RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants

Clement Ma, Tom Blackwell, Michael Boehnke, Laura J. Scott,* and the GoT2D investigators

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

Received 25 February 2013; Revised 12 May 2013; accepted revised manuscript 20 May 2013.
Published online 20 June 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21742

ABSTRACT: In genome-wide association studies of binary traits, investigators typically use logistic regression to test common variants for disease association within studies, and combine association results across studies using meta-analysis. For common variants, logistic regression tests are well calibrated, and meta-analysis of study-specific association results is only slightly less powerful than joint analysis of the combined individual-level data. In recent sequencing and dense chip based association studies, investigators increasingly test low-frequency variants for disease association. In this paper, we 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. We 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. We demonstrate for low-count variants (roughly minor allele count [MAC] < 400) 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. We also establish MAC as the key parameter determining test calibration for joint and meta-analysis.
Genet Epidemiol 37:539–550, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: meta-analysis; joint analysis; single variant tests; single nucleotide polymorphisms; low-frequency 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 $N = 4{,}000$ individuals, or MAF < 0.01 for $N = 20{,}000$. For a given study design with $N > 2{,}000$, 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

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

*Correspondence to: Laura J. Scott, Department of Biostatistics and Center for Statistical Genetics, University of Michigan, M4134 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: ljst@umich.edu

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

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

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

$$W = \hat{\beta} \big/ SE(\hat{\beta}) \quad (2)$$

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

$$LR = -2[l(\tilde{\alpha}, 0) - l(\hat{\alpha}, \hat{\beta})] \quad (3)$$

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

$$S = U_\beta \Big/ \sqrt{\text{var}(U_\beta)} \quad (4)$$

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

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

where $I_{AB} = -\partial l^2 / \partial A \partial B$ is the AB component of the observed Fisher's 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^2_1 distribution.

In logistic regression models, "separation" occurs when cases and controls can be perfectly explained by a nontrivial 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 (e.g., 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 $\pm\infty$. Firth [1993] proposed a penalized likelihood function to correct the first-order asymptotic bias of parameter estimates that is especially relevant for separated datasets. The Firth bias-corrected log-likelihood function is

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

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

$$F = -2[l^*(\tilde{\alpha}^*, 0) - l^*(\hat{\alpha}^*, \hat{\beta^*})] \quad (5)$$

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^2_1 distribution. We modified Ploner's R implementation of the bias-corrected logistic regression test [Ploner et al., 2010] to increase computational efficiency, and include 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)–(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 (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)–(5)) accordingly.

In meta-analysis, we perform a separate association test within each study and combine the study-level association results (e.g., using P-values and directions of effect, transformed into z-scores). For each asymptotic test (equations (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 from a χ^2_1 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 [Han and Eskin, 2011; Mantel and Haenszel, 1959].

Analytical Calculation of Type I Error Rate for Joint Analysis

For joint analysis, we calculate type I error rates for significance levels $\alpha = 5 \times 10^{-5}$ 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 [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

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

where $\Pr[T_1 = i, T_0 = j]$ is the probability for the (i,j)th configuration and $I[P - value_{ij} \leq \alpha]$ 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 ($\alpha = 5 \times 10^{-8}$) 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 (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 10,000/10,000, 8,000/12,000, 5,000/15,000, and 1,000/19,000 cases and controls, respectively. We subdivide each full dataset into $K = 10$ equal-sized substudies with identical case-control ratios, analyze each substudy separately, and meta-analyze the substudy association results. We perform up to 10 million simulation replicates under the null model (OR = 1) to estimate type I error rates at $\alpha = 5 \times 10^{-4}$ or 5×10^{-5}, and 10,000 replicates under alternative models (OR > 1) to estimate power at $\alpha = 5 \times 10^{-8}$.

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 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

$\alpha = 5 \times 10^{-8}$. For sample-size weighted meta-analysis, we estimate type I error using simulation at a less stringent threshold ($\alpha = 5 \times 10^{-4}$ [Supplementary Fig. S2] or 5×10^{-5}) due to computational constraints. For both joint and meta-analysis, we estimate power using simulation at $\alpha = 5 \times 10^{-8}$ 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 ($\alpha = 5 \times 10^{-8}$) for a single balanced study with 10,000 cases and 10,000 controls (Fig. 1A). For high-count variants (expected MAC > 400; MAF > 0.01 for $N = 20{,}000$), we focus on type I error estimates for a variant with expected MAC = 2,000 (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 ($\alpha = 5 \times 10^{-5}$) for meta-analysis of 10 balanced substudies each with 1,000 cases and 1,000 controls (Fig. 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; Fig. 1D) is less conservative than the score test based meta-analysis (blue dashed line; Fig. 1G). For example, at E[MAC] = 40 (MAF = 0.001), the empirical type I error rate (at $\alpha = 5 \times 10^{-5}$) 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 5,000/15,000 (1:3) and 1,000/19,000 (1:19) cases and controls, respectively (Fig. 1B and 1C). 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 analysis type I error rates at $\alpha = 5 \times 10^{-8}$ across a wide range of case-control ratios for high count (Fig. 2A) and low-count (Fig. 2B) variants.

Finally, we examine type I error rates for meta-analysis of 10 unbalanced substudies each with 500/1,500 (1:3) or 100/1,900 (1:19) cases and controls. For high-count variants, in a 1:3 study, all meta-analysis tests are well calibrated (Fig. 1H); in a 1:19 study, meta-analysis of Firth, score, and likelihood ratio test results can be slightly anti-conservative (Fig. 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 (Fig. 2F).

Power of Joint and Meta-Analysis Tests

We first examine the power ($\alpha = 5 \times 10^{-8}$) for joint and meta-analysis tests in balanced studies. For high-count variants (E[MAC] = 2,000; MAF = 0.05), all tests have near identical power for both joint and meta-analysis, as expected [Lin and Zeng, 2010] (Fig. 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 (Fig. 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; Fig. 3B) or similar (1:19 study; Fig. 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 (Fig. 3E and 3F). Nonetheless, we again observe some correspondence between increased test conservativeness and reduced test power in unbalanced studies.

Consistent Test Calibration With Fixed Total MAC

All of the results shown so far (Figs. 1–3) refer to analyses with a total sample size of $N = 20{,}000$ individuals. Here, we examine joint analysis (Fig. 4, Supplementary Fig. S1; $\alpha = 5 \times 10^{-8}$) and meta-analysis (Supplementary Fig. S2; $\alpha = 5 \times 10^{-4}$) 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 ($N = 2{,}000$–$50{,}000$) 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 $N > 2{,}000$).

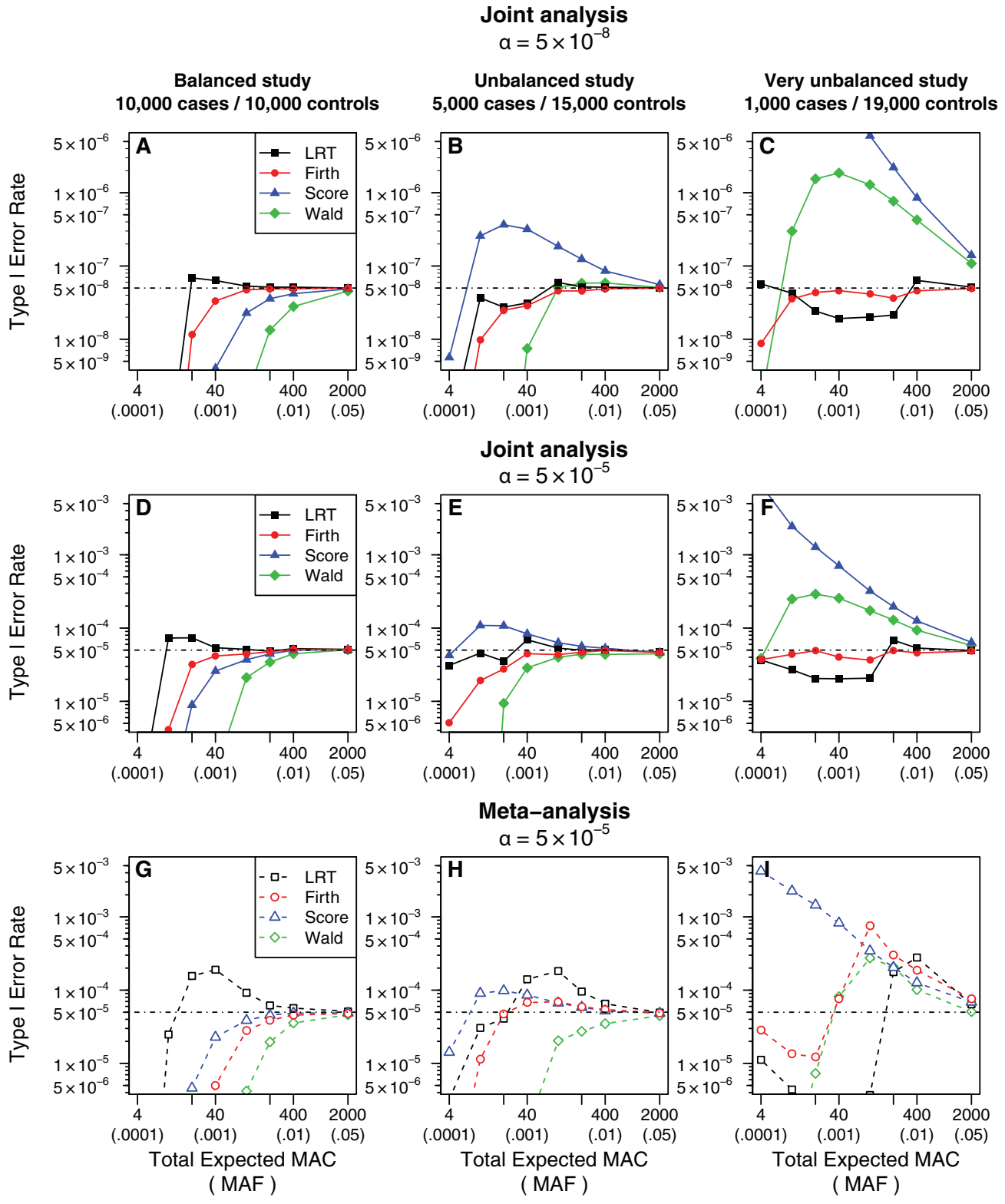

Figure 1. Type I error rates by minor allele count (MAC) for logistic regression tests in joint and meta-analysis. (A–C) Analytically calculated type I error rates ($\alpha = 5 \times 10^{-8}$) for joint analysis; (D–F) empirical type I error rates ($\alpha = 5 \times 10^{-5}$) for joint analysis; and (G–I) empirical type I error rates ($\alpha = 5 \times 10^{-5}$) for sample-size weighted meta-analysis. Type I error rates for joint analysis are estimated for studies with 10,000/10,000, 5,000/15,000, and 1,000/19,000 total cases and controls; meta-analysis is based on partitioning the full dataset into 10 equal-sized substudies. 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.

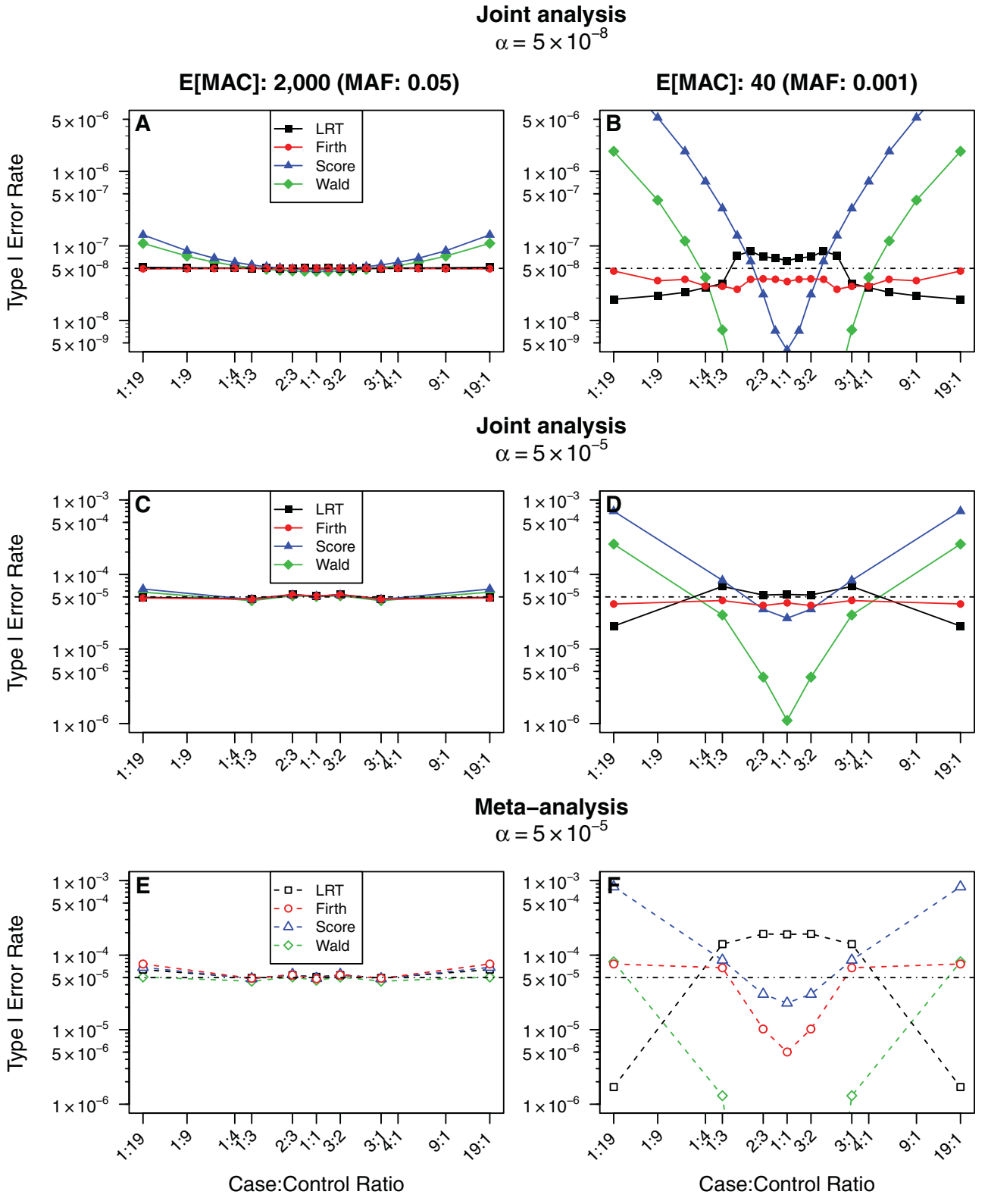

Figure 2. Type I error rates by case-control ratio for logistic regression tests in joint and meta-analysis. (A, B) Analytically calculated type I error rates ($\alpha = 5 \times 10^{-8}$) for joint analysis; (C, D) empirical type I error rates ($\alpha = 5 \times 10^{-5}$) for joint analysis; and (E, F) empirical type I error rates ($\alpha = 5 \times 10^{-5}$) for sample-size weighted meta-analysis. Type I error rates are estimated for a high count (expected MAC = 2,000; MAF = 0.05), and low-count (E[MAC] = 40; MAF = 0.001) variant, in studies with N = 20,000 individuals and varying case-control ratios. The horizontal dotted line denotes the corresponding nominal significance threshold.

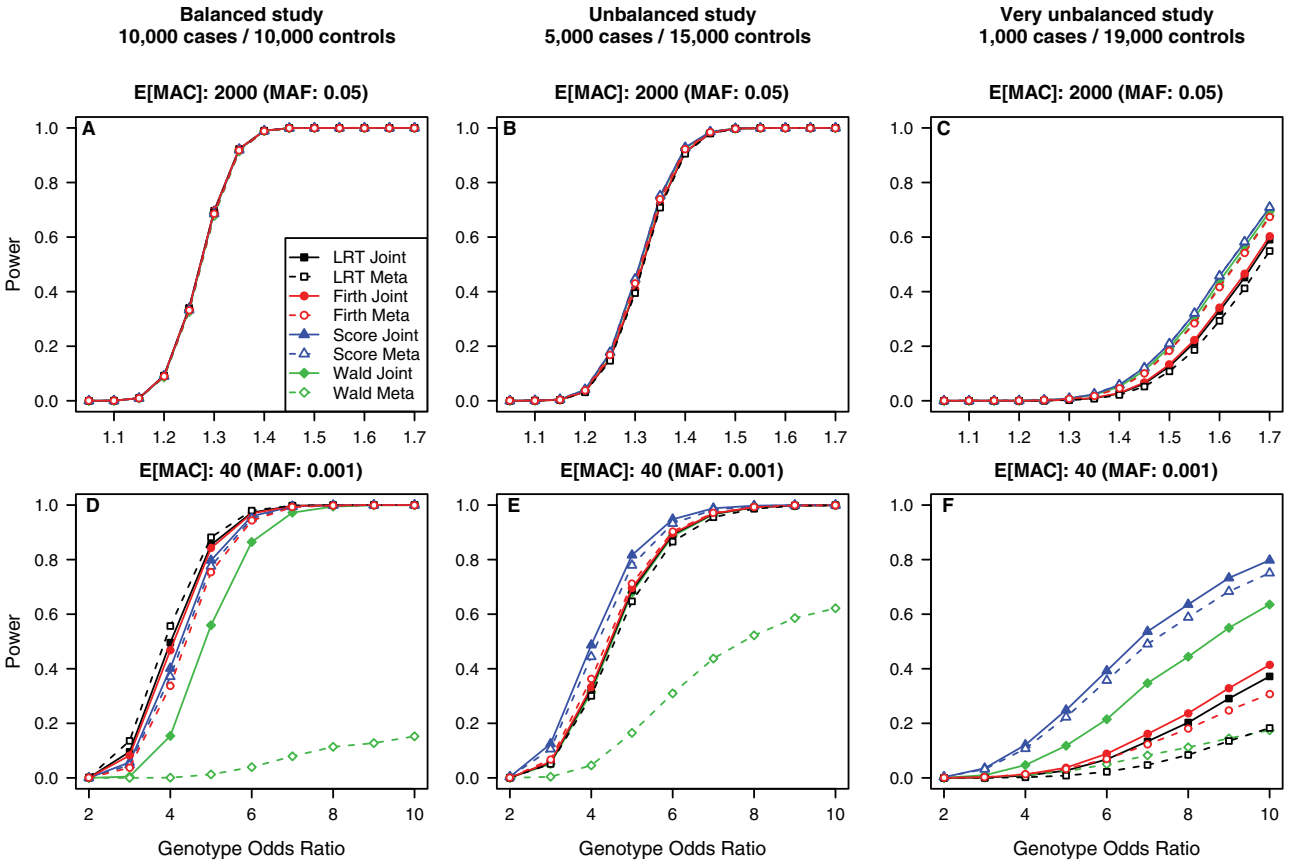

Figure 3. Simulation-based power curves for joint and meta-analysis. Simulated power ($\alpha = 5 \times 10^{-8}$) in joint analysis and sample-size weighted meta-analysis for (A–C) a high-count variant (expected MAC = 2,000; MAF = 0.05); and (D–F) a low-count variant (E[MAC] = 40; MAF = 0.001). Power for joint analysis is estimated for studies with 10,000/10,000, 5,000/15,000, and 1,000/19,000 total cases and controls; meta-analysis is based on partitioning the full dataset into 10 equal-sized substudies.

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 (Fig. 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 higher 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 observed MAC = 40 in a study of N = 20,000 individuals (Fig. 5, upper panels). In Figure 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).

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 (Fig. 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 (Fig. 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).

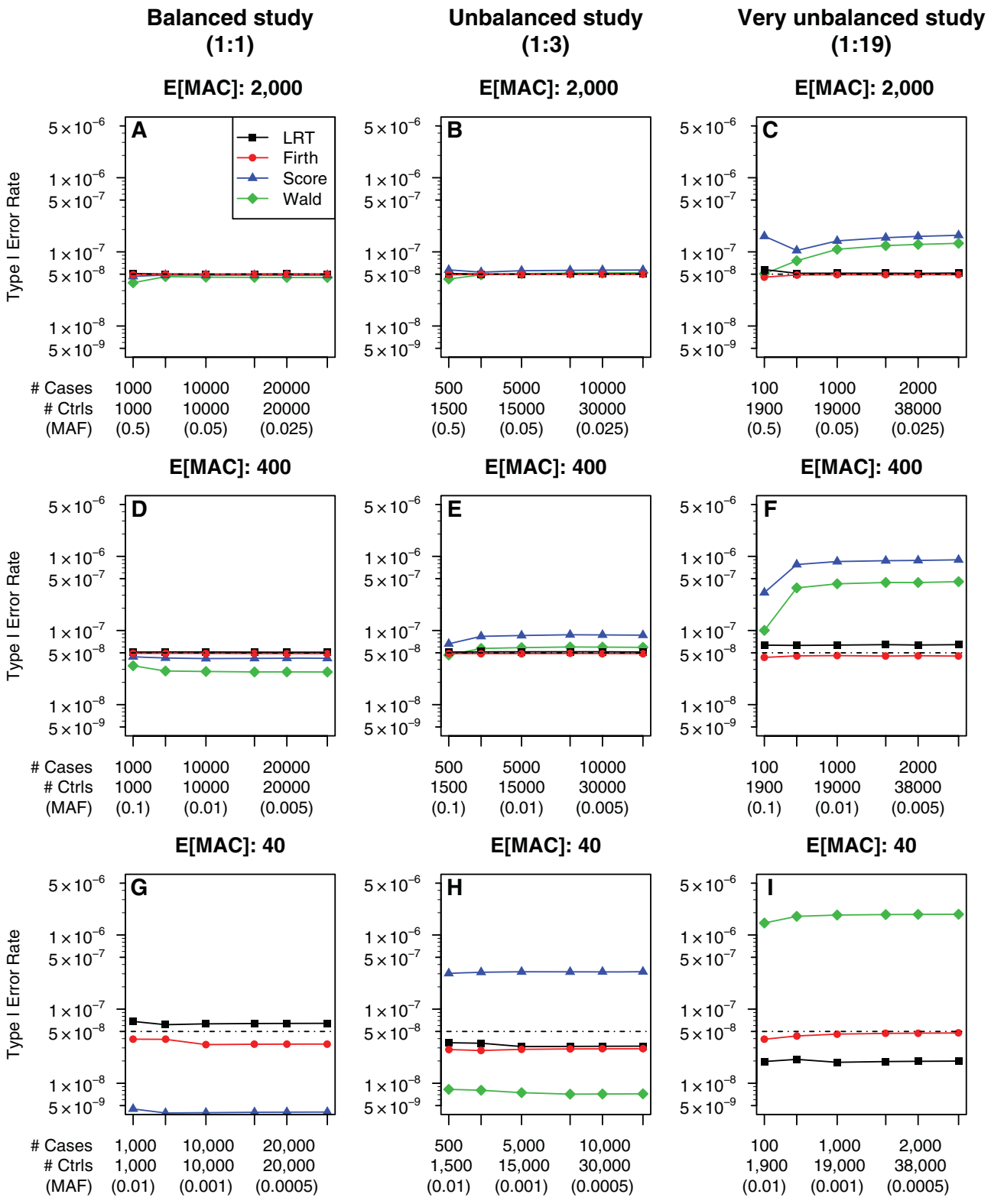

Figure 4. Joint analysis type I error rates by sample size for fixed expected minor allele count (MAC). Analytically calculated joint analysis type I error rates for single balanced (case-control ratio 1:1), unbalanced (1:3), and very unbalanced studies (1:19) of various sample sizes. For each study, variant allele frequencies are selected so that variants have (A–C) expected MAC = 2,000; (D–F) expected MAC = 400; or (G–I) expected MAC = 40. The horizontal dotted line denotes the corresponding nominal significance threshold ($\alpha = 5 \times 10^{-8}$). Very conservative or anti-conservative tests with type I error rates that exceed the vertical axis scale are not displayed.

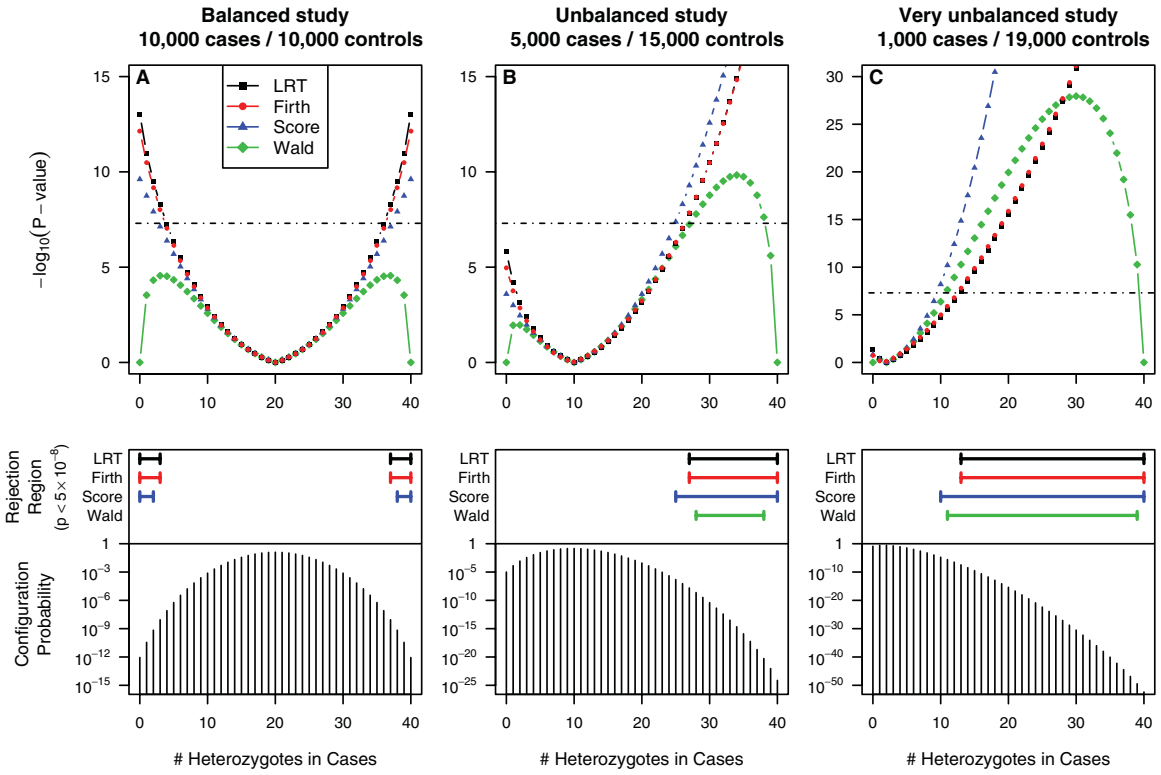

Figure 5. Logistic regression *P*-value distributions for fixed total minor allele count (MAC). For a variant with MAC = 40, the upper panels display *P*-values for all 41 possible allele configurations for each test in a single study of (A) 10,000/10,000, (B) 5,000/15,000, and (C) 1,000/19,000 cases and controls, respectively. The horizontal dotted line denotes the corresponding nominal significance threshold ($\alpha = 5 \times 10^{-8}$). The lower panels display horizontal bars indicating the rejection region (*P*-value $< 5 \times 10^{-8}$) for each test and hypergeometric probabilities of each allele configuration.

Tests with the highest to lowest type I error rates (likelihood ratio, Firth, score, Wald; Fig. 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 (Fig. 5C, upper panel). Again, this corresponds to the total configuration probability encompassed by the rejection regions (Fig. 5C, lower panel), and the least to most conservative tests (Fig. 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 (i.e., [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; Fig. 3D–F) and should not be used.

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 comprised 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 in the total sample.

For high-count variants (400 < MAC ≤ 908), score test based meta-analysis and Firth test based joint analysis produce similar *P*-values (Fig. 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 (Fig. 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 Supporting Information (Supplementary Figs. S3 and S4).

Discussion

Recommendations

For analysis of high-count variants (MAC > 400), in balanced and moderately unbalanced (1:3) studies, joint and

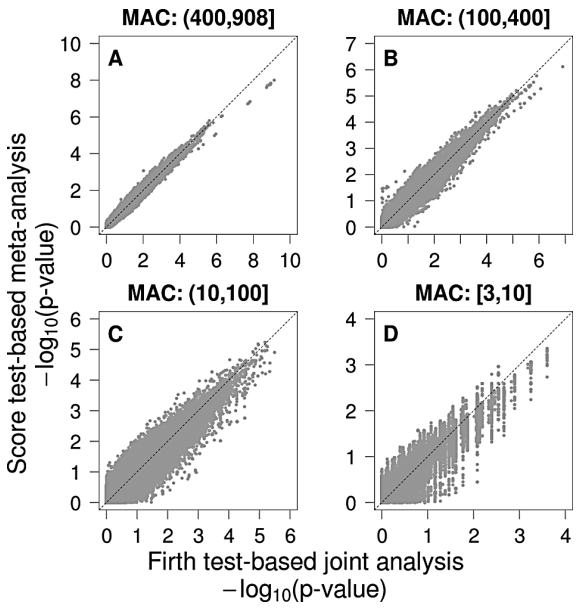

Figure 6. Comparison of score test based meta-analysis and Firth test based joint analysis *P*-values in the GoT2D study. Results for 8.58 million variants are shown, with 2.4, 2.4, 2.6, and 1.2 million variants in the highest to lowest minor allele count (MAC) categories, respectively.

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. In particular, the score test is not recommended for studies with case-control ratios < 1:3 or > 3:1.

Use of MAC Rather Than MAF in Describing Test Calibration

We present our recommendations using a rough MAC threshold, rather than an MAF threshold, since test calibration remains consistent as long as MAC is constant (given $N > 2{,}000$, 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 total MAC 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 substudy, since even variants with a single observed minor allele may contribute to the overall meta-analysis. Imposing a more stringent study-level MAC filter leads to more conservative and less-powerful meta-analysis results (Supplementary Fig. S5). 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 ($\alpha < 5 \times 10^{-8}$). For a given fixed total N, we observe that meta-analysis of many small substudies is more conservative and less powerful than meta-analysis of a few larger substudies (Supplementary Fig. S6). Smaller substudies are more likely to be monomorphic for low-count variants, and so are effectively removed from the meta-analysis. Practically, the time and effort needed to analyze and prepare a very small study for meta-analysis may outweigh the potential contribution of that study.

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 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 a disease with 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 disease 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, though slightly more variable, 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 (Supplementary Figs. S7–S9), 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 the Appendix); simulations show that inverse-variance weighted meta-analysis of Firth or Wald test results in unbalanced studies is also anti-conservative (Supplementary Figs. S7–S9). 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 substudy 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 2,000 cases and 18,000 controls, randomly removing 12,000 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 (e.g., 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 higher 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.

Acknowledgments

We thank Hyun Min Kang for helpful discussions and for including relevant tests in his EPACTS software, Georg Heinze and Peter X. K. Song for helpful discussions regarding bias-corrected logistic regression, and our GoT2D colleagues for allowing us to use an early sequence data freeze. This research was supported by the National Institutes of Health grants HG000376 and DK088389 to M.B.

References

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

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.

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

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.

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.

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

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

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.

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.

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.

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

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

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/

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

Upton GJG. 1982. A comparison of alternative tests for the 2×2 comparative trial. *J R Stat Soc Ser A* 145:86–105.

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

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 R Stat Soc* 1:217–235.

APPENDIX

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:

$$\tilde{\beta}_{IV} = \sum_{k=1}^{K} V_k^{-1}\beta_k \Big/ \sum_{k=1}^{K} V_k^{-1}$$

$$SE\left(\tilde{\beta}_{IV}\right) = \left[\sqrt{\sum\nolimits_{k=1}^{K} V_k^{-1}}\right]^{-1}$$

$$Z_{IV} = \tilde{\beta}_{IV} \Big/ SE\left(\tilde{\beta}_{IV}\right)$$

This method is only applicable for statistical tests that estimate the effect size and its standard error, and so cannot be used for the score test or FET.