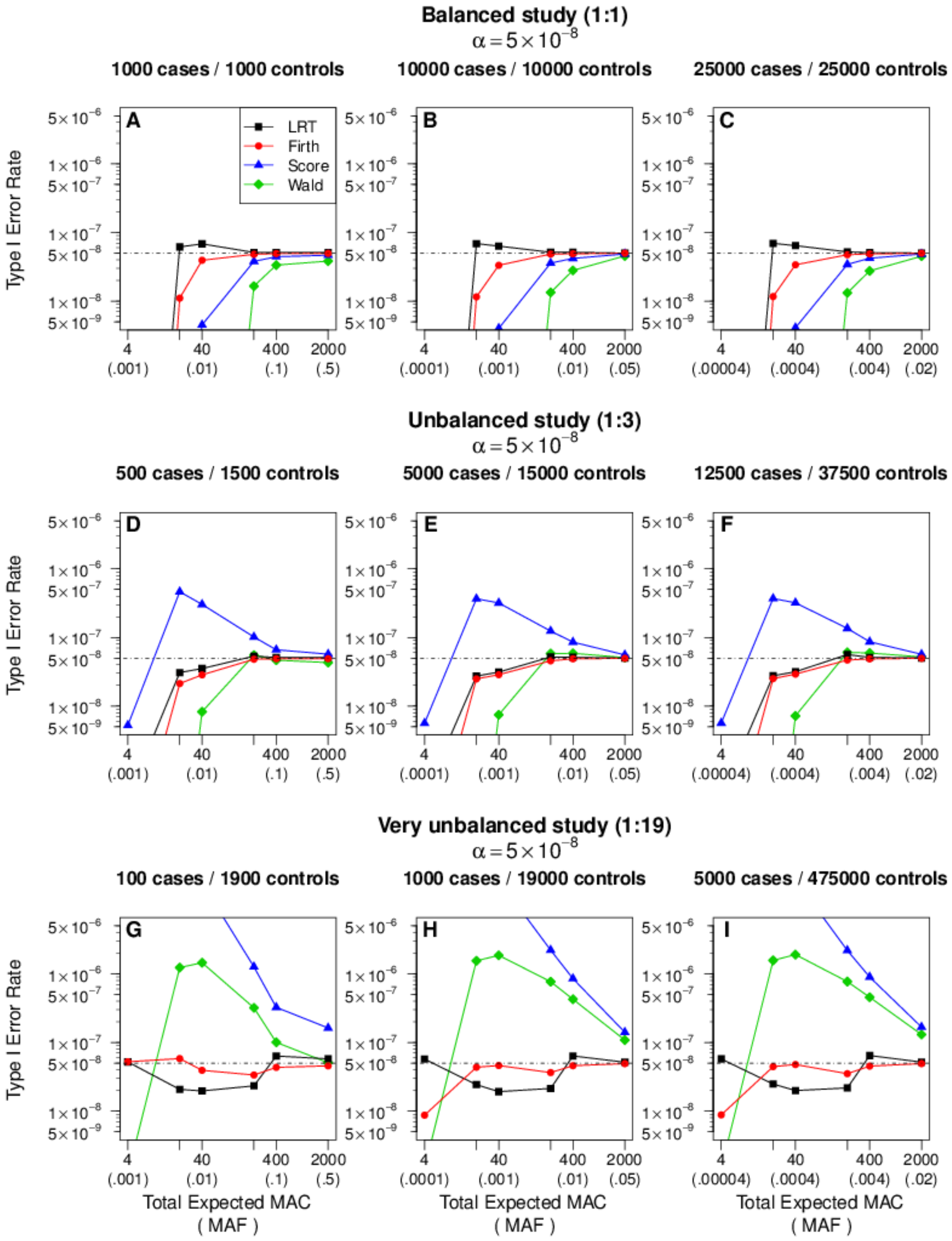


Figure S2.2: Meta-analysis type I error rates by sample size for fixed expected minor allele count (MAC).

Simulation-based sample-size weighted meta-analysis type I error rates ($\alpha = 5\times10^{-4}$) for balanced (case-control ratio 1:1), unbalanced (1:3), and very unbalanced studies (1:19) with of various sample sizes. For each study, variant allele frequencies are selected so that the expected MAC = 2000 (A - C), 400 (D - F), or 40 (G - I). The horizontal dotted line denotes the corresponding nominal significance threshold ($\alpha = 5\times10^{-4}$). Very conservative or anti-conservative tests with type I error rates that exceed the vertical axis limits are not displayed.

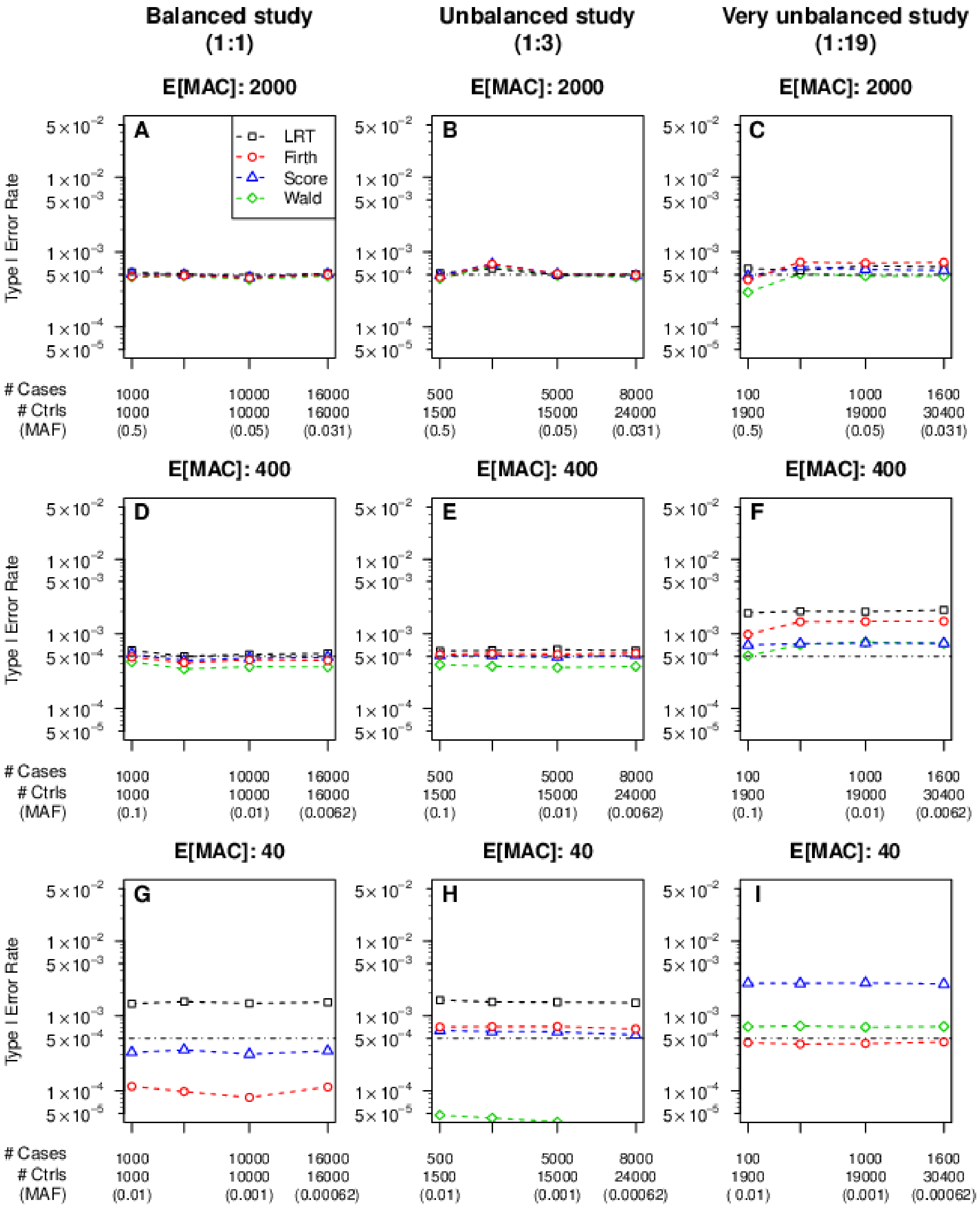


Figure S2.3: Comparison of score and Firth test association p-values in the GoT2D study.

For different minor allele count (MAC) categories, comparison of score and Firth test-based (A-D) joint analysis p-values and (E-H) meta-analysis p-values.

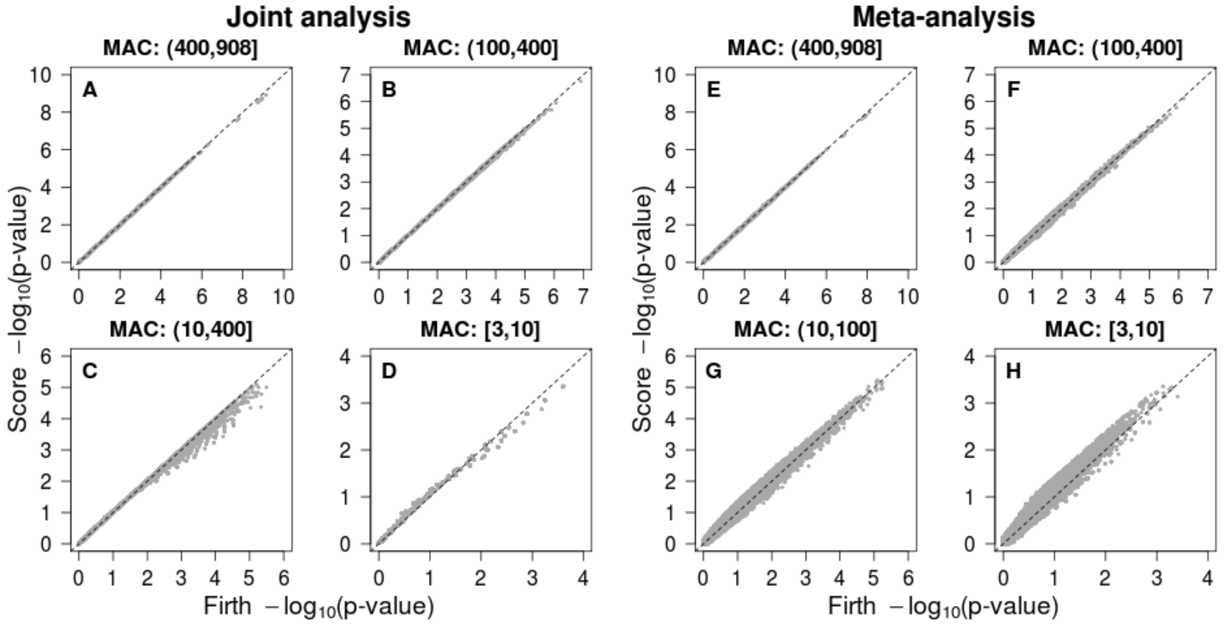


Figure S2.4: Comparison of joint and meta-analysis p-values in the GoT2D study.

For different minor allele count (MAC) categories, comparison of joint and meta-analysis p-values using the (A-D) Firth test and (E-H) score test.

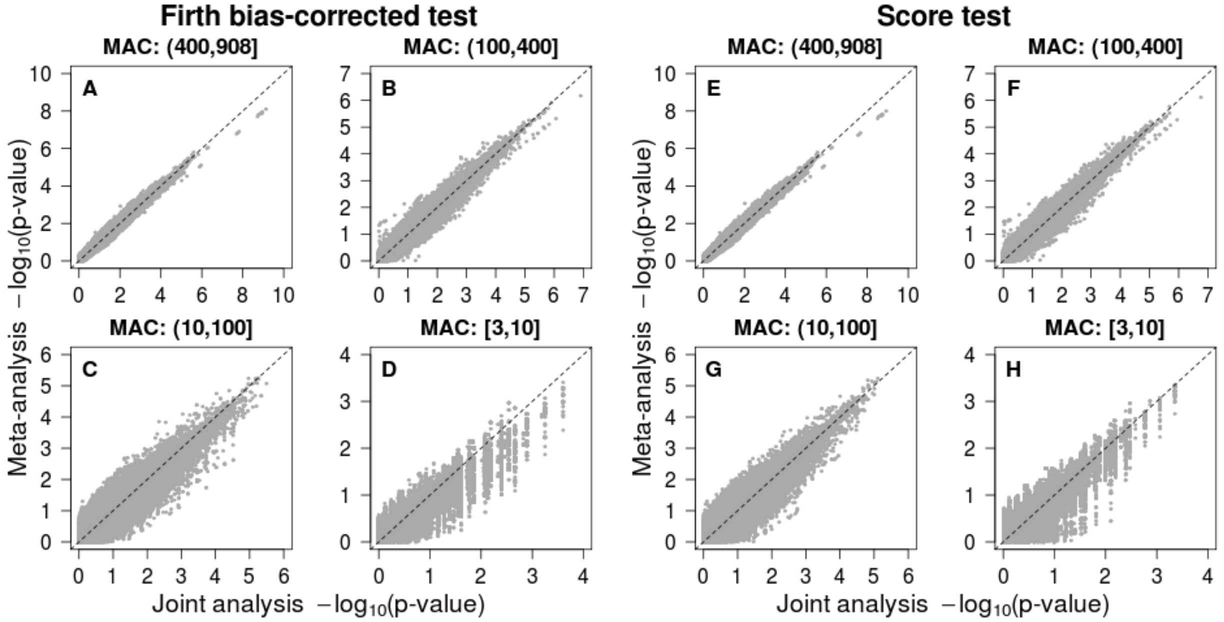


Figure S2.5: Score test type I error rate and power with study-level minor allele count (MAC) filters.

(A) Empirical type I error rates (α = 5×10^{-5}) for score test-based joint and sample-size weighted meta-analysis, with varying degrees of study-level MAC filters. Type I error rates for joint analysis are estimated for studies with 10000/10000 total cases and controls; meta-analysis is based on partitioning the full dataset into 10 equal-sized sub-studies. The horizontal dotted line denotes the corresponding nominal significance threshold. (B - C) Simulated power at α = 5×10^{-8} for a variant with: expected MAC = 40 (MAF = 0.001); and E[MAC] = 20 (MAF = 0.0005), for the same study design.

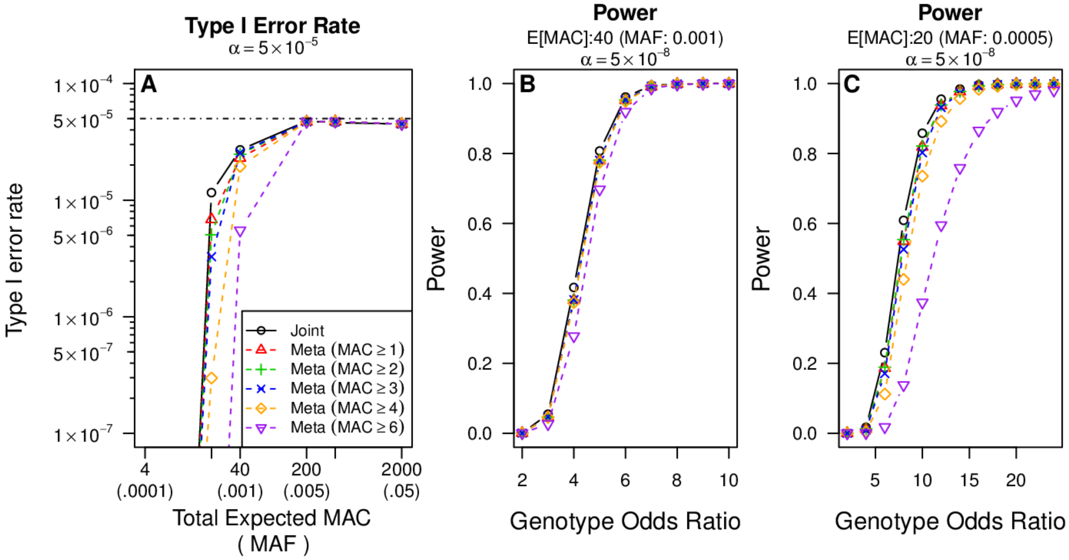


Figure S2.6: Score test type I error rate and power curves for meta-analysis of *K* = 10 and 50 sub-studies.

(A) Empirical type I error rates ($\alpha = 5\times10^{-5}$) for score test-based joint analysis with 10000/10000 total cases and controls (black); sample-size weighted meta-analysis with *K* = 10 sub-studies of 1000/1000 cases and controls (red); and *K* = 50 sub-studies of 200/200 cases and controls (green). (B) Simulated power ($\alpha = 5\times10^{-8}$) for a variant with expected minor allele count = 40 (MAF = 0.001) for the same study design.

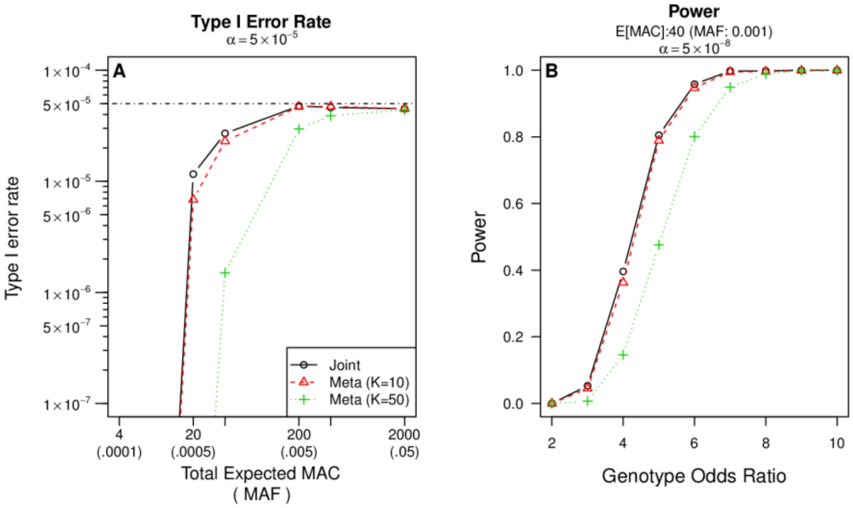


Figure S2.7: Type I error rates by minor allele count (MAC) for logistic regression tests and Fisher's exact test in joint and meta-analysis.
(A - C) Analytically calculated type I error rates (α = 5x10^{-8}) for joint analysis; (D - F) empirical type I error rates (α = 5x10^{-5}) for joint analysis; and (G - I) empirical type I error rates (α = 5x10-^{5}) for sample-size weighted (dashed) and inverse-variance weighted (dotted) meta-analysis. Type I error rates for joint analysis are estimated for studies with 10000/10000, 5000/15000 and 1000/19000 total cases and controls; meta-analysis is based on partitioning the full dataset into 10 equal-sized sub-studies. The horizontal dotted line denotes the corresponding nominal significance threshold.

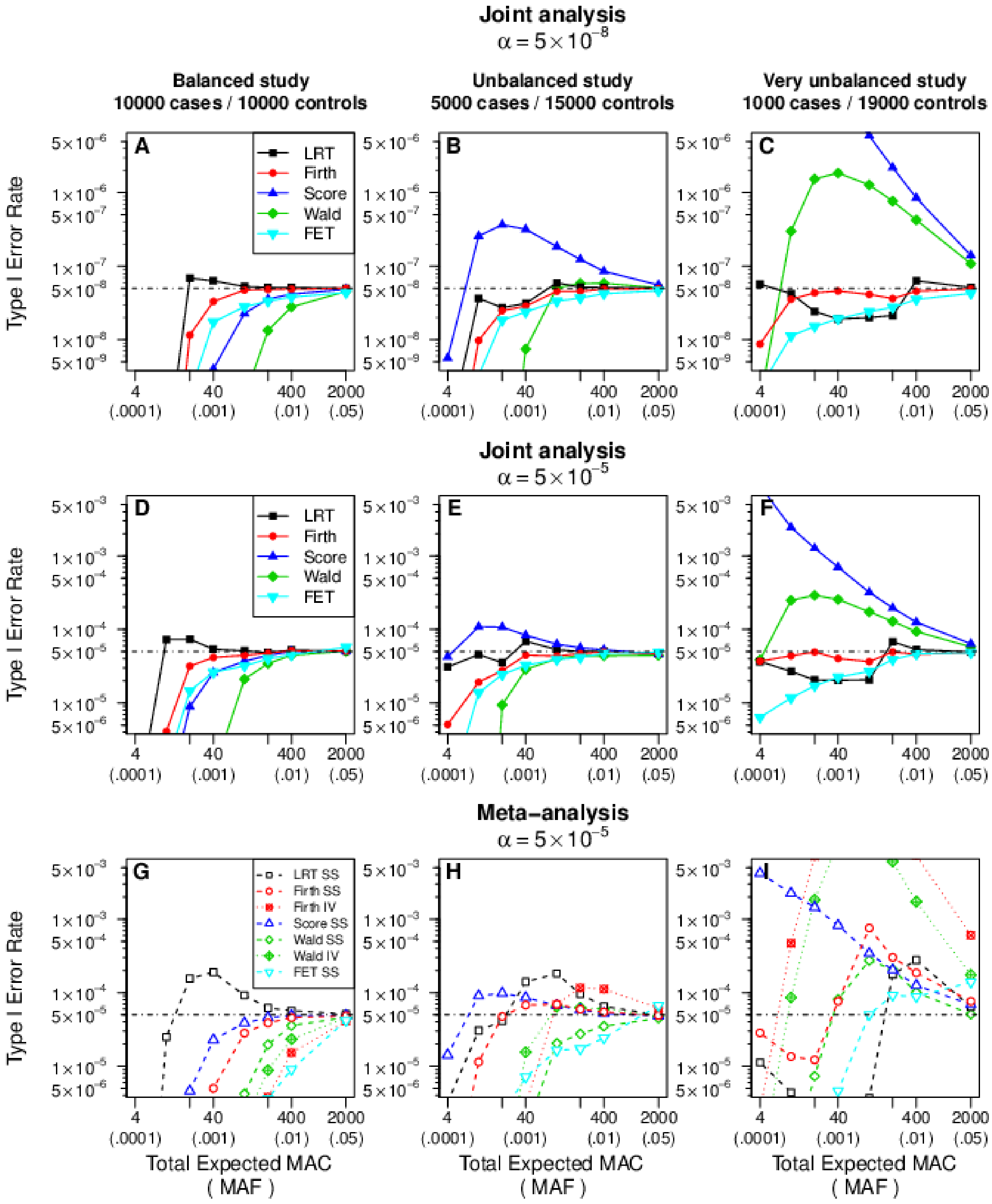


Figure S2.8: Type I error rates by case-control ratio for logistic regression and Fisher's exact tests in joint and meta-analysis.

(A, B) Analytically calculated type I error rates (α = $5x10^{-8}$) for joint analysis; (C, D) empirical type I error rates (α = $5x10^{-5}$) for joint analysis; and (E, F) empirical type I error rates (α = 5x10-5) for sample-size weighted (dashed) and inverse-variance weighted (dotted) meta-analysis. Type I error rates are estimated for a high count (expected MAC = 2000; MAF = 0.05), and low count (E[MAC] = 40; MAF = 0.001) variant, in studies with N = 20000 individuals with varying case-control ratios. The horizontal dotted line denotes the corresponding nominal significance threshold.

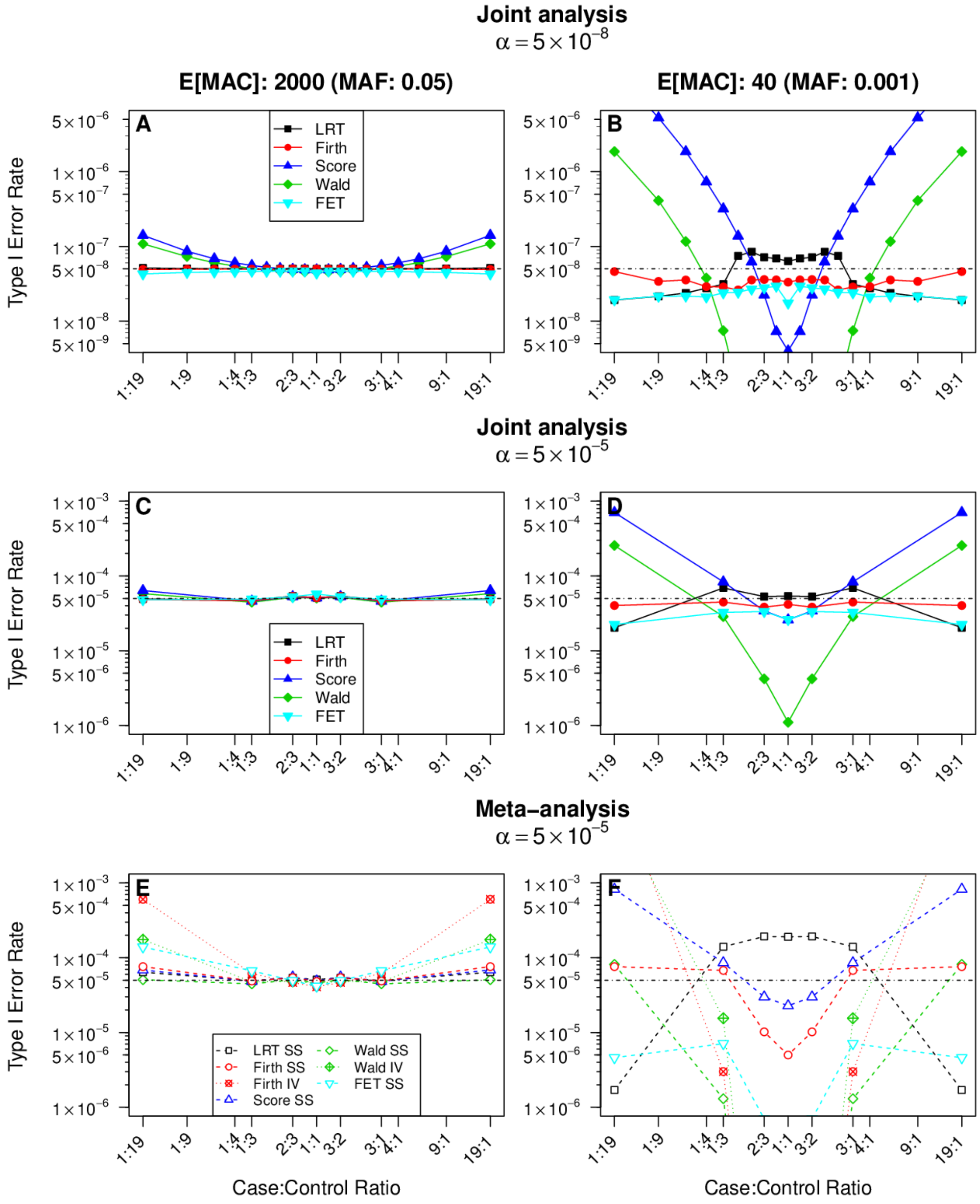
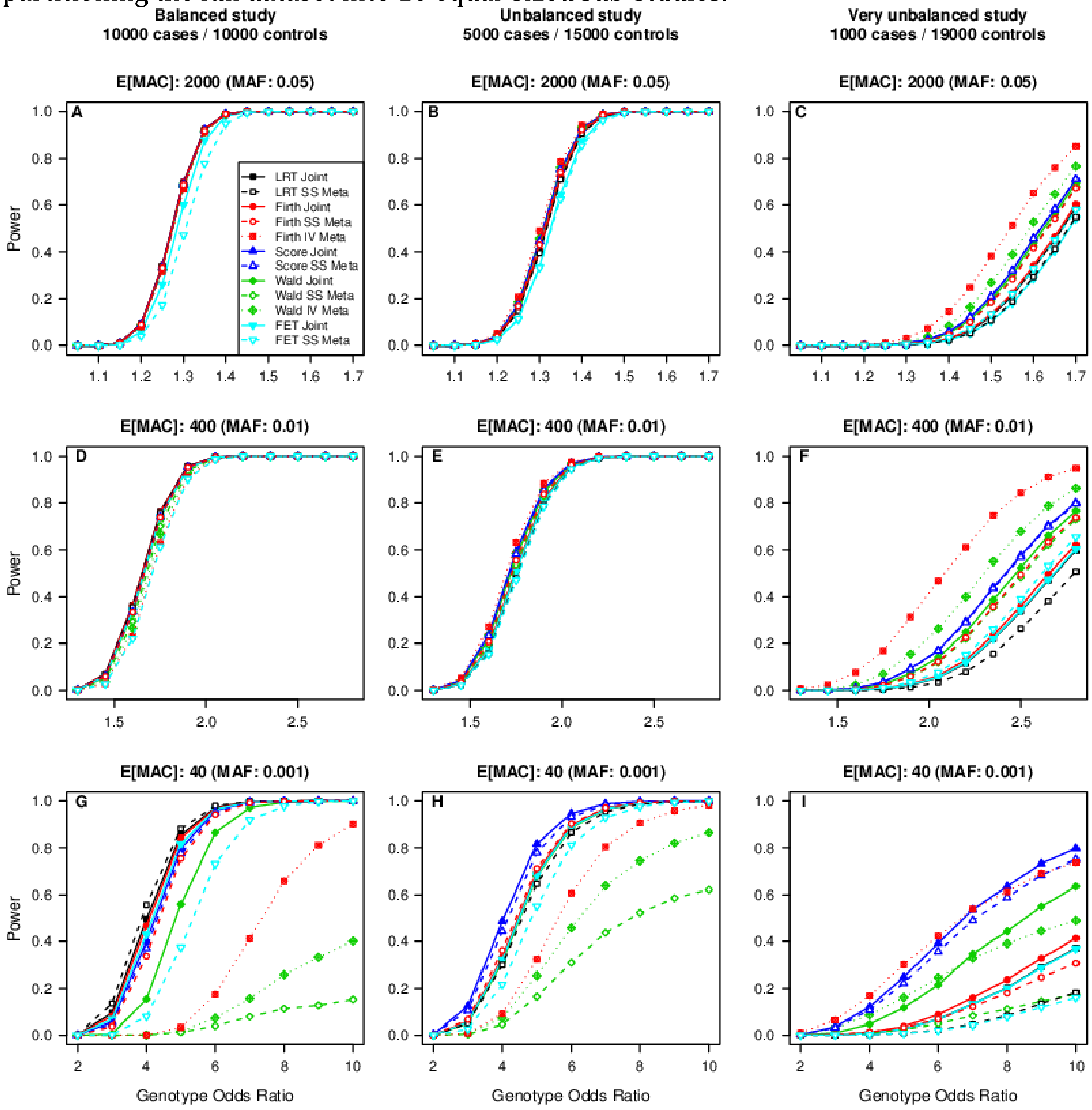


Figure S2.9: Simulated power curves for joint and meta-analysis.
Simulated power ($\alpha = 5\times10^{-8}$) in joint analysis (solid), sample-size weighted (dashed) and inverse-variance weighted (dotted) meta-analysis for a variant with: (A - C) expected MAC = 2000 (MAF = 0.05); (D - F) expected MAC = 400 (MAF = 0.01); and (G - I) expected MAC = 40 (MAF = 0.001). Power for joint analysis is estimated for studies with 10000/10000, 5000/15000, and 1000/19000 total cases and controls; meta-analysis is based on partitioning the full dataset into 10 equal-sized sub-studies.



Chapter 3: Near equivalent calibration and power of joint and meta-analysis for association analysis of quantitative traits

Introduction

Genome-wide association studies (GWAS) have identified thousands of common genetic variants (minor allele frequency [MAF] ≥ 5%) associated with hundreds of human diseases and traits [Hindorff et al., 2012]. For association analysis of quantitative traits (QTs), investigators typically test each variant for association using linear regression [Diabetes Genetics Initiative et al., 2007]. However, individual association studies are often too small to detect modest genetic effects. To combine information across multiple studies, meta-analysis of study-level association results or joint analysis of the combined individual-level data can be used [Scott et al., 2007; Schizophrenia Psychiatric Genome-Wide Association Study Consortium, 2011]. Meta-analysis is frequently used since there are fewer logistical constraints in sharing summary-level data, and because meta-analysis is only slightly less powerful than joint analysis for association analysis of common variants [Lin and Zeng, 2010].

While GWAS arrays generally have limited coverage of low-frequency variants [Zeggini et al., 2005], whole-genome or whole-exome sequencing, dense-array genotyping [Voight et al., 2012], and imputation using sequence-based reference panels [The 1000 Genomes Project Consortium, 2010] provide opportunities to identify trait-associated low-frequency variants [Huyghe et al., 2013; Steinthorsdottir et al., 2014; Wang et al., 2014b]. For the analysis of low-frequency variants (MAF < 5%), single-marker tests (e.g. linear regression) allows detection of individual low-frequency variants with strong effect on the QT. However, asymptotic assumptions for joint and meta-analysis of linear regression may no longer hold for analysis of low-frequency variants, as shown in a similar investigation for logistic regression test-based joint and meta-analysis of binary traits [Ma et al., 2013].

Specifically, logistic regression analysis of variants with low minor allele count [MAC] can be poorly calibrated, and joint analysis can be more powerful than meta-analysis in certain scenarios. The calibration of linear regression and the relative power of joint and meta-analysis has not been investigated in the context of association analysis of low-frequency variants.

In this study, we evaluate the calibration and power of linear regression in joint and meta-analysis for QT association testing. Using simulation, we estimate type I error rates and power for studies of different sample sizes, for normally and non-normally distributed QTs, and with and without additional covariates. For QTs inverse-normalized within study, for variants with expected minor allele count E[MAC] ≥ 10, we show that linear regression-based joint and meta-analysis are near-equivalently well-calibrated and powered, but inverse-variance weighted meta-analysis of small-sized studies can be slightly anti-conservative; for E[MAC] < 10, joint and meta-analysis are increasingly conservative. For non-normally distributed QTs, joint and meta-analysis are equally very anti-conservative, but this calibration problem can be resolved by inverse-normal transformation of the QT. However, applying inverse-normal transformation to QTs of any distribution reduces power for low-frequency variants. These results continue to hold when adjusting for covariates.

Subjects and Methods

Notation

First, consider a single study of *N* individuals. For individual *i*, let y_i denote the QT value, $G_i = \{0,1,2\}$ the number of minor alleles for a specific variant, and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ic})^T$ the vector of *c* covariates. For simplicity, assume an additive genetic model.

Linear regression

We assume the linear regression model is:

$$y_i = \alpha + \beta G_i + \gamma^T Z_i + \varepsilon_i \qquad \text{(Equation 3.1)}$$

where α is the intercept, β is the regression coefficient for the genetic effect, γ is a vector of regression coefficients, and ε_i is normally distributed with mean zero, and variance σ^2. We wish to test the null hypothesis of no genetic association: H$_0$: $\beta = 0$. The usual linear regression partial F-test statistic is:

$$F = \frac{SSE_0 - SSE_1}{SSE_1 / (N - c - 2)} \qquad \text{(Equation 3.2)}$$

where the residual sums of squares under the null and alternative hypotheses are:

$$SSE_0 = \sum_{i=1}^{N} \left(y_i - (\tilde{\alpha} + \tilde{\gamma}^T Z_i) \right)^2$$

$$SSE_1 = \sum_{i=1}^{N} \left(y_i - (\hat{\alpha} + \hat{\beta} G_i + \hat{\gamma}^T Z_i) \right)^2$$

and $(\tilde{\alpha}, \tilde{\gamma})$ and $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ are the least squares (and maximum likelihood) parameter estimates under the null and alternative hypotheses, respectively. This *F* statistic is evaluated relative to the *F* distribution on *1* and *N-c-2* degrees of freedom.

Combining data across studies in joint and meta-analysis

Next, consider *K* studies in which study *k* is comprised of N_k individuals. For joint analysis, we analyze the individual-level genotype and phenotype data combined across all *K* studies (with $N = \sum_k N_k$ individuals) using the same linear regression partial F test (Equation 3.2). To account for population structure between studies, we can adjust for population-specific covariates, such as study indicators or principal components [Price et al., 2006], in the linear regression model (Equation 3.1).

In meta-analysis, we analyze each study separately using linear regression, and combine study-level association results. Two commonly used meta-analysis methods are: (1) sample-size weighted meta-analysis, and (2) inverse-variance weighted meta-analysis.

The sample-size weighted meta-analysis test statistics is:

$$Z_{SS} = \sum_{k=1}^{K} \sqrt{N_k} Z_k \Big/ \sqrt{\sum_{k=1}^{K} N_k} \qquad \text{(Equation 3.3)}$$

where the study-specific z-score $Z_k = \pm\sqrt{q_k}$ is calculated from the corresponding quantile q_k of the χ_1^2 distribution with upper tail probability equal to the association p-value, and the sign is based on direction of effect.

The inverse-variance weighted meta-analysis pooled effect size, its standard error, and corresponding z-score are:

$$\overline{\beta}_{IV} = \sum_{k=1}^{K} \hat{V}_k^{-1} \hat{\beta}_k \Big/ \sum_{k=1}^{K} \hat{V}_k^{-1}\ ;\quad SE(\overline{\beta}_{IV}) = \left[\sqrt{\sum_{k=1}^{K} \hat{V}_k^{-1}} \right]^{-1}\ ;\quad Z_{IV} = \overline{\beta}_{IV} / SE(\overline{\beta}_{IV}) \qquad \text{(Equation 3.4)}$$

where $\hat{\beta}_k$ and $\hat{V}_k$ are the study-specific estimates of the genetic effect regression coefficient and its variance.

Simulation-based estimation of type I error and power

We simulated QT datasets under the null and alternative hypotheses to estimate joint and meta-analysis type I error and power for normally and non-normally distributed QTs, with and without additional covariates. For joint analysis, each dataset had N individuals in total; for meta-analysis, we partitioned the same dataset into K equal-sized studies with n individuals each (e.g. $N = K \times n$). We simulated a single, additive, biallelic, causal variant with population-level MAF p and proportion of trait variance explained σ_G^2. In some simulations, we also included one binary covariate Z_1 distributed as Bernoulli with probability $f = 0.3$ and one continuous covariate Z_2 distributed as normal with mean zero and variance one; Z_1 and Z_2 have trait-variances explained $\sigma_{Z_1}^2$ and $\sigma_{Z_2}^2$, respectively.

We simulated the normally distributed QT y_i based on the linear regression model (equation e.1), assuming variant effect size $\beta = \sqrt{\sigma_G^2 / [2p(1-p)]}$, and covariate effect sizes (if included) $\gamma_1 = \sqrt{\sigma_{z_1}^2 / [f(1-f)]}$ and $\gamma_2 = \sqrt{\sigma_{z_2}^2}$, with normally distributed residual error with mean zero and variance = $1 - \sigma_G^2 - \sigma_{Z_1}^2 - \sigma_{Z_2}^2$. To mimic transformations used in actual data analysis, within each of the K studies (of size n) we first regressed out the QT-covariate effects (if included), and transformed the regression residuals $r_i = y_i - (\tilde{\alpha} + \tilde{\gamma}^T Z_i)$ using the rank-based, inverse-normal transformation: $t_i = \Phi^{-1}(rank(r_i)/(n+1))$, where Φ^{-1} is the inverse cumulative distribution function (CDF) of the standard normal distribution. To simulate non-normally distributed QTs, we transformed the regression residuals r_i to the target non-normally distributed QT as $s_i = F^{-1}\left(\Phi(r_i)\right)$, where F^{-1} is the inverse CDF of the target non-normal distribution. We explored the impact of distributions that have heavy tails or were skewed, and simulated *t* distributions on 1, 2, and 5 degrees of freedom, and chi-squared distributions on 1, 5, and 10 degrees of freedom.

For joint analysis, we performed association analysis using the transformed residuals combined across all *K* studies; for meta-analysis, we combined study-level association results using sample-size and inverse-variance weights (equations 3, 4). For all scenarios (e.g. normally and non-normally distributed QTs, with and without covariates), simulated datasets had N = 10,000 samples partitioned into K = 10 sub-studies. In addition, for the scenario with normally distributed QTs and without covariates, we simulated datasets with N = 1,000 and 100,000 samples partitioned into K = 10 sub-studies, and N = 10,000 partitioned into K = 50 sub-studies. We performed 10 million replicates under the null model to evaluate type I error at significance threshold $\alpha = 5\times10^{-5}$, and 10,000 replicates under the alternative model to evaluate power at $\alpha = 5\times10^{-8}$. Due to computational constraints, we are unable to evaluate type I error rates at the typical genome-wide significance threshold ($\alpha = 5\times10^{-8}$)

Genetics of type 2 diabetes (GoT2D) study

To assess these methods in the context of real data, we analyzed preliminary low-pass sequencing data from the Genetics of Type 2 Diabetes (GoT2D) study, which aims to investigate the role of low-frequency variation on type 2 diabetes (T2D) risk and variability in related QTs. We analyzed a subset of individuals (N = 789) with high density lipoprotein cholesterol (HDL) measurements from three studies: (1) 165 Swedish and Botnian Finnish individuals from the Diabetes Genetics Initiative (DGI); (2) 567 Finnish individuals from the Finland-United States Investigation of NIDDM Genetics (FUSION) study; and (3) 57 British individuals from the Wellcome Trust Case-Control Consortium (WTCCC). The DGI study was itself comprised of four sub-studies (Table 3.1). To account for differences in HDL levels between sub-studies, and between T2D cases and healthy controls (Table 3.1), and to normalize the skewed HDL distribution, for each sub-study, in cases and controls separately, we regressed HDL on age, age^2, and sex, and inverse-normalized the residuals. We performed joint analysis on the transformed residuals for all 789 individuals, and sample-size and inverse-variance weighted meta-analysis on the association results from the three studies using EPACTS [Kang, 2012] for association testing and METAL [Willer et al., 2010] for meta-analysis. We included no additional covariates except study indicators for joint analysis.

Table 3.1: Sample-sizes and untransformed HDL values for GoT2D studies and sub-studies

		Cases		**Controls**	
Study	**Sub-study**	**N**	**Mean HDL (SD)**	**N**	**Mean HDL (SD)**
DGI	Botnia	68	1.19 (0.29)	10	1.25 (0.34)
	Diabetes Registry	0	NA	35	3.22 (0.87)
	Helsinki Sib-pairs	0	NA	35	1.32 (0.39)
	Malmo Sib-pairs	0	NA	17	1.24 (0.43)
FUSION	FUSION	267	1.55 (0.43)	300	1.21 (0.34)
WTCCC	WTCCC	57	1.47 (0.39)	0	NA

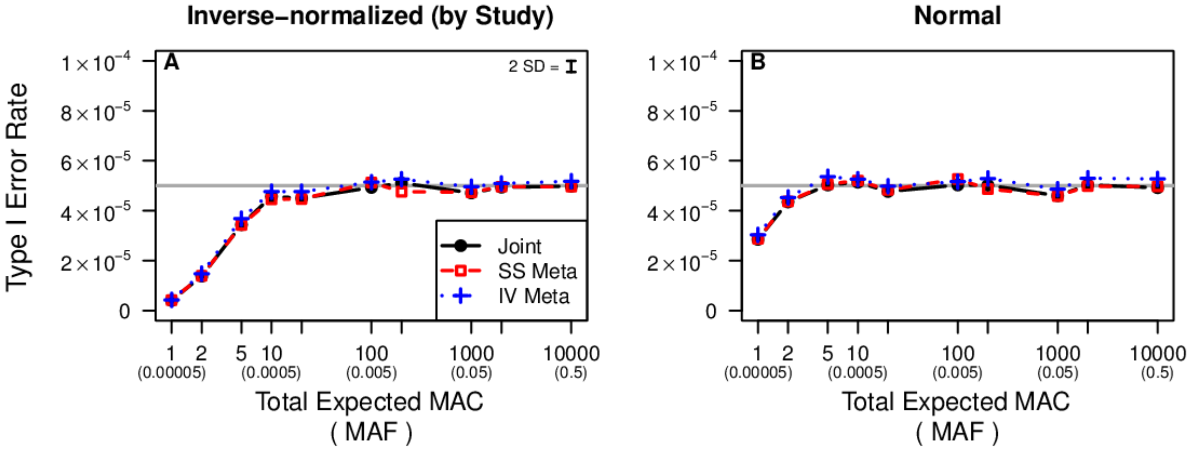
Results

Type I error rates for joint and meta-analysis

First, to examine the effect of rank-based inverse-normal transformation, we evaluate type I error rates (at α = 5×10^{-5}) for joint and meta-analysis of QTs with inverse-normalization within study and normally-distributed QTs without transformation (Figure 3.1). For variants with E[MAC] ≥ 10, joint and meta-analysis of QTs with inverse-normalization (Figure 3.1A) and without (Figure 3.1B) are both equally well-calibrated; for E[MAC] < 10, joint and meta-analysis is slightly more conservative for QTs with inverse-normalization than without. For example, for a low-frequency variant with E[MAC] = 5 (MAF = 2.5×10^{-4}), joint analysis is slightly more conservative with inverse-normal transformation (type I error rate = 3.4×10^{-5}; Figure 3.1A), than without (5.0×10^{-5}; Figure 3.1B). The increased conservativeness is due to shrinkage of QT outliers towards the mean by the rank-based inverse-normal transformation. Despite the increased conservativeness, investigators typically normalize QTs within study in real meta-analyses to account for differences in QT distributions between studies, and to correct for non-normally distributed QTs. Hence, subsequent simulation results will be inverse-normalized within study to mimic real data analysis.

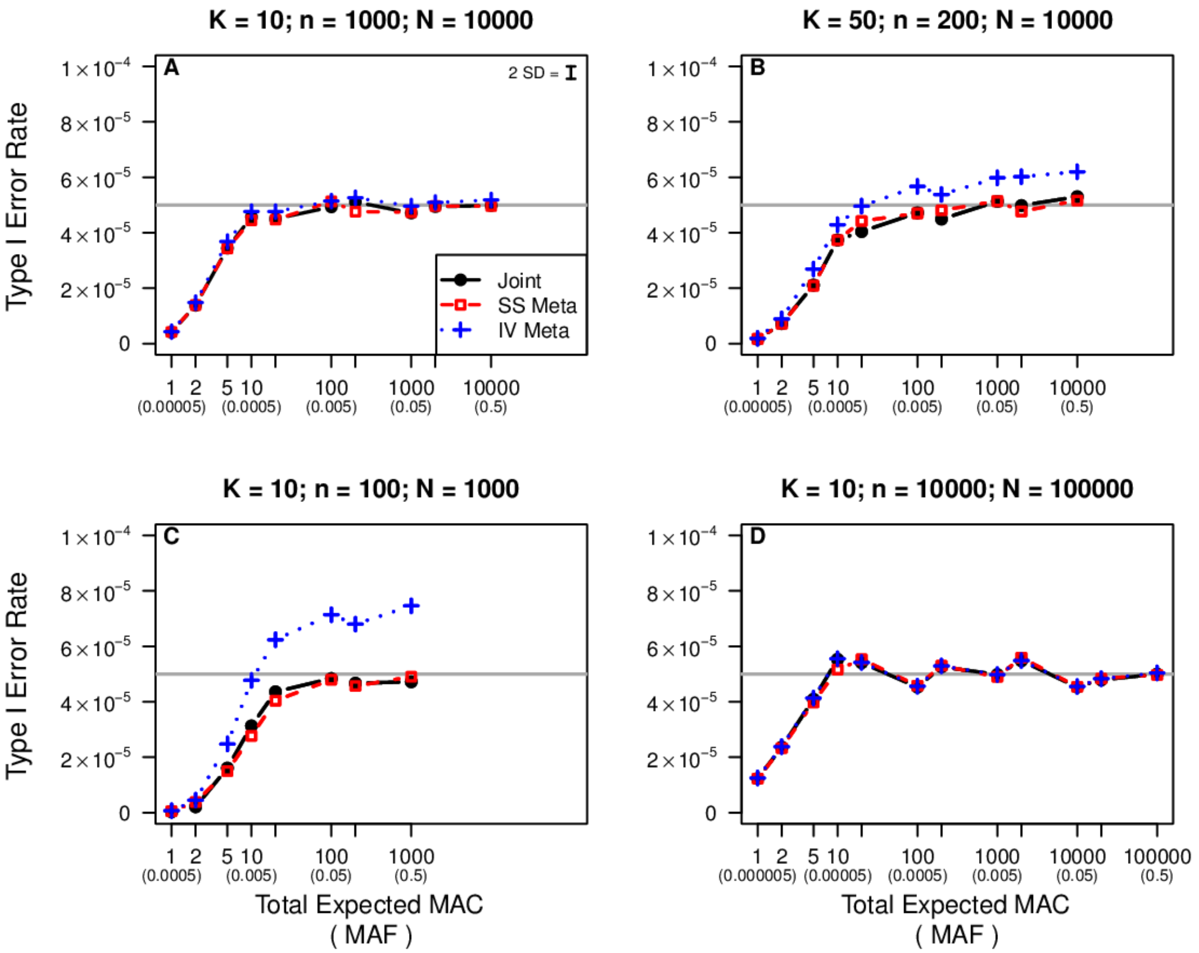
Figure 3.1: Type I error rates of inverse-normalized and normally distributed quantitative traits (QTs) for linear regression in joint and meta-analysis.

Sample-size (SS) and inverse variance (IV) weighted meta-analysis for analyses of $K = 10$ studies with $n = 1{,}000$ individuals each. QTs are (A) inverse-normalized within each study, and (B) normally distributed (without transformation). Joint analysis corresponds to analysis of the transformed or untransformed QTs combined from $N = n \times K$ individuals. All points are based on 10^7 simulation replicates, and the error bar (in panel A) represents two standard deviations (SD) from the nominal significance threshold $\alpha = 5\times10^{-5}$.



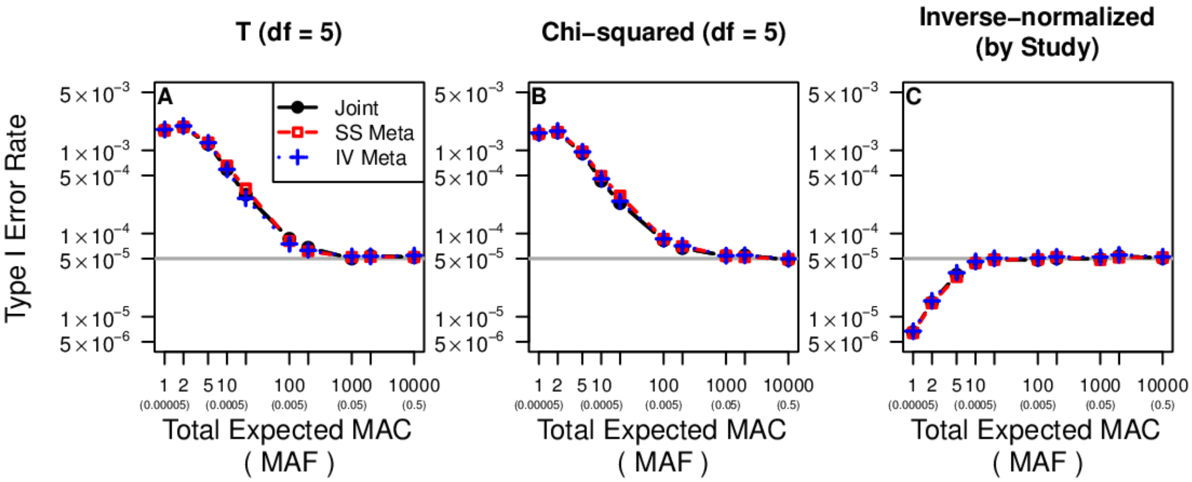
Second, we examine the effect of study-level sample size (*n)* and number of studies (*K*) on calibration (Figure 3.2). For E[MAC] ≥ 10, joint analysis and sample-size weighted meta-analysis are equally well-calibrated for all combinations of *n* and *K* considered (Figure 3.2A-D), but inverse-variance weighted meta-analysis can be slightly anti-conservative for smaller-sized studies with $n \leq 200$ (Figure 3.2B-C). For K=50 studies with n=200 individuals each, inverse-variance weighted meta-analysis has slightly inflated type I error rate = 6.0×10^{-5} (Figure 3.2B). This inflation occurs in meta-analysis of small-sized studies because we evaluate the inverse-variance weighted meta-analysis z-score Z_{IV} (equation 4) against the standard normal distribution, when Z_{IV} is actually a weighted sum of K variables each with t_{n-c-2} distributions; these effects are diminished in larger study sample-sizes since the t_{n-c-2} distribution converges to the standard normal. For E[MAC] < 10, joint and meta-analysis become increasingly conservative. These results hold true when we include additional covariates (Supplementary Figure S3.1).

Figure 3.2: Type I error rates of inverse-normalized quantitative traits (QTs) for linear regression in joint and meta-analysis.
Sample-size (SS) and inverse variance (IV) weighted meta-analysis for analyses of (A) K = 10 studies with n = 1,000 individuals each; (B) K = 50, n = 200; (C) K = 10, n = 100; and (D) K = 10, n = 10,000. QTs are inverse-normalized within each study. Joint analysis corresponds to analysis of the inverse-normalized QTs combined from $N = n \times K$ individuals. All points are based on 10^7 simulation replicates, and the error bar (in panel A) represents two standard deviations (SD) from the nominal significance threshold $\alpha = 5\times10^{-5}$.



Third, we examine the effect of non-normally distributed QTs on calibration (Figure 3.3). For QTs having heavy tailed (Figure 3.3A) or skewed (Figure 3.3B) distributions, joint and meta-analysis have near-identical, and substantially anti-conservative type I error rates at low frequencies. For a variant with E[MAC] = 5 (MAF = 2.5×10^{-4}) and a t_5-distributed QT, joint and meta-analysis are equally anti-conservative (type I error rates = 1.2×10^{-3}; Figure 3.3A). Applying the inverse-normal transformation within each study restores accurate calibration for joint and meta-analysis for variants with E[MAC] ≥ 10 (Figure 3.2C), but is somewhat conservative for low-frequency variants with E[MAC] < 10. Additional type I error results for other non-normal distributions are found in Supplemental Figure S3.2; results remain consistent when we include additional covariates (data not shown).

Figure 3.3: Type I error rates of non-normally distributed quantitative traits (QTs) for linear regression in joint and meta-analysis.
Sample-size (SS) and inverse variance (IV) weighted meta-analysis for analyses of K = 10 studies with n = 1,000 individuals each; joint analysis corresponds to analysis of $N = n \times K$ individuals. Non-normally distributed QTs with (A) t distribution (df = 5), and (B) chi-squared distribution (df = 5) are transformed from a normally distributed QT using the inverse cumulative distribution function of the target non-normal distribution. For comparison, type I error rates are displayed for (C) inverse-normalized QTs within each study. The horizontal line denotes the nominal significance threshold α = 5×10^{-5}.



Power of joint and meta-analysis

Overall, power generally reflects the relative conservativeness of the aggregation method - conservative methods typically have lower power. First, we examine the effect of QT inverse-normalization on power (at α = 5×10^{-8}; Figure 3.4E-F). For a common variant with E[MAC] = 1,000 (Figure 3.4E), joint and meta-analysis have near-equivalent power with and without inverse-normalization; for a low-frequency variant with E[MAC] = 5 (Figure 3.4F). joint and meta-analysis have reduced power for QTs with inverse-normalization than without, reflecting the increased conservativeness of inverse-normalized QTs for low-frequency variants (Figure 3.1B).

Second, we evaluate the effect of study sample-size (n) and number of studies (K) on power (Figure 3.3A-B). For E[MAC] = 1,000 (Figure 3.4A), joint and sample-size weighted meta-analysis have very similar power across different combinations of n and K; for small-sized studies ($n \leq 200$), inverse-variance weighted meta-analysis is slightly more powerful, but increased power corresponds to the anti-conservativeness of this method (Figure 3.2B-C). For E[MAC] = 5 (Figure 3.4B), joint and meta-analysis have reduced power versus a common variant (E[MAC] = 1,000; Figure 3.4A), especially in small-sized studies. In the presence of trait-associated covariates (Figure 3.4C-D), regressing the QT on the covariates does not alter the relative power between joint and meta-analysis, but by explaining more of the variation, can increase power for genetic association.

Third, we evaluate power for non-normally distributed QTs (Figure 3.4E-F). For E[MAC] = 1,000 (Figure 3.3E), QTs with heavy tailed (t_5) or skewed distributions (χ_5^2) have comparable power to normally distributed QTs in joint and meta-analysis. However, for E[MAC] = 5 (Figure 3.4F), non-normally distributed QTs have increased power, but the increased power is due to the severely inflated type I error rates (Figure 3.3A-B). Inverse-normal transformation of the QT can recover most of the power compared to the untransformed QT (Figure 3.4F).

Figure 3.4: Power of linear regression in joint and meta-analysis.
(A-B) Power for normally distributed QTs (inverse-normalized within study) without additional covariates in joint analysis, and sample-size (SS) and inverse-variance (IV) weighted meta-analysis. The genotypic variance explained (σ^2_G) is varied inversely to sample size so that power is approximately 50% for a common variant with expected minor allele count = 1,000. (C-D) Power for normally distributed QTs with binary (σ^2_{Z1}) and normally distributed (σ^2_{Z2}) covariate effects regressed out, and inverse-normalized within each study. (E-F) Power for non-normally distributed QTs having t (df=5) and chi-squared (df=5) distributions, and normally distributed QTs with and without inverse-normal transformation. Unless noted in the figure, meta-analysis corresponds to datasets having K = 10 studies with n = 1,000 individuals each; joint analysis corresponds to analysis of $N = n \times K$ individuals; and $\sigma^2_G = 0.003$. Power is evaluated at nominal significance threshold $\alpha = 5\times10^{-8}$, and each estimate is based on 10^4 simulation replicates. Error bars denote ± two standard deviations of the power estimate. Anti-conservative tests are marked with an "x" symbol.

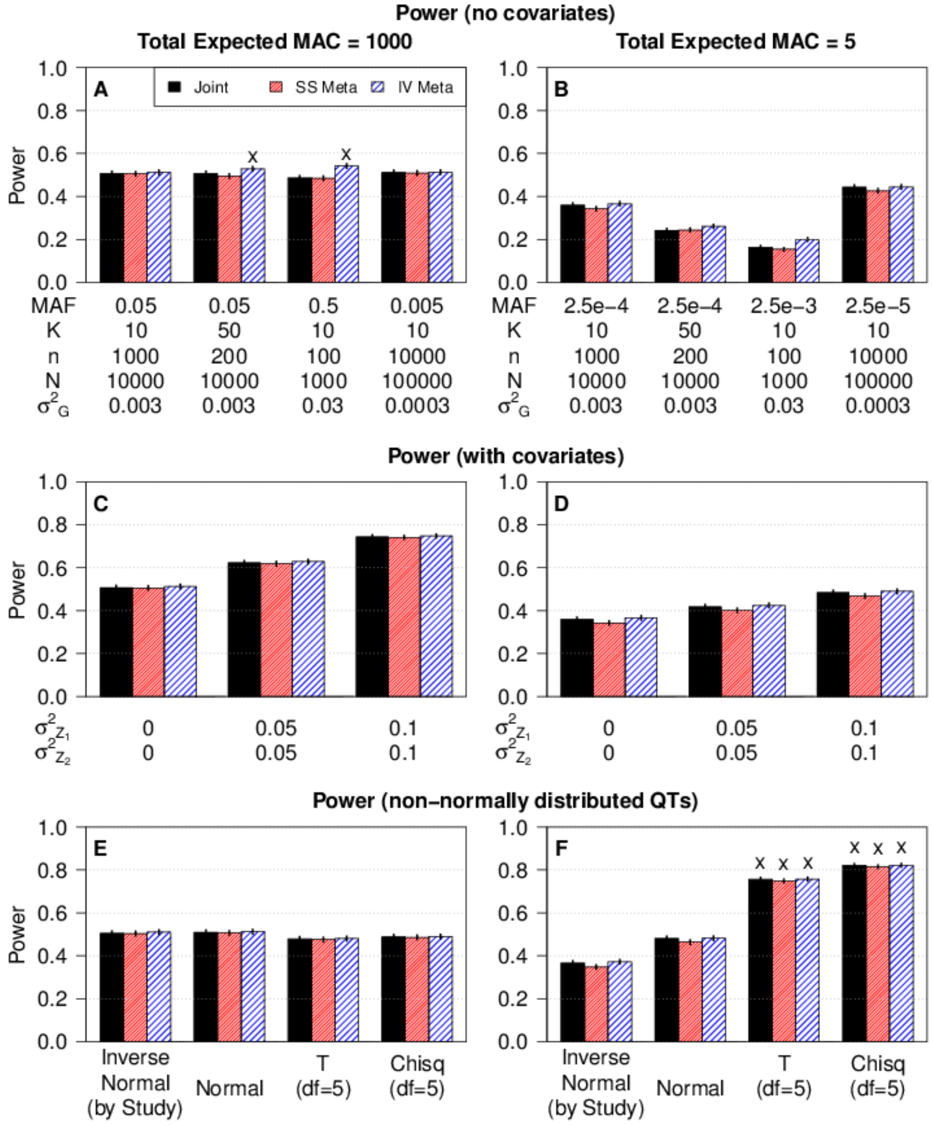
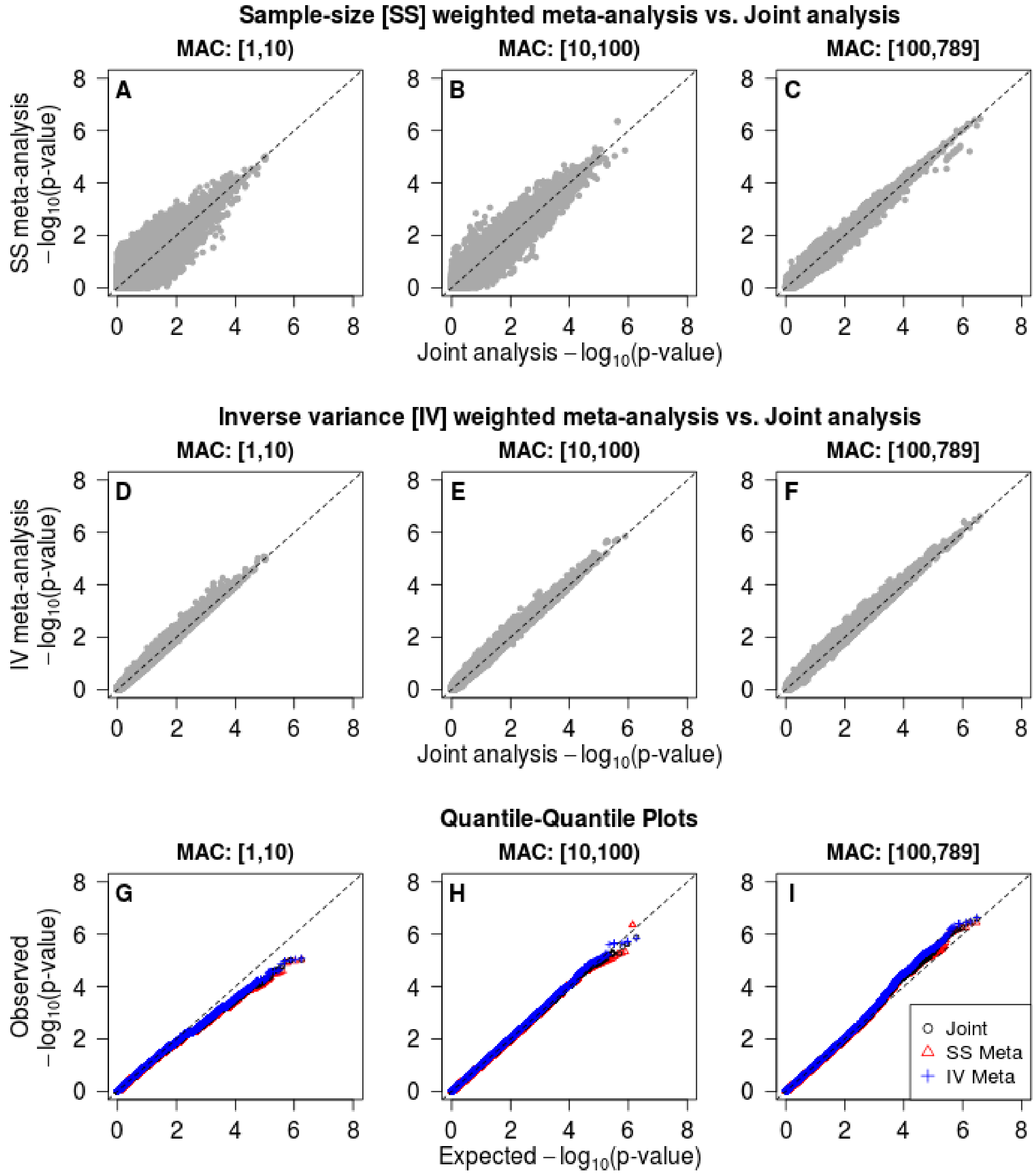


Figure 3.5: Joint and meta-analysis of high density lipoprotein (HDL) in the GoT2D study.
Effects of age, age^2, and sex on HDL are regressed out, and regression residuals are inverse-normalized within each in sub-study cases and controls separately. Joint analysis is based on transformed residuals from N = 789 individuals; meta-analysis is based on three studies with n = 165/567/67 individuals respectively. Scatterplots compare joint [JT] analysis p-values with (A-C) sample-size [SS] weighted, and (D-F) inverse variance [IV] weighted meta-analysis p-values, in different minor allele count [MAC] categories. (G-I) Quantile-Quantile plots compare expected and observed p-value distributions.



Joint and meta-analysis of GoT2D HDL data

To further examine the differences between joint and meta-analysis in real data, we tested for association between HDL cholesterol and 10 million variants (MAC ≥ 1) detected in low pass sequencing data from the GoT2D study. This GoT2D data subset was comprised of N = 789 individuals from three Northern European studies: DGI, FUSION, and WTCCC with n = 165/567/57 individuals (Table 3.1). DGI was itself comprised of four sub-studies with n = 17/35/35/78 individuals. To account for differences in mean HDL levels and remedy the skewed HDL distribution, in cases and controls separately, we first regressed the effects of age, age^2, and sex, then inverse-normalized the regression residuals within each sub-study. We performed joint analysis on all GoT2D individuals (N = 789), and performed sample-size and inverse-variance weighted meta-analysis using association results from the three studies.

For all MAC categories, joint analysis and sample-size weighted meta-analysis p-values are distributed symmetrically across the diagonal line (Figure 3.5A-C). In comparison, inverse-variance weighted meta-analysis p-values tend to be more significant than joint analysis (Figure 3.5D-F), and are less variable than comparing p-values between sample-size weighted meta-analysis and joint analysis (Figure 3.5A-C). In the quantile-quantile (QQ) plots (Figure 3.5G-I), p-values from inverse-variance weighted meta-analysis are slightly more significant than those for joint analysis (black) and sample-size weighted meta-analysis. These patterns are consistent with our simulation results, joint and sample-size weighted meta-analysis are equally well-calibrated and powered, but inverse-variance meta-analysis can be slightly anti-conservative, especially for meta-analysis of smaller studies.

Discussion

For association analysis of normally distributed QTs, joint analysis and sample-size weighted meta-analysis are well-calibrated and equally powered for variants with E[MAC] ≥ 10, but inverse-variance weighted meta-analysis can be slightly anti-conservative for small-sized studies; both joint and meta-analysis are increasingly conservative for low-frequency variants with E[MAC] < 10. For non-normally distributed QTs, joint and meta-analysis are very anti-conservative for low-frequency variants. Rank-based inverse-normal transformation can be used to recover well-calibrated type I error rates, but transforming QTs of any distribution reduces power for low-frequency variants. Inclusion of additional, independent covariates has no effect on calibration.

We focused our investigation on the commonly-used linear regression partial F test. This test is identical to the linear regression Wald and likelihood ratio tests when the trait variance (σ^2) is unknown, and asymptotically equivalent to the score test. Our simulation results demonstrate that the score test is slightly more conservative and less powerful than the Wald and likelihood ratio tests (data not shown).

For a low-frequency variant with the same MAC, we find that linear regression test-based joint and meta-analysis of QTs are generally better calibrated than logistic regression test-based joint and meta-analysis of binary traits (given the same sample sizes and significance thresholds). For normally-distributed QTs, we showed that joint and meta-analysis are slightly conservative for variants with E[MAC] < 10; for binary traits, joint and meta-analysis may be poorly calibrated for variants with E[MAC] ≤ 400 [Ma et al., 2013]. While a few extreme QT values can lead to a significant association using linear regression, a significant logistic regression association requires a minimum number of minor alleles to be present, for example, MAC ≥ 26 to achieve $\alpha \leq 5\times10^{-8}$ given equal numbers of cases and controls and no covariates.

For joint or meta-analysis of study data, it is critical to properly control for differences in QT values between studies, and remedy non-normally distributed QTs through appropriate transformation. We explored an alternative way to adjust for differences in HDL levels between the GoT2D studies. In the initial analysis, we only controlled for HDL differences between the three main studies (e.g. DGI, FUSION, and WTCCC), and did not account for differences within the four DGI sub-studies. This led to HDL levels in DGI with much greater variance than the other studies, and the resulting joint and meta-analysis p-values were more discordant than expected from our simulation results.

We presented power estimates based on nominal significance thresholds, e.g. power is the proportion of simulated p-values (under the alternative hypothesis) more significant than the nominal significance threshold $\alpha = 5 \times 10^{-8}$. By leveraging our type I error rate simulations, we also explored power based on empirical thresholds, at a less stringent $\alpha = 5 \times 10^{-5}$. For each test, the empirical threshold is the α^{th} quantile of the 10^7 p-values simulated under the null hypothesis; empirical power is the proportion of simulated p-values less than the empirical threshold. Overall, empirical and nominal power estimates were comparable, but in scenarios where a test was very conservative or anti-conservative, using an empirical threshold corrected for poor test calibration. For example, for a dataset having $K = 10$ studies with $n = 1{,}000$ individuals each, a low-frequency variant with E[MAC] = 10, and a skewed QT, nominal joint analysis power (= 0.63) is inflated, since joint analysis is anti-conservative for low-frequency variants with skewed QTs. In comparison, the empirical joint analysis power is 0.50, which is the same as the power of joint analysis for the inverse-normalized QT. The empirical threshold properly accounts for the test's poor calibration. However, in real data analysis, empirical thresholds are rarely used, since it is computationally infeasible to calculate empirical thresholds for all analysis scenarios at stringent thresholds ($\alpha = 5 \times 10^{-8}$).

In this study, we did not investigate the impact of differently sized studies within the same meta-analysis. Nor did we investigate the effects of population stratification. However, when aggregating data across several divergent populations, we can first adjust for within-study stratification using study specific covariates, and use meta-analysis instead of joint analysis, with minimal power loss.

We focused our investigation on rank-based inverse normal transformation, and did not evaluate other methods to transform non-normally distributed QTs, such as taking the natural logarithm of the QT. One disadvantage of inverse-normal transformation is that the regression parameter estimates are difficult to interpret in terms of differences in the mean of the QT [Buzkova, 2013]. Some studies have argued that transformations of non-normal QTs are not always necessary in genome-wide association studies. Buzkova [2013] demonstrates for single studies that heteroscedasticity (e.g. different population subgroups have different phenotypic variances from each other), MAF, and sample size can be more important in determining good test calibration, compared to the non-normality of the QT (and residual errors). Lumley et al. [2002] show that linear regression does not require the assumption of normally-distributed QT for studies with greater than 100-500 individuals (depending on the severity of the skewness). Beasley et al. [2009] explain that inverse-normal transformation does not necessarily guarantee good calibration, and reduces power in certain scenarios. However, in this study, we examined the impact of transformation in more extreme scenarios found in current genome-wide association studies: we examined very non-normal QTs and low-frequency variants at stringent significance thresholds. Without transformation, we find that joint and meta-analysis of non-normal QTs can be equally anti-conservative for low-frequency variants, but inverse-normal transformation (within study) can recover good calibration, with only minimal power loss.

In practice, the availability of individual-level data across studies dictates the use of joint or meta-analysis in genome-wide association studies. It is typically more difficult to share individual-level data between studies due to ethical requirements and studies with data sharing restrictions [Lin and Zeng, 2010]. We show that for analysis of (normally-distributed) QTs, sample-size and inverse-variance weighted meta-analysis are well-calibrated and efficient alternatives to joint analysis, for variants with E[MAC] ≥ 10. In most scenarios, meta-analysis using either sample-size or inverse-variance weights are near-equivalent, except for analysis of small-sized studies where inverse-variance meta-analysis can be slightly anti-conservative. In additional simulations, we observed that inverse variance meta-analysis of studies with sample sizes n ≥ 500 are well-calibrated (data not shown). Parameter estimates from inverse-variance weighted meta-analysis can be directly interpreted as differences in the mean of the QT, while sample-size weighted meta-analysis cannot. Overall, we recommend meta-analysis with inverse-variance or sample-size weights for most scenarios.

In summary, for study specific inverse normalized QTs, we demonstrate that linear regression-based joint analysis and sample-size weighted meta-analysis are well-calibrated and have equivalent power for variants with E[MAC] ≥ 10, while inverse-variance meta-analysis can be slightly anti-conservative for small-sized studies; joint and meta-analysis can be conservative for low-frequency variants with E[MAC] < 10. For non-normally distributed QTs, joint and meta-analysis can be equally anti-conservative especially for low frequencies, but inverse-normal transformation restores good calibration, but the transformation reduces power slightly when applied to any QT distribution. Hence, given normally distributed QTs, we recommend either joint and meta-analysis using linear regression for association analysis of variants of any frequency.

Supplemental Figures

Figure S3.1: Type I error rates of normally distributed quantitative traits (QTs) for linear regression in joint and meta-analysis with covariates.

Sample-size (SS) and inverse variance (IV) weighted meta-analysis correspond to datasets having K = 10 studies with n = 1,000 individuals each; joint analysis corresponds to analysis of $N = n \times K$ individuals. Datasets have one binary (Z_1) and one normally distributed (Z_2) covariate, with (A) no effect on the phenotype; (B) 5% and 0.5% variance explained; and (C) 10% and 10% variance explained. Covariate effects on the QT are regressed out, and regression residuals are inverse-normalized within each study. All points are based on 10^7 simulation replicates, and the error bar (in panel A) represents two standard deviations (SD) from the nominal significance threshold $\alpha = 5\times10^{-5}$.

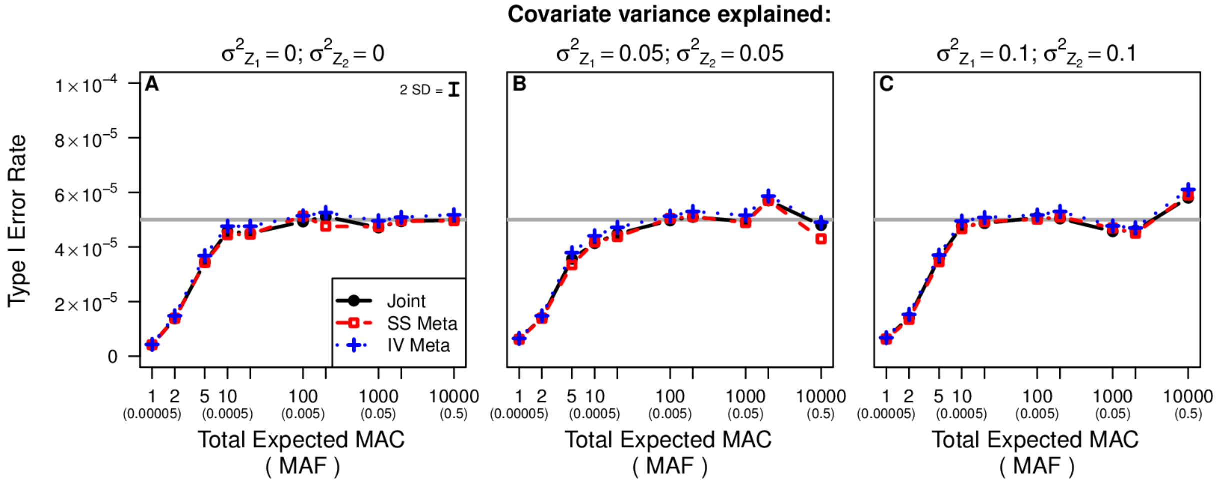
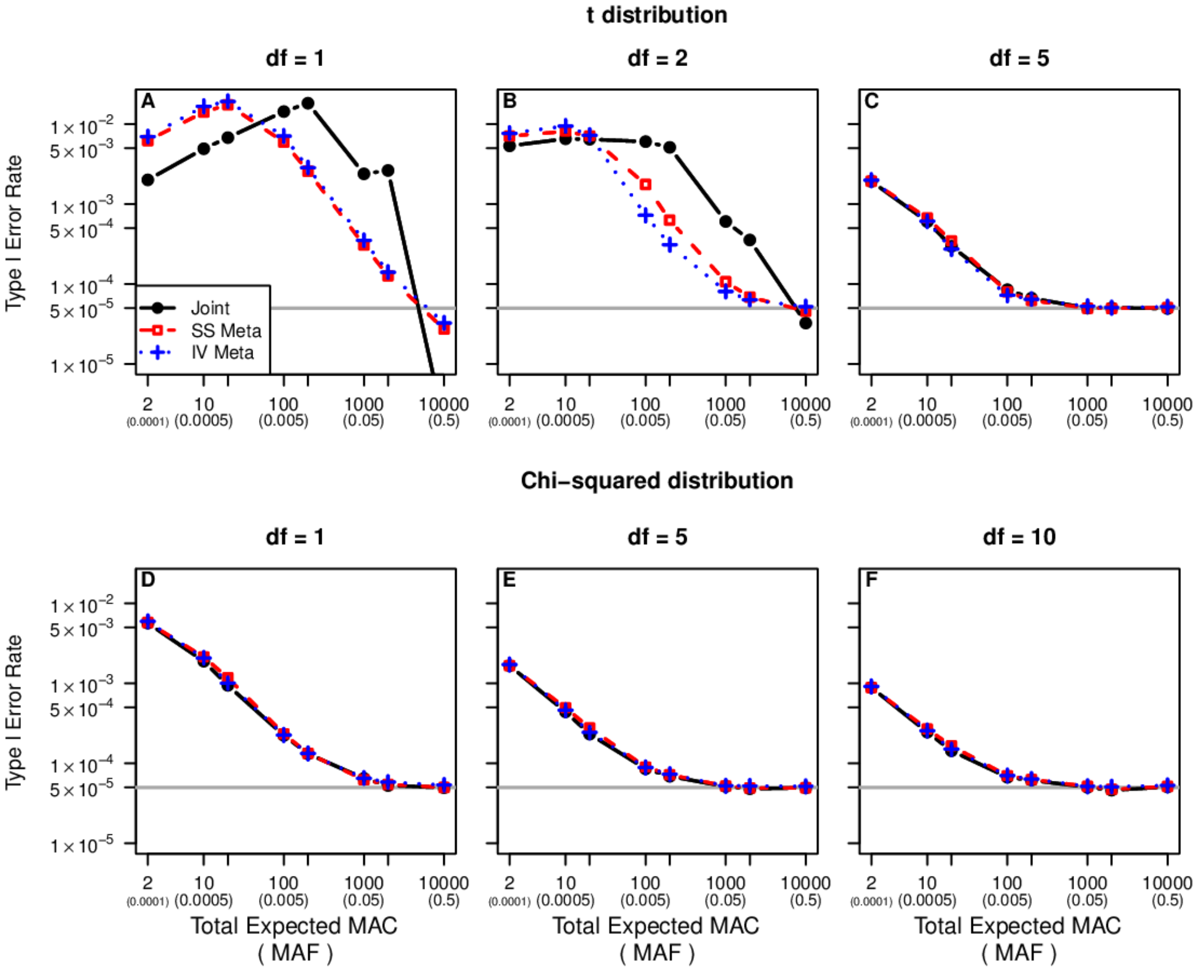


Figure S3.2: Type I error rates of additional non-normally distributed quantitative traits (QTs) for linear regression in joint and meta-analysis.

Sample-size (SS) and inverse variance (IV) weighted meta-analysis correspond to datasets having *K* = 10 studies with *n* = 1,000 individuals each; joint analysis corresponds to analysis of *N* = *n* × *K* individuals. Non-normally distributed QTs with (A-C) t distribution (df = 1, 2, or 5), and (B) chi-squared distribution (df = 1, 5, 10) are transformed from a normally distributed QT using the inverse cumulative distribution function of the target non-normal distribution. The horizontal line denotes the nominal significance threshold α = 5×10^{-5}.



Chapter 4: Evaluating the calibration and power of three gene-based association tests for the X chromosome

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 autosomal chromosomes [Wise et al., 2013]. Analysis of X chromosome association requires proper treatment of diploid female and haploid male participants. Whereas we can code female genotypes 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 [Marchini et al., 2007; Howie et al., 2012], and association analysis [Zheng et al., 2007; Clayton, 2008]. 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, many 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 allele-based 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 recommends 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. Notably, Loley showed that Clayton's non-sex-stratified tests can be anti-conservative when allele frequencies differ between the sexes. Hickey and Bahlo [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.

Recent GWAS use genome or exome sequencing [The GoT2D Consortium] or specialized genotyping arrays [Huyghe et al., 2013] to better assay low-frequency (0.5% < minor allele frequency [MAF] < 5%) and rare genetic variants (MAF < 0.5%). Single-marker tests have low power to test for association with rare variants unless the sample and/or effect size is very large [Asimit and Zeggini, 2010]. In contrast, gene-based tests in which multiple markers are analyzed jointly as a unit can be more powerful for analyzing rare variants [Lee et al., 2014]. The calibration and power of gene-based tests have not been evaluated in the context of analyzing rare variants on the X chromosome. Using simulated binary and quantitative trait (QT) datasets, we describe, apply, and evaluate three gene-based tests for the X chromosome: burden, SKAT, and optimal unified SKAT (SKAT-O) [Lee et al., 2012]. Specifically, 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. Power for association testing is comparable for studies with different ratios of males and females in cases and controls, but as expected, power depends on the underlying genetic architecture of the genomic region analyzed. In most scenarios, power is slightly increased when we analyze data using the true model to code male genotypes, but power is only decreased slightly under the incorrect model. 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 rare variants for both binary and quantitative traits.

Methods

Notation

Consider *n* individuals sequenced at *m* variants in a genomic region of interest. For individual *i*, let $\mathbf{X}_\mathbf{i}' = (x_{i1},\ldots,x_{is})'$ be the vector of *s* covariates (including a covariate for sex) and $\mathbf{G}_\mathbf{i}' = (g_{i1},\ldots,g_{im})'$ be the vector of genotypes. For (diploid) female *i*, 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} = {0,1} when assuming X-inactivation in the females and (2) g_{ij} = {0,2} 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 bi-allelic 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:

$$\text{logit}(\Pr(y_i = 1)) = \text{logit}(\pi_i) = \gamma_0 + \mathbf{X}_\mathbf{i}'\boldsymbol{\gamma}_\mathbf{1} + \mathbf{G}_\mathbf{i}'\boldsymbol{\beta} \qquad \text{(Equation 4.1)}$$

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

$$y_i = \gamma_0 + \mathbf{X}_\mathbf{i}' \boldsymbol{\gamma}_\mathbf{1} + \mathbf{G}_\mathbf{i}' \boldsymbol{\beta} + \varepsilon_i \qquad \text{(Equation 4.2)}$$

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

Since 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 (4.1) and (4.2) become:

$$\text{logit}(\pi_i) = \gamma_0 + \mathbf{X}_\mathbf{i}' \boldsymbol{\gamma}_\mathbf{1} + \beta_c \left(\sum_{j=1}^{m} w_j g_{ij} \right) \qquad \text{(Equation 4.3)}$$

$$y_i = \gamma_0 + \mathbf{X}_\mathbf{i}' \boldsymbol{\gamma}_\mathbf{1} + \beta_c \left(\sum_{j=1}^{m} w_j g_{ij} \right) + \varepsilon_i \qquad \text{(Equation 4.4)}$$

We use weights w_j= *Beta*(p_j, $\alpha = 1$, $\beta = 25$), having beta density function with shape parameters $\alpha = 1$ and $\beta = 25$, to up-weight rarer variants, where p_j is the MAF of variant *j* [Wu et al., 2011]. To test the gene-based null hypothesis $H_o : \beta_c = 0$, the burden score statistic is:

$$Q_B = \left(\sum_{j=1}^{m} w_j S_j \right)^2 \qquad \text{(Equation 4.5)}$$

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 SNP *j* in the regression model, and $\hat{\mu}_i$ is the estimated mean of y_i under H_0. The burden score statistic is evaluated relative to a scaled χ_1^2 distribution.

SKAT assumes the β_j*'s* follow an arbitrary distribution with mean zero and variance $w_j^2\tau$. Testing the null hypothesis H₀: **β**=0 is equivalent to testing H₀: τ=0. The SKAT score statistic is:

$$Q_S = \sum_{j=1}^{m} w_j^2 S_j^2 \qquad \text{(Equation 4.6)}$$

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 optimal unified association test 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:

$$Q_\rho = \rho Q_B + (1-\rho) Q_S, \quad 0 \le \rho \le 1 \qquad \text{(Equation 4.7)}$$

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 ρ = {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 using numerical integration [Lee et al., 2012].

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

Numerical simulations

To generate simulated genomic regions, we use 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 select 3Kb region from a single random haplotype. For each simulated female individual, we select a 3Kb region from two random haplotypes and pair 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 simulate case-control datasets with N_{cases} = 1,000 and N_{ctrls} = 1,000 under the logistic regression model:

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \beta_1 g_{i1} + \cdots + \beta_s g_{is} \qquad \text{(Equation 4.8)}$$

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 $g_{i1}, \ldots, g_{is}$ (but under the null model, we set genetic effects to zero $\beta_1 = \ldots = \beta_s = 0$). 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 explore a broad range of male:female ratios in cases and controls by sampling the exact number of males and females (Table 4.1) from the simulated cases and controls.

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

$$y_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \beta_1 g_{i1} + \cdots + \beta_s g_{is} + \varepsilon_i \qquad \text{(Equation 4.9)}$$

with X_{1i}, X_{2i}, X_{3i}, and $g_{i1}, \ldots, g_{is}$ as for equation (4.8). 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 $\sigma^2_{X2} = (\gamma_2)^2 f(1-f)$; the effect of sex is zero ($\gamma_3 = 0$). The normally-distributed residual error has mean zero and variance = $1 - \sigma^2_{X1} - \sigma^2_{X2}$. We sample the desired number of males and females (Table 4.2) from the simulated individuals.

Table 4.1: Sample sizes for simulated case-control datasets

	# Cases	**# Controls**
Simulation	**(Males:Females)**	**(Males:Females)**
A	500:500	500:500
B	900:100	500:500
C	100:900	500:500
G	500:500	900:100
H	500:500	100:900
I	900:100	900:100
J	100:100	100:900

Table 4.2: Sample sizes for simulated quantitative trait datasets

	# Individuals
Simulation	**(Males:Females)**
D	1,000:1,000
E	200:1,800
F	1,800:200

We analyze each simulated dataset using the six combinations of three gene-based tests (Equation 4.5, 4.6, 4.7) and two coding schemes for male genotypes. To increase computational efficiency, we simulate 100,000 independent datasets per simulation scenario, and resample the phenotype 1,000 times per independent dataset, resulting in a total of 100 million simulation replicates. We evaluate 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 estimate type I error as the proportion of simulation replicates with a p-value < $\alpha = 2.5\text{x}10^{-6}$, corresponding to Bonferroni correction for association testing of the approximately 20,000 genes in the human genome.

Power simulations

Within the 3Kb region, we select 10% or 50% of variants with MAF < 0.03 as causal. Using the same simulation settings as for type I error simulations, we simulate case-control datasets using the logistic regression model (Equation 4.7); we simulate QT datasets using the linear (Equation 4.8) regression model assuming the normally-distributed residual error ε_i has mean zero and variance = $1-\sigma_{X1}^2-\sigma_{X2}^2-\sum_{j=1}^{s}\sigma_j^2$ where $\sigma_j^2=(\beta_j)^2 2p_j(1-p_j)$ is the proportion of trait variance explained by variant *j*.

We simulate datasets under the alternative hypothesis assuming with and without X-inactivation coding for male genotypes. We consider genetic effect sizes proportional to the variant's MAF $|\beta_j|=c|\log_{10}p_j|/2$, and adjust the tuning parameter *c* so that power estimates will not be too close to 1 or 0. For binary traits, when 10% of variants are causal, *c* = log(15), which gives an odds ratio of 15 when MAF = 0.01; when 50% variants were causal, *c* = log(3) or log(5). For QTs, when 10% of variants are causal, *c* = log(7), which gives a linear regression coefficient of approximately 1.95 when MAF = 0.01; when 50% of variants are causal, c = log(1.8). We assume that either all causal variants are deleterious, or that 50% are deleterious and 50% are protective. We simulate 1,000 independent replicates per simulation scenario, and evaluate power as the proportion of replicates with p-value < $\alpha = 2.5\text{x}10^{-6}$.

Table 4.3: Type I error rates for burden, SKAT, and SKAT-O tests in binary and quantitative trait studies.
Type I error estimates are based on 10^8 simulation replicates so that the nominal significance threshold of $\alpha = 2.5\text{x}10^{-6}$ corresponds to 250 rejections. Empirical type I error rates between $2.2\text{x}10^{-6}$ and $2.8\text{x}10^{-6}$ have 95% confidence intervals which include the nominal value.

Binary Traits

				Type I error rate (×10^{-6})		
Sim.	**# Cases (Males:Females)**	**# 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	3.7
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	4.6
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	4.6

Quantitative Traits

			Type I error rate (×10^{-6})		
Sim.	**# Individuals (Males:Females)**	**Coding for Male Genotypes**	**Burden**	**SKAT**	**SKAT-O**
D	1000:1000	No X-inactivation g={0,1}	2.5	2.1	2.8
		X-inactivation g={0,2}	2.5	2.4	2.8
E	1800:200	No X-inactivation g={0,1}	2.7	2.5	3.3
		X-inactivation g={0,2}	2.5	2.4	3.1
F	200:1800	No X-inactivation g={0,1}	2.6	2.8	3.4
		X-inactivation g={0,2}	2.7	2.8	3.7

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 4.3). For each gene-based test, we examine 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 burden (non-X-inactivation type I error rate = 2.4×10^{-6} vs. X-inactivation $=1.8 \times 10^{-6}$) and SKAT (2.6×10^{-6} vs. 1.9×10^{-6}). However, the opposite is true for SKAT-O (2.4×10^{-6} vs. 3.7×10^{-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 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 (Supplemental Figure S4.1).

We also examine 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 (Supplemental Figure S4.1), 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 demonstrate 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 (Supplemental Figure 4.2A-C) and without resampling (Supplemental Figure 4.2D-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 4.3). Type I error rates are nearly identical between the two male coding schemes.

Power

We examine 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 50% protective). For binary traits, power results (Figure 4.1, 4.2) 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 rare variants are causal and have the same direction of effect (Figure 4.1D-F). SKAT is more powerful when 10% of rare variants are causal with the same or opposite direction of effect (Figure 4.1A-C), or when 50% of causal variants have opposite direction of effect (Figure 4.2). 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 non-adjusted power estimates are comparable (data not shown).

Next, we investigate 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, except for

SKAT-O which is unexpectedly more powerful for the misspecified X-inactivation coding (g_{ij} ={0,2}) (Figure 4.1, 4.2). However, when simulating variants assuming X-inactivation, X-inactivation coding (g_{ij} ={0,2}) is slightly more powerful in all scenarios (Supplemental Figure 4.3 and 4.4). 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 (Figure 4.1E), and the average difference is 2.3%. Power results are comparable for studies with different male:female ratios in cases and controls. Finally, for QTs, power comparisons are very similar to those for binary traits (Figure 4.3, 4.4).

Figure 4.1: 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 3Kb 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. In cases, there are (A,D) 500/500, (B,E) 900/100, and (C,F) 100/900 males and females respectively. In controls, all simulated datasets have 500 males and 500 females. Power estimates (at $\alpha = 2.5\text{x}10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

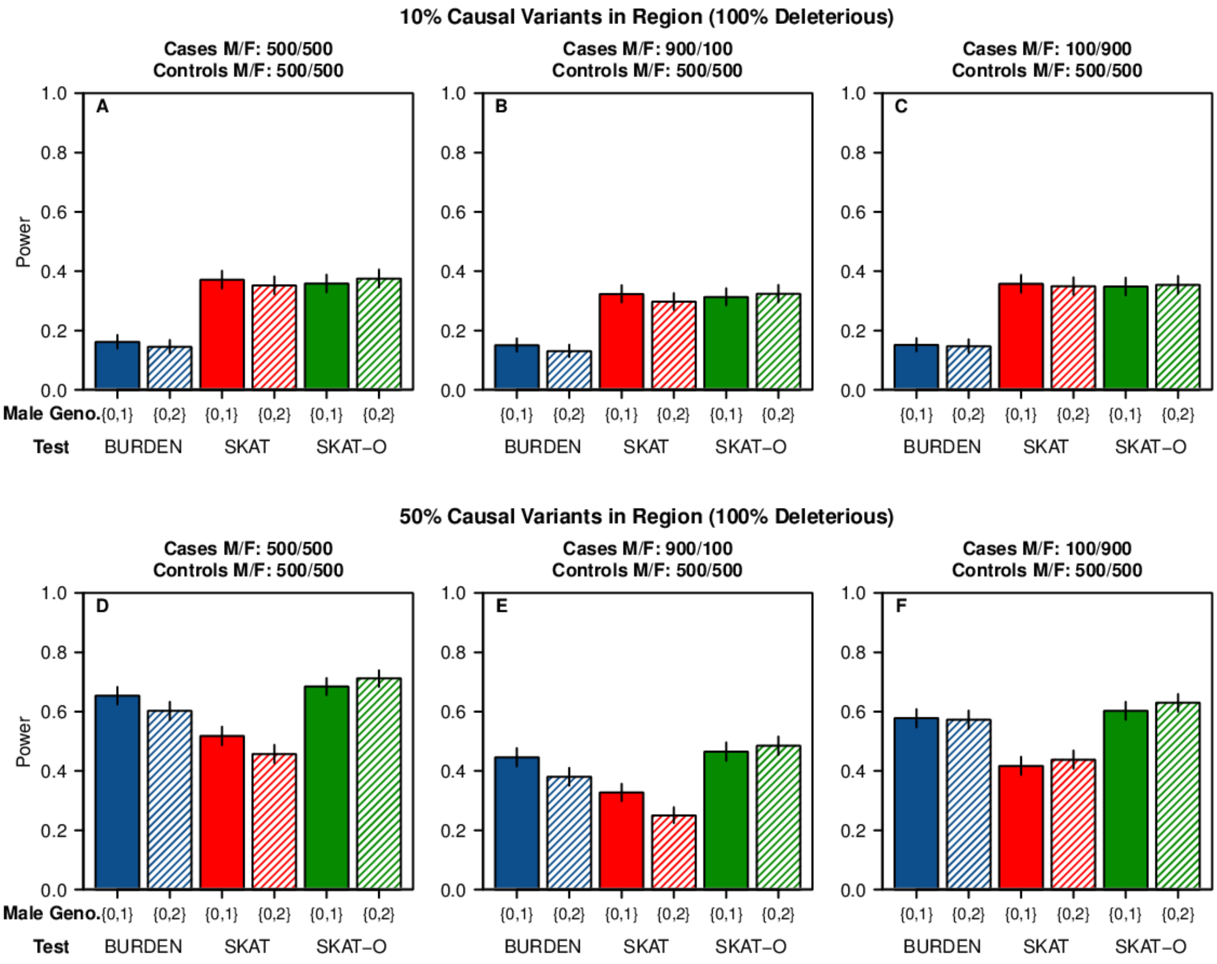


Figure 4.2: Power for gene-based tests in case-control studies assuming causal variants are 50% deleterious and 50% protective.

Causal variants are simulated with non-X-inactivation coding of male genotypes g_{ij} = {0,1}. Within each simulated 3Kb 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. In cases, there are (A,D) 500/500, (B,E) 900/100, and (C,F) 100/900 males and females respectively. In controls, all simulated datasets have 500 males and 500 females. Power estimates (at $\alpha = 2.5x10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

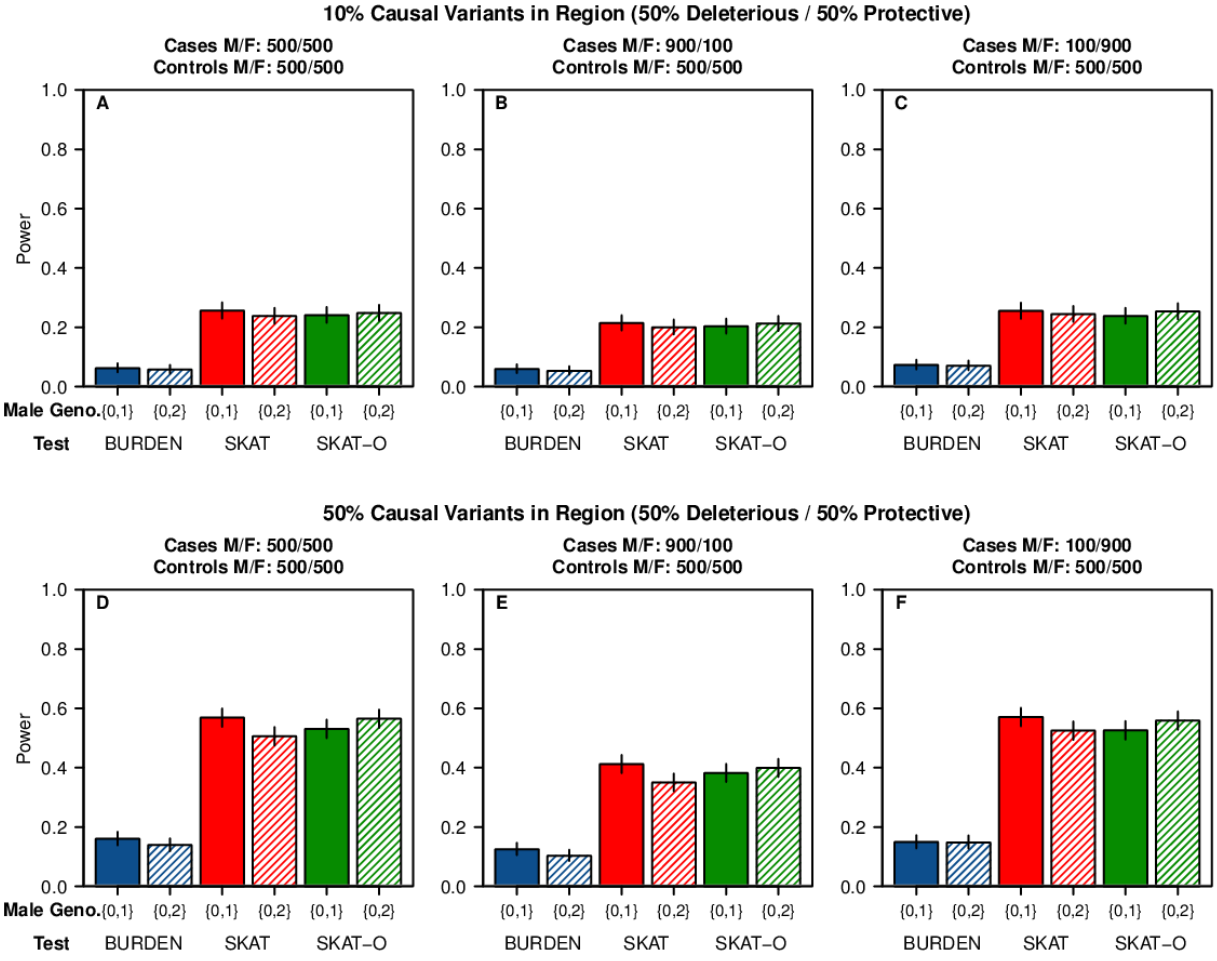


Figure 4.3: Power for gene-based tests in QT studies assuming all causal variants are deleterious.

Causal variants are simulated with non-X-inactivation coding of male genotypes $g_{ij} = \{0,1\}$. Within each simulated 3Kb 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 $\left|\beta_j\right| = c\left|\log_{10} p_j\right|/2$, and is proportional to MAF (pj) 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 $\alpha = 2.5\text{x}10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

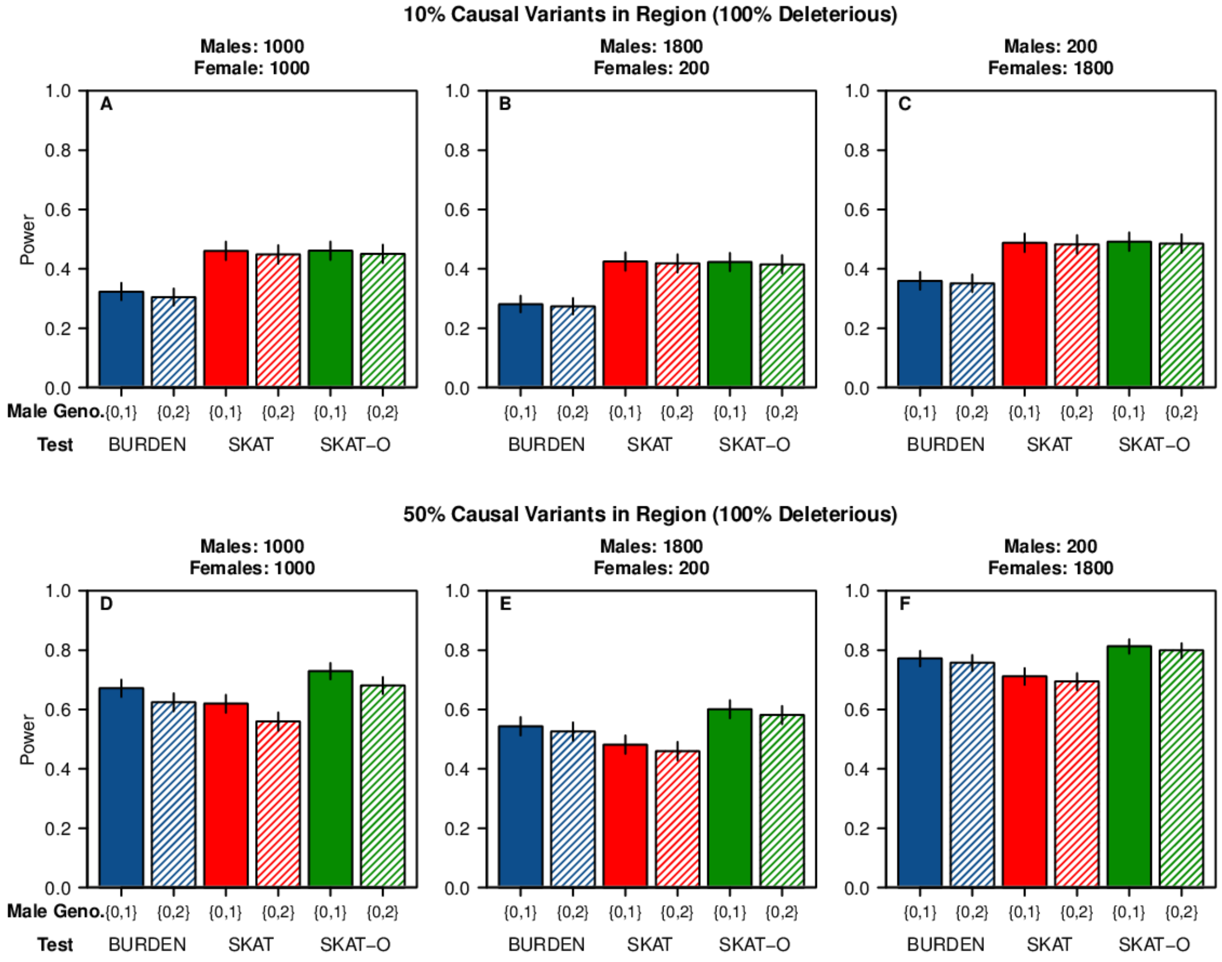
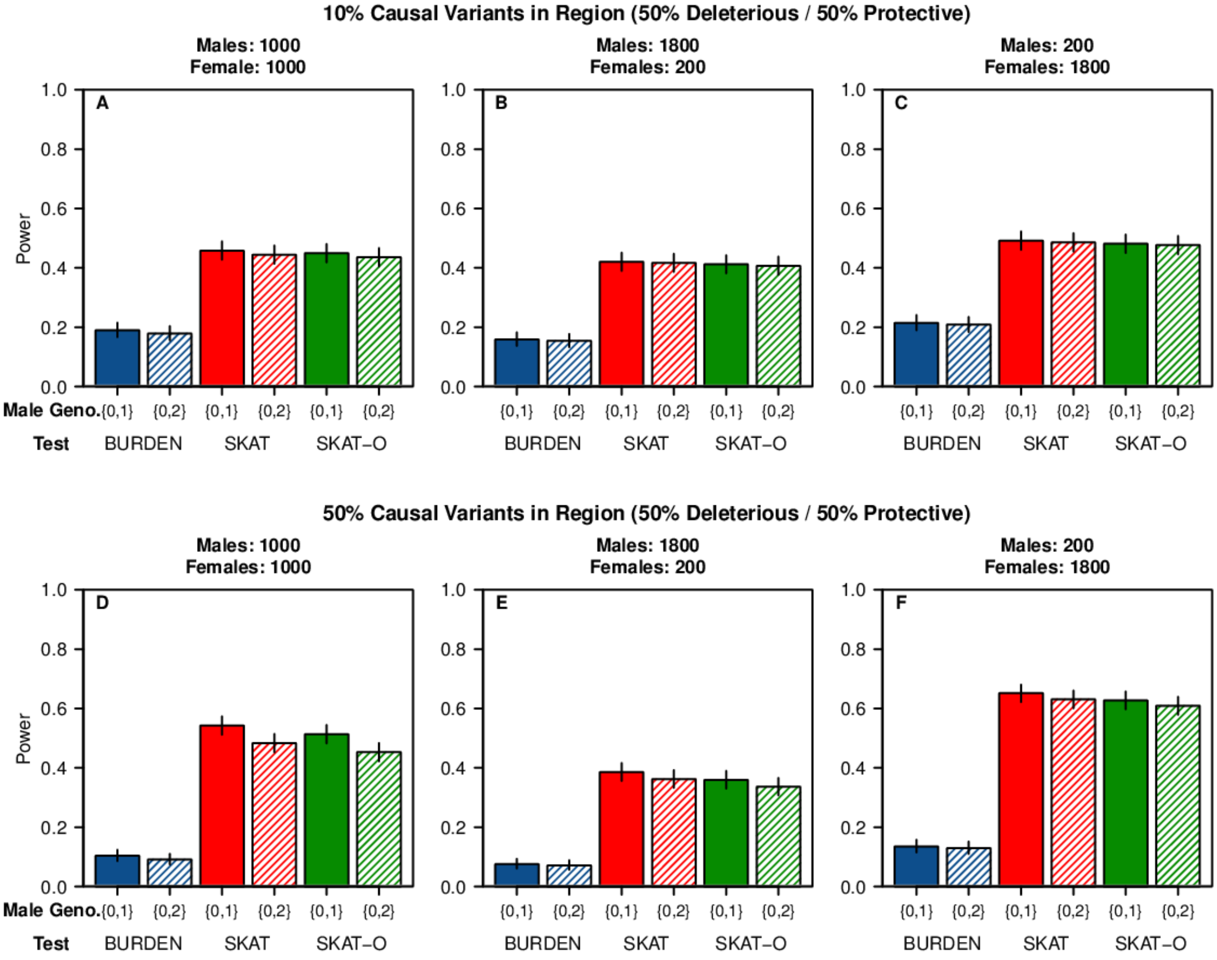


Figure 4.4: Power for gene-based tests in QT studies assuming causal variants are 50% deleterious and 50% protective.
Causal variants are simulated with non-X-inactivation coding of male genotypes $g_{ij} = \{0,1\}$. Within each simulated 3Kb genomic region, (A-C) 10% or (D-F) 50% of variants with MAF < 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 $\alpha = 2.5\text{x}10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.



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 quantitative trait data. For binary traits, all tests are well-calibrated or slightly anti-conservative for all simulation scenarios. Power differences reflect 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 coding scheme. Differences in male:female ratio in cases and controls have little effect on power. For QTs, the burden and SKAT tests are well-calibrated, and SKAT-O is very slightly anti-conservative; power results are similar to those for binary traits.

While 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 estimated power using nominal significance thresholds, e.g. power is the proportion of simulation replicates (under the alternative hypothesis) with p-values more significant than the nominal significance threshold $\alpha = 2.5\times10^{-6}$. By using our type I error simulations, we investigated the power using empirical thresholds. For each test, the empirical threshold is the α^{th} quantile of the 10^8 p-values simulated (with resampling) under the null hypothesis; empirical power is the proportion of simulated p-values more significant than the empirical threshold. Overall, power using either nominal or empirical thresholds are near-identical since tests are relatively well-calibrated across all scenarios, and the empirical thresholds (range = 1×10^{-6} — 3.5×10^{-6}) are very similar to the nominal significance threshold (data not shown). However, we observe that using empirical thresholds can partially resolve unexpected power results by accounting for poorly calibrated tests. When power is calculated using nominal thresholds, assuming a non-X-inactivated region, SKAT-O is unexpectedly more powerful using the misspecified X-inactivation coding (g_{ij} ={0,2}) compared to the non-X-inactivated coding (g_{ij} ={0,1}), (Figure 4.1, 4.2). However, when we use empirical thresholds, this effect is removed or reduced (data not shown). In practice, it is usually computationally infeasible to use empirical thresholds in real data analysis.

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 could be extended to other gene-based tests. 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 non-random or skewed X-inactivation [Amos-Landgraf et al., 2006], where >75% of cells have one allele inactivated. While a unified approach to account for both random and non-random X-inactivation may be more robust and powerful, as demonstrated by Wang et al. [2014a] 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 demonstrate that the burden, SKAT, and SKAT-O 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 rare variants on the X chromosome.

Supplemental Figures

Figure S4.1: Complete type I error rates for the burden (BURD), SKAT, and SKAT-O tests in case-control studies.

Type I error rates are for studies with (A-C) male/female ratio = 1; (D-E) male/female ratio > 1; (F-G) male/female ratio < 1. The suffixes "1" and "2" denote coding male genotypes as g_{ij} = {0,1} or {0,2} respectively. The suffix "ADJ" denotes analysis with small-sample adjustment. Type I error rates (at $\alpha = 2.5\text{x}10^{-6}$) are based on 10^8 simulation replicates, and vertical bars denote the exact binomial 95% confidence interval.

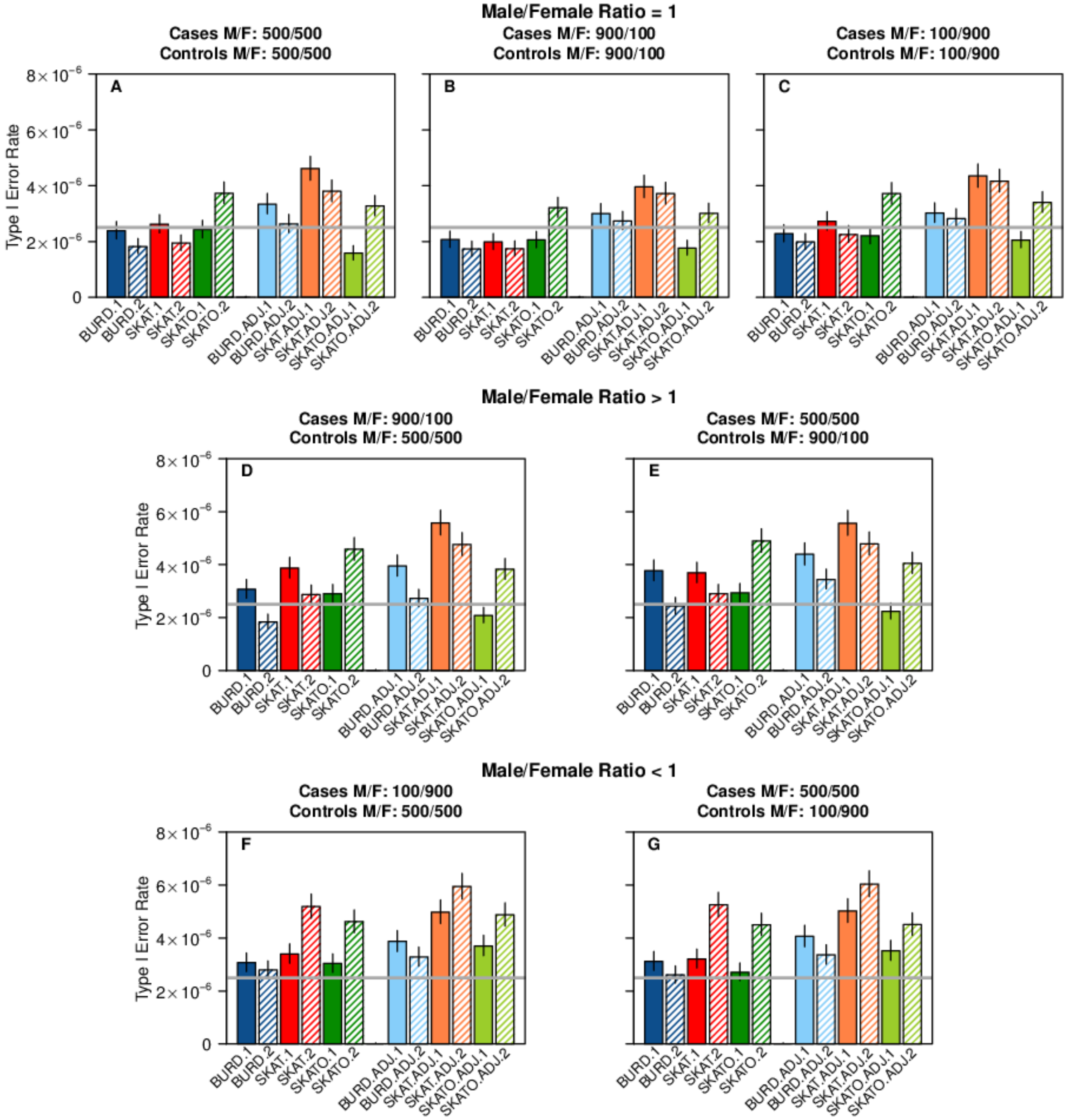


Figure S4.2: Type I error rates based on simulated datasets with re-sampling and without re-sampling .

Type I error rates are based on simulations with (A-C) 1,000 independent replicates each with 1,000 resampled phenotypes (total replicates = 10^6); and (D-E) 10^6 independent replicates. In cases, there are (A,D) 500/500, (B,E) 900/100, and (C,F) 100/900 males and females respectively. In controls, all simulated datasets have 500 males and 500 females. Type I error rates are evaluated at α = $5x10^{-4}$, and vertical bars denote the exact binomial 95% confidence interval.

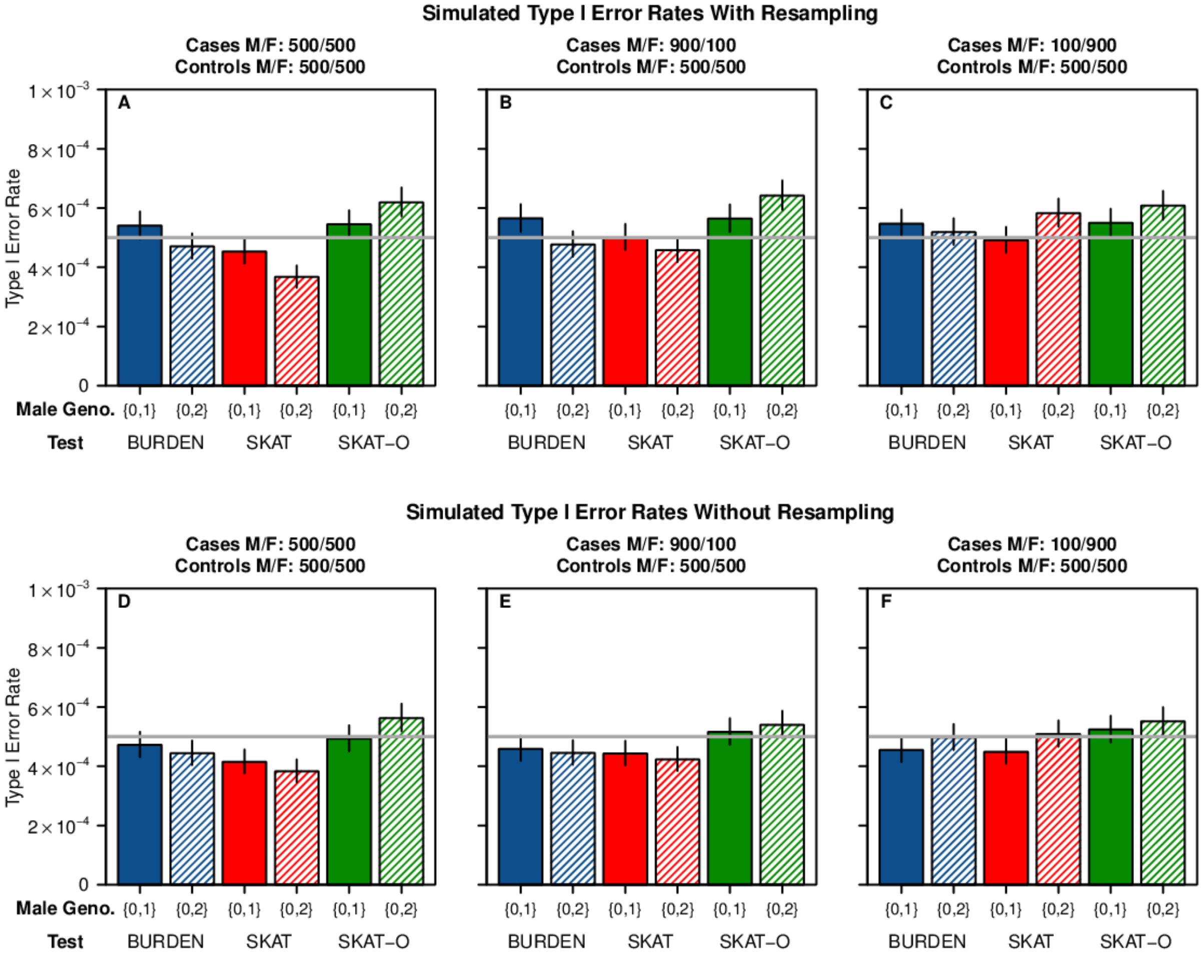


Figure S4.3: Power simulated with X-inactivation for gene-based tests in case-control studies assuming all causal variants are deleterious.
Causal variants are simulated with X-inactivation coding of male genotypes $g_{ij} = \{0,2\}$. Within each simulated 3Kb genomic region, (A-C) 10% or (D-F) 50% of 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. In cases, there are (A,D) 500/500, (B,E) 900/100, and (C,F) 100/900 males and females respectively. In controls, all simulated datasets have 500 males and 500 females. Power estimates (at $\alpha = 2.5\text{x}10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.

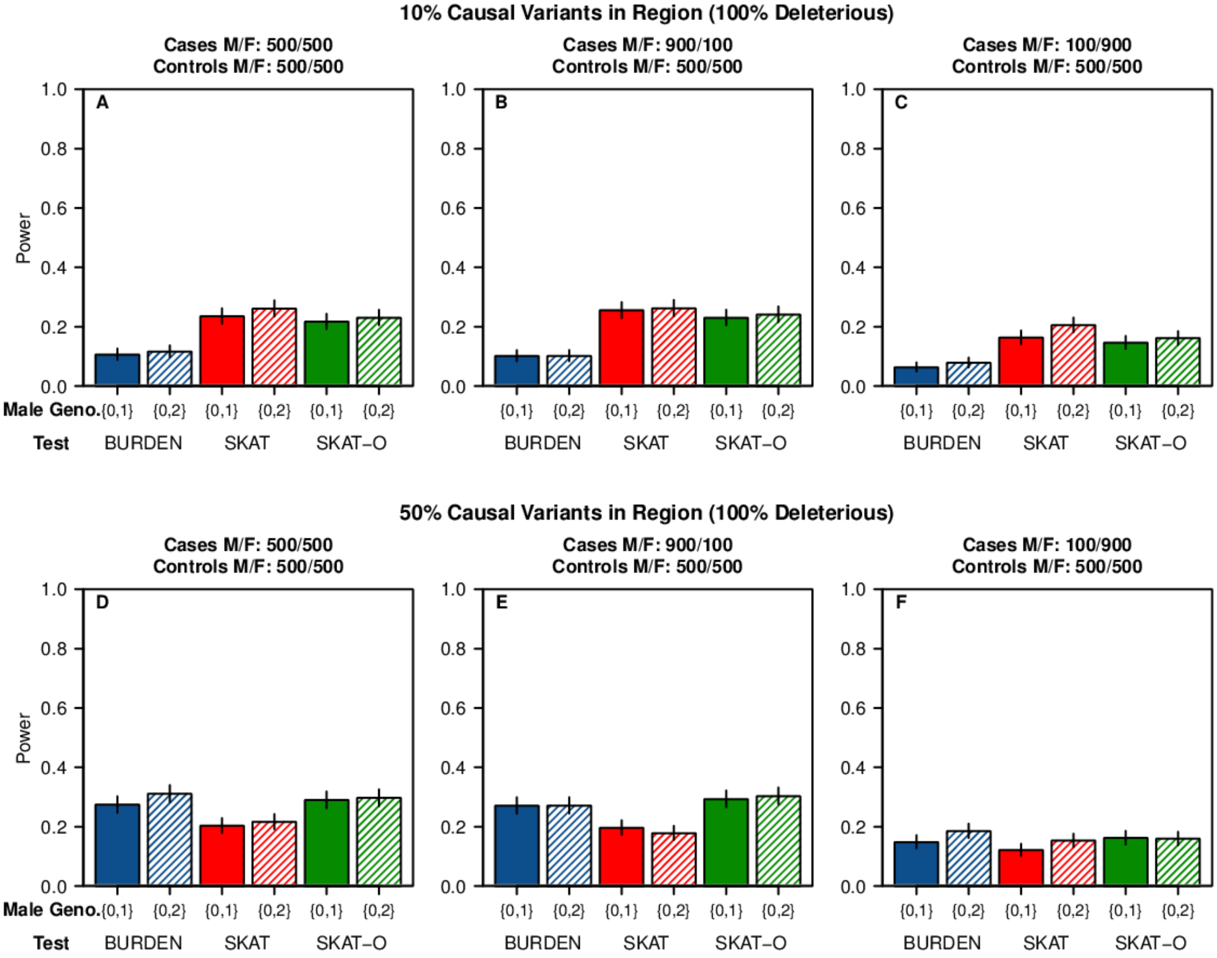
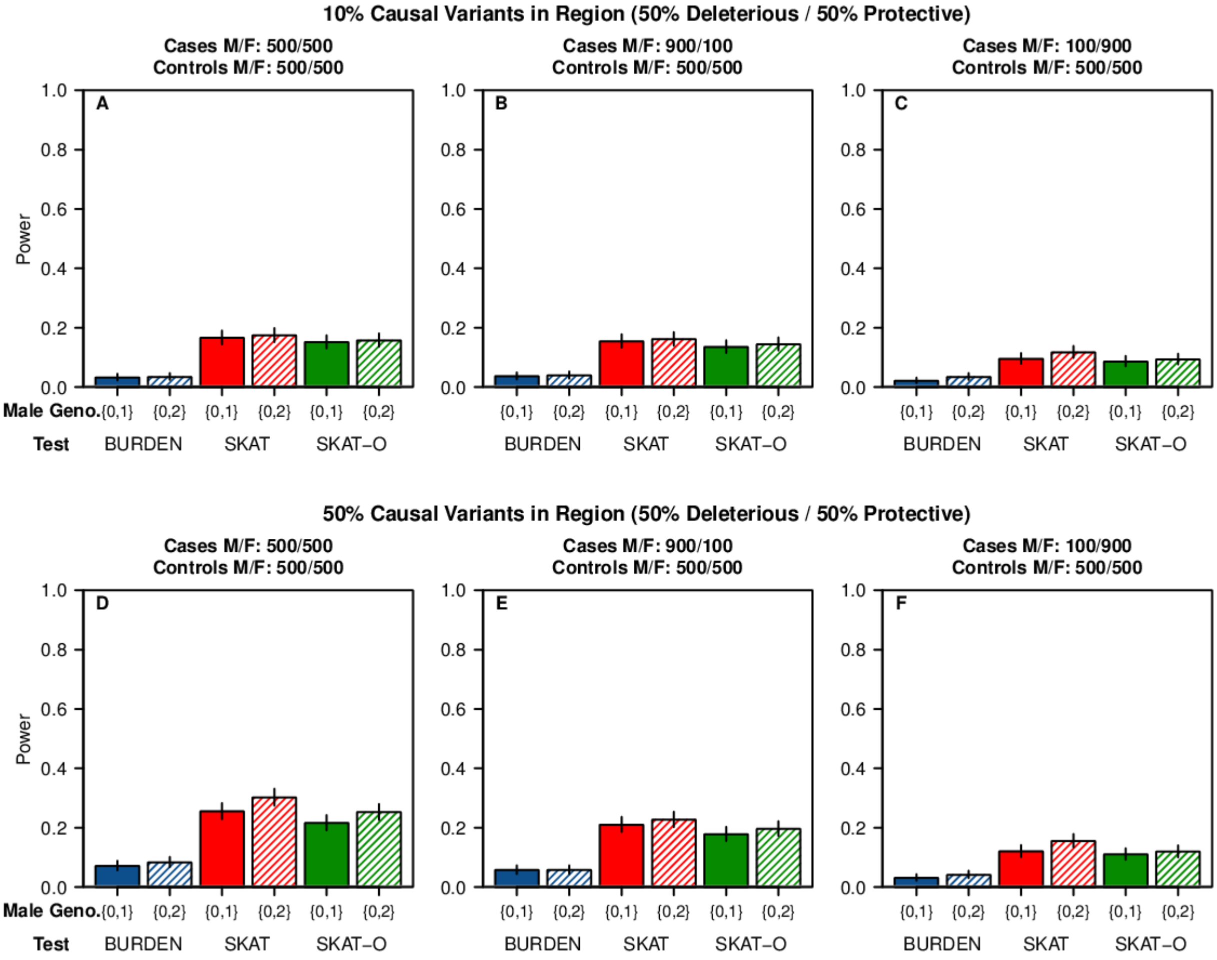


Figure S4.4: Power simulated with X-inactivation for gene-based tests in case-control studies assuming causal variants are 50% deleterious and 50% protective. Causal variants are simulated with X-inactivation coding of male genotypes $g_{ij} = \{0,2\}$. Within each simulated 3Kb genomic region, (A-C) 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); causal variants are simulated as 50% deleterious and 50% protective. In cases, there are (A,D) 500/500, (B,E) 900/100, and (C,F) 100/900 males and females respectively. In controls, all simulated datasets have 500 males and 500 females. Power estimates (at $\alpha = 2.5\text{x}10^{-6}$) are based on 1,000 simulation replicates; vertical bars denote the exact binomial 95% confidence intervals.



Chapter 5: Summary, discussion, and future directions

Results summary

In my dissertation, I investigate the calibration and power of commonly used statistical methods in genome-wide association studies (GWAS) applied to new situations. In Chapter 2, I examine the calibration and power of logistic regression tests in joint and meta-analysis of low-frequency genetic variants for case-control data. I discover that logistic regression-based joint and meta-analysis can be poorly calibrated and/or have low power when analyzing variants at a combination of three extremes: low MAF, stringent significance threshold α, and severe case-control imbalance. Based on simulation results, the recommended test for analyzing low-frequency variants differs depending on the availability of individual-level data across studies.

In Chapter 3, I examine the calibration and power of linear regression in joint and meta-analysis across a wide range of MACs for normally-distributed and non-normally distributed QTs, with and without adjusting for the effects of additional covariates. In contrast to binary traits, for normally-distributed QTs, linear regression-based joint and meta-analysis are relatively well-calibrated and have good power for variants with MAC ≥ 10 (given sufficiently large sample-sizes). For non-normally distributed QTs, joint and meta-analysis can be equally anti-conservative especially for low-frequency variants. Inverse-normal transformation of the QT can restore accurate calibration but inverse-normal transformation of QTs of any distribution reduces power.

In Chapter 4, I examine the calibration and power of the burden, SKAT, and SKAT-O aggregation association tests for analyzing rare variants on the X chromosome, in both case-control and QT datasets. For case-control datasets, all tests are relatively well-calibrated across all simulation scenarios. Power depends on the underlying genetic architecture of the genomic region analyzed. Power increases slightly when the coding scheme for male genotypes matches the underlying genetic model (with or without X-inactivation), but power loss is small if the coding scheme is misspecified. Power is comparable across different male:female ratios in cases and controls. For QT datasets, patterns of calibration and power are very similar to those for case-control data.

Strategies for numerical simulations

In each of the three projects, I performed extensive numerical simulations to evaluate type I error rates at stringent significance thresholds. To put into perspective the sheer number of replicates required, we need 10^{10} simulation replicates to expect 500 rejections under the null hypothesis at $\alpha = 5\times10^{-8}$; the resulting type I error estimate would have 95% exact binomial confidence intervals of $[4.57\times10^{-8}, 5.46\times10^{-8}]$. Then, we need to repeat the 10^{10} simulations for every combination of simulation parameters (e.g. MAF, case-control ratio, etc.). To overcome this immense computational burden, I employed a number of simulation strategies and algorithms to evaluate type I error rate at, or close to, the desired significance threshold.

In Chapter 2, I devised an exact calculation to analytically compute type I error rates (at $\alpha = 5\times10^{-8}$) for logistic regression based-joint analysis in case-control studies, similar to a method described by Upton [1982]. I assume a dominant genetic model, which is a good approximation to the multiplicative model for low-frequency variants, to reduce a case-control study to a 2 by 2 contingency table (assuming no additional covariates). The exact calculation enumerates all possible tables for each MAC, and the analytical type I error rate is simply the sum of the probabilities of tables rejecting the null hypothesis. This analytical calculation can compute type I error rates for any desired significance level in realistic timeframes. For example, in a single case-control study with 10,000 cases and 10,000 controls, and a variant with MAF = 0.0005, I can calculate type I error rates (at $\alpha = 5\times10^{-8}$) in 12 CPU-minutes using exact computation, compared to 9.5 CPU-years using direct simulation (assuming 10^{10} simulation replicates). However, the exact approach is limited to variants with MAF ≤ 0.05, where the dominant genetic model is a good approximation to the commonly used multiplicative model.

I devised an exact calculation of type I error rates for meta-analysis of $K = 2$ equal-sized studies for variants with MAC < 100; the calculation was computationally infeasible for $K >$ 2 or variants with MAC ≥ 100. The algorithm enumerates all possible ways to partition the MACs from the joint sample into the K studies, and calculates type I error rates by summing the probabilities of configurations that reject the null hypothesis. For example, for a variant with MAC = 50 in both cases and controls, and for $K = 2$, there are 51 possible ways to partition 50 alleles in cases, and 51 partitions for controls, thus resulting in $(51)^2$ configurations for cases and controls combined. For MAC < 100 and $K = 2$, the exact algorithm produces type I error rates very similar to those estimated by direct simulation. However for MAC ≥ 100 or K > 2, the number of ways to partition MACs increases extremely rapidly and the exact calculation becomes computationally infeasible; for example, for MAC = 50 and $K = 6$, there are $(3{,}478{,}761)^2$ possible configurations. In these scenarios, direct simulation remains the only feasible way to evaluate type I error rates for meta-analysis.

In Chapter 4, I used a resampling approach to evaluate the calibration of three aggregation association tests at $\alpha = 2.5\times10^{-6}$. For this method, I simulate 10,000 independent datasets, and resample the phenotype 1,000 times for each independent dataset, resulting in 10^8 total simulation replicates. Since generating the dataset is time-consuming, I can save computation time by resampling the phenotype for each simulated dataset. For example, for study with 1,000 cases and 1,000 controls, I can estimate type I error rates (at $\alpha = 2.5\times10^{-6}$) in 1.6 CPU-years using the resampling approach, compared to 6.3 CPU-years using independent simulations. The resampling approach produces comparable results to independent simulation.

Future work

I want to continue to develop and evaluate novel statistical methods for analyzing different types of genomic data. I will initially focus on: (1) meta-analysis method for unbalanced case-control studies, (2) genetic association methods for survival outcomes, and (3) adaptive aggregation test for analyzing rare variants on the X chromosome.

Meta-analysis method for unbalanced case-control studies

In Chapter 2, I have shown that all logistic regression tests can be anti-conservative for meta-analysis of unbalanced studies, especially for low-frequency variants. Large case-control genetics consortia often include unbalanced population-based case-control studies, which could potentially skew the meta-analysis results. Furthermore, even though study-level association results are well-calibrated or conservative in unbalanced studies, meta-analysis across studies can still be anti-conservative. Thus the calibration problem arises from the meta-analysis procedure, and not the logistic regression test. Investigating further, for low frequency variants in unbalanced studies, the empirical variance of the meta-analysis test statistic (Z_{meta}) is not well-approximated by the asymptotic variance. Hence, evaluating Z_{meta} against the asymptotic null distribution is inaccurate.

I will explore a meta-analysis method based on adjusting the reference null distribution for determining meta-analysis p-values inspired by the moments adjustment for the sequence kernel association test (SKAT) [Lee et al., 2012]. Specifically, I will readjust the moments of the null distribution of Z_{meta}, and determine meta-analysis p-values based on the adjusted (recalibrated) null distribution. I will compare the type I error rate of this method when adjusting for the first, second, and fourth moments (mean, variance, and kurtosis). I will modify the moment adjustment algorithm from the SKAT software, incorporate it into my existing case-control simulation program, and evaluate this method through additional case-control simulation.

Genetic association methods for survival outcomes

Most genome-wide association studies to date have focused on binary or quantitative traits, but there have been relatively few studies investigating the genetic effects on survival outcomes. To date, published GWAS on the survival outcomes have relatively small sample sizes, as in studies of survival in patients with small-cell lung cancer (N=139; [Han et al., 2014]) and pancreatic adenocarcinoma (N=1,005; [Wu et al., 2014]). As sequencing costs continue to decrease, I anticipate that large sequencing studies of survival outcomes will become feasible. There are relatively few statistical tools specialized for analyzing genetic studies of survival outcomes. *ProABEL* [Aulchenko et al., 2010] can analyze single variants using the Cox proportional hazards model, and *coxKM* [Lin et al., 2011] can perform SNP-set analysis based on the kernel machine Cox regression framework. I wish to (1) evaluate the calibration and power of the Cox proportional hazards regression in common and low-frequency variants, and (2) develop an efficient algorithm for the Cox proportional hazards model to feasibly test millions of genetic markers in large datasets.

Unified aggregation test for analyzing rare variants on the X chromosome

In Chapter 4, I have shown that power is slightly greater (up to 7.7%) if the coding scheme for male genotypes (e.g. with or without X-inactivation) matches the underlying genetic model, compared to when the coding scheme is misspecified. Since the true underlying genetic model is generally unknown, I propose a unified aggregation association test that is robust across different genetic models for the X chromosome, similar to Wang's [2014a] method for analyzing skewed X-inactivation. In the proposed method, I will: (1) analyze each genomic region using both X-inactivation and non-X-inactivation coding for male genotypes, (2) select the coding scheme that produces the more significant association p-value, and (3) evaluate empirical significance using numerical permutations.

Conclusion

It is a truly exciting time for human genetics research. Technological advances and decreasing costs in genotyping and sequencing have allowed investigators to examine the impact of common, low-frequency, and rare genetic variation on human traits and diseases, in increasingly larger sample sizes. Nonetheless, careful statistical analysis also plays an important role by maximizing power while guarding against false positives in each study. To this end, I evaluated the calibration and power of a variety of statistical methods for analyzing low-frequency and rare variants in genetic association studies. Based on these findings, I provided practical guidelines for analyzing different data types. These recommendations increase power to identify novel, trait-associated genetic loci, providing new targets for fine-mapping and functional studies to identify the true causal variants. With a better understanding of the trait's underlying genetic and biological mechanisms, we can ultimately develop new preventative interventions, diagnostic tests, and effective therapies to combat disease.

References

Albert A, Anderson JA. 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71:1-10.

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.

Aulchenko YS, Struchalin MV, van Duijn CM. 2010. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 11:134.

Beasley TM, Erickson S, Allison DB. 2009. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet* 39:580-95.

Buzkova P. 2013. Linear regression in genetic association studies. *PLoS ONE* 8:e56976.

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

Cox DR, Hinkley DV. 1974. Theoretical Statistics. London: Chapman and Hall.

Dersimonian R, Laird N. 1986. Meta-analysis in clinical-trials. *Control Clin Trials* 7:177-188.

Diabetes Genetics Initiative, Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of Biomedical Research, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316:1331-6.

Firth D. 1993. Bias reduction of maximum-likelihood-estimates. *Biometrika* 80:27-38.

Gottlieb MS, Root HF. 1968. Diabetes mellitus in twins. *Diabetes* 17:693-704.

Han B, Eskin E. 2011. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 88:586-598.

Han JY, Lee YS, Shin ES, Hwang JA, Nam S, Hong SH, Ghang HY, Kim JY, Yoon SJ, Lee JS. 2014. A genome-wide association study of survival in small-cell lung cancer patients treated with irinotecan plus cisplatin chemotherapy. *Pharmacogenomics J* 14:20-27.

Hauck WW, Donner A. 1977. Wald's test as applied to hypotheses in logit analysis. *J Am Stat Assoc* 72:851-853.

Heinze G, Schemper M. 2002. A solution to the problem of separation in logistic regression. *Stat Med* 21:2409-2419.

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 et al. 2013. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* 45:197-201.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437:1299-1320.

Kang HM. 2012. EPACTS: efficient and parallelizable association container toolbox. University of Michigan: Department of Biostatistics and Center for Statistical Genetics. Available at: http://www.sph.umich.edu/csg/kang/epacts/.

Kaprio J, Tuomilehto J, Koskenvuo M, Romanov K, Reunanen A, Eriksson J, Stengard J, Kesaniemi YA. 1992. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* 35:1060-7.

Kulzer JR, Stitzel ML, Morken MA, Huyghe JR, Fuchsberger C, Kuusisto J, Laakso M, Boehnke M, Collins FS, Mohlke KL. 2014. A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am J Hum Genet* 94:186-97.

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.

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.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311-321.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34:816-834.

Lin DY, Zeng D. 2010. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet Epidemiol* 34:60-66.

Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, Lin X. 2011. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genet Epidemiol* 35:620-631.

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

Lumley T, Diehr P, Emerson S, Chen L. 2002. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health* 23:151-69.

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

Ma C, Blackwell T, Boehnke M, Scott LJ. 2013. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol* 37:539-550.

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

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22:719-748.

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.

Mehta CR, Patel NR. 1995. Exact logistic-regression - theory and examples. *Stat Med* 14:2143-2160.

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.

Ploner M, Dunkler, D., Southworth, H., Heinze, G. 2010. logistf: Firth's bias reduced logistic regression. Version Version 1.10. Medical University of Vienna: Center for Medical Statistics, Informatics and Intelligent Systems. Available at: http://CRAN.R-project.org/package=logistf.

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 et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-575.

R Development Core Team. 2012. R: A language and environment for statistical computation. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.r-project.org/.

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.

Schizophrenia Psychiatric Genome-Wide Association Study Consortium. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 43:969-976.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341-1345.

Shea J, Agarwala V, Philippakis AA, Maguire J, Banks E, Depristo M, Thomson B, Guiducci C, Onofrio RC, Kathiresan S et al. 2011. Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat Genet* 43:801-5.

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, De Lange M, Harris JR, Hjelmborg JV et al. 2003. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res* 6:399-408.

Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S et al. 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 39:770-775.

Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadottir HT, Johannsdottir H, Magnusson OT, Gudjonsson SA et al. 2014. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* 46:294-8.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.

The GoT2D Consortium. Interrogation of type 2 diabetes loci, variants, and mechanisms through whole genome sequencing of 2,657 cases and controls. *(In preparation)*.

Upton GJG. 1982. A comparison of alternative tests for the 2x2 comparative trial. *J Roy Statist Soc Ser A* 145:86-105.

Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J et al. 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 8:e1002793.

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

Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, Zong X, Laplana M, Wei Y, Han Y et al. 2014b. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* 46:736-41.

Willer CJ, Li Y, Abecasis GR. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26:2190-2191.

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 C, Kraft P, Stolzenberg-Solomon R, Steplowski E, Brotzman M, Xu M, Mudgal P, Amundadottir L, Arslan AA, Bueno-de-Mesquita HB et al. 2014. Genome-wide association study of survival in patients with pancreatic adenocarcinoma. *Gut* 63:152-160.

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.

Xing G, Lin CY, Wooding SP, Xing C. 2012. Blindly using Wald's test can miss rare disease-causal variants in case-control association studies. *Ann Hum Genet* 76:168-177.

Yates F. 1934. Contingency tables involving small numbers and the χ^2 test. *Supp J Roy Statist Soc* 1:217-235.

Zeggini E, Rayner W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI. 2005. An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat Genet* 37:1320-1322.

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