Received: 4 July 2018 | Revised: 9 January 2019 | Accepted: 1 February 2019

DOI: 10.1002/gepi.22197

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Robust meta-analysis of biobank-based genome-wide association studies with unbalanced binary phenotypes

Rounak Dey[1] | Jonas B. Nielsen[2] | Lars G. Fritsche[1] | Wei Zhou[3] | Huanhuan Zhu[4] | Cristen J. Willer[3,5,6] | Seunggeun Lee[1]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan

[2]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark

[3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan

[4]Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan

[5]Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, Michigan

[6]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan

Correspondence
Seunggeun Lee, Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor 48109, Michigan.
Email: leeshawn@umich.edu

Funding information
National Human Genome Research Institute, Grant/Award Number: R01HG008773

Abstract

With the availability of large-scale biobanks, genome-wide scale phenome-wide association studies are being instrumental in discovering novel genetic variants associated with clinical phenotypes. As increasing number of such association results from different biobanks become available, methods to meta-analyse those association results is of great interest. Because the binary phenotypes in biobank-based studies are mostly unbalanced in their case–control ratios, very few methods can provide well-calibrated tests for associations. For example, traditional Z-score-based meta-analysis often results in conservative or anticonservative Type I error rates in such unbalanced scenarios. We propose two meta-analysis strategies that can efficiently combine association results from biobank-based studies with such unbalanced phenotypes, using the saddlepoint approximation-based score test method. Our first method involves sharing the overall genotype counts from each study, and the second method involves sharing an approximation of the distribution of the score test statistic from each study using cubic Hermite splines. We compare our proposed methods with a traditional Z-score-based meta-analysis strategy using numerical simulations and real data applications, and demonstrate the superior performance of our proposed methods in terms of Type I error control.

KEYWORDS

biobank, case–control studies, GWAS, meta-analysis, saddlepoint approximation

1 | INTRODUCTION

Genome-wide scale phenome-wide association analysis (Hebbring, 2014) is gaining increasing attention in the human genetics community in the recent years. The availability of detailed phenotypic information from the electronic health record (EHR) systems in large biobanks as well as the recent advancements in genotyping and imputation technologies (Das et al., 2016) are allowing researchers to phenotype thousands of traits and genotype tens of millions of variants in large cohort studies. Several biobank studies, including UK-Biobank (Bycroft et al., 2017), Michigan Genomics Initiative and Nord-Trøndelag Health Study (Krokstad et al., 2013) currently attempt to test for associations in all genotype–phenotype pairs, which results in billions of tests. These large-scale analyses have great potential to find novel genotype–phenotype associations, which will help uncover underlying molecular mechanism of clinical phenotypes.

In a typical phenome-wide association study (PheWAS) in biobanks, most of the phenotypes are binary with unbalanced (1:5) or often extremely unbalanced (1:500) case–control ratios, which results in performing 1,000s of unbalanced case–control genome-wide association studies (GWASs). For example, ~1,400 case–control studies in the UK Biobank interim release data have more than 100 controls/case (see Histogram in Figure S1 in Supporting Information Material A). Under such case–control imbalance, the standard asymptotic tests such as the Wald test, score test, and likelihood ratio test can severely inflate the Type I errors resulting in several spurious associations, especially for the low frequency (0.01 < minor allele frequency [MAF] < 0.05) and rare (MAF < 0.01) variants (Dey, Schmidt, Abecasis, & Lee, 2017; Ma, Blackwell, Boehnke, Scott, & investigators, 2013). To obtain well-calibrated *P* values in such situations, Ma et al. (2013) proposed to use the Firth's penalised likelihood ratio test (Firth, 1993). Since the Firth's test is computationally too expensive to be used for billions of association tests, Dey et al. (2017) developed a fast saddlepoint approximation (SPA)-based score test, fastSPA, which is computationally much faster than the Firth's test.

As more and more association results from different biobanks become available, meta-analysing (Evangelou & Ioannidis, 2013) the results from the unbalanced GWASs is the logical next step to improve the power to detect novel genotype–phenotype associations. *Z*-score-based approach (Cooper, Hedges, & Valentine, 2009) which converts *P* values to normal *Z* scores for combining multiple study *P* values, has been a standard meta-analysis method in GWASs (Evangelou & Ioannidis, 2013). However, even though *P* values from fastSPA and Firth's test are well calibrated in a single study, combining them through *Z*-score method can fail to control for Type I errors. Ma et al. (2013) has shown that combining Firth's test-based *P* values through *Z*-score method can produce conservative or anticonservative behaviours especially when the case–control ratio is unbalanced and the variant minor allele count (MAC) is small. This may be because the study-specific *P* values have discrete distribution due to case–control imbalance and small MAC. As shown in our simulation studies, the same problem also occurs in the meta-analysis using fastSPA-based *P* values. To facilitate the meta-analysis of the biobank-based GWASs, we need a robust method to control for Type I errors regardless of case–control ratios and MAC.

In this paper, we first evaluate the performance of the *Z*-score-based meta-analysis procedure using the fastSPA test-based *P* values under extensive simulation settings and real data sets, and propose two alternative meta-analysis strategies to obtain well-calibrated meta-analysis *P* values. The first method involves sharing the overall number of homozygous minor and heterozygous genotypes for each genetic variant, in addition to the case–control sample size and *P* value shared in the *Z*-score-based meta-analysis strategy. The second method involves sharing the observed within-study score statistics and the cumulant generating functions (CGF) of those score statistics using a spline-based approach, which will be used to carry out SPA to obtain the meta-analysis *P* value. The additional information facilitates approximating the distributions of the study-specific score statistics, which can be discrete, asymmetric and different from the traditionally used normal distribution. Through extensive simulation studies and an analysis of the UK Biobank data, we show that the proposed methods can control the Type I error rates and retain similar power as a joint analysis as well as being scalable to large-scale PheWASs.

2 | METHODS

2.1 | Model for single study association test and SPA

We consider J case–control studies, where the jth study has sample size n_j. Within each individual study, we follow the regression model and testing procedure described in Dey et al. (2017). For the ith subject in the jth study, let $Y_i^{(j)} = 1$ or 0 denote the case–control status, $X_i^{(j)}$ denote the $k \times 1$ vector of nongenetic covariates (including the intercept) and $G_i^{(j)} = 0, 1, 2$ denote the number of minor alleles of the variant to be tested. Let $\beta^{(j)}$ be the $k \times 1$ vector of coefficients for the nongenetic covariates and $\gamma^{(j)}$ be the genotype log odds ratio. We use the following logistic regression model to perform association test in the jth study.

$$\text{logit }\left[\Pr\left(Y_i^{(j)} = 1 | X_i^{(j)}, G_i^{(j)}\right)\right] = X_i^{(j)^T}\beta^{(j)} + G_i^{(j)}\gamma^{(j)} \text{ for } i = 1, 2, \ldots, n_j. \quad (2.1)$$

Let $\hat{\mu}_i^{(j)}$ be the maximum likelihood estimator of $\mu_i^{(j)} = \Pr(Y_i^{(j)} = 1 | X_i^{(j)})$ under the null hypothesis $H_0 : \gamma^{(j)} = 0$. Further, let $X^{(j)} = (X_1^{(j)^T}, \ldots, X_{n_j}^{(j)^T})$ be the $n_j \times k$ matrix of covariates, $G^{(j)} = (G_1^{(j)}, \ldots, G_n^{(j)})^T$ be the genotype vector, $W^{(j)}$ be a diagonal matrix with ith diagonal element $\hat{\mu}_i^{(j)}(1 - \hat{\mu}_i^{(j)})$, and $\tilde{G}^{(j)} = G^{(j)} - X^{(j)}(X^{(j)^T}W^{(j)}X^{(j)})^{-1} X^{(j)^T}W^{(j)}G^{(j)}$ be the covariate-adjusted genotype vector. Then, the score statistic for testing $H_0 : \gamma^{(j)} = 0$ will be $S^{(j)} = \sum_{i=1}^{n_j} \tilde{G}_i^{(j)}(Y_i^{(j)} - \hat{\mu}_i^{(j)})$. To apply the SPA-based score test, we first need to calculate the CGF of the score statistic and its first and second derivatives given by

$$K^{(j)}(t) = \sum_{i=1}^{n_j} \log\left(1 - \hat{\mu}_i^{(j)} + \hat{\mu}_i^{(j)} e^{\tilde{G}_i^{(j)} t}\right) - t \sum_{i=1}^{n_j} \tilde{G}_i^{(j)} \hat{\mu}_i,$$

$$K'^{(j)}(t) = \sum_{i=1}^{n_j} \frac{\hat{\mu}_i^{(j)} \tilde{G}_i^{(j)}}{\left(1 - \hat{\mu}_i^{(j)}\right) e^{-\tilde{G}_i^{(j)} t} + \hat{\mu}_i^{(j)}} - \sum_{i=1}^{n_j} \tilde{G}_i^{(j)} \hat{\mu}_i^{(j)},$$

$$K''^{(j)}(t) = \sum_{i=1}^{n_j} \frac{(1 - \hat{\mu}_i^{(j)}) \hat{\mu}_i^{(j)} \tilde{G}_i^{(j)2} e^{-\tilde{G}_i^{(j)} t}}{\left[\left(1 - \hat{\mu}_i^{(j)}\right) e^{-\tilde{G}_i^{(j)} t} + \hat{\mu}_i^{(j)}\right]^2}.$$

Using the SPA method (Barndorff-Nielsen, 1990; Daniels, 1954), the distribution function of $S^{(j)}$ at the observed score statistic s can be approximated by

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

where $w = \mathrm{sgn}(\hat{t})\sqrt{2(\hat{t}s - K^{(j)}(\hat{t}))}$, $v = \hat{t}\sqrt{K''^{(j)}(\hat{t})}$, $\hat{t}$ is the solution to the equation $K'^{(j)}(\hat{t}) = s$, and Φ is the standard normal distribution function. The fastSPA (Dey et al., 2017) test implements a faster version of this SPA method, which can be applied to obtain the P value $p^{(j)}$. One of the steps implemented in the fastSPA test is to apply the SPA method only if the score statistic lies outside a certain standard deviation (*SD*) threshold from the mean. If the score statistic lies inside the *SD* threshold, then the fastSPA test uses the normal approximation to calculate the P values because the normal approximation behaves well near the mean. In this paper, we will consider the P values using two such *SD* threshold, 2 and 0.1, and will denote the tests by fastSPA-2 and fastSPA-0.1, respectively.

2.2 | *P* value-based meta-analysis and normal distribution-based *Z*-score method

We first introduce a framework for P value-based meta-analysis. In this framework, the study-specific P values ($p^{(j)}s$) are inverted to obtain the signed scores $R^{(j)}s$ using some distributions $F^{(j)}$s, for $j = 1,...,J$, where the signs are determined by the directions of associations. We call $F^{(j)}$s reference distributions. Then, the meta-analysis score is given by $R_{\text{meta}} = \sum_{j=1}^{J} R^{(j)}$ where each $R^{(j)} \sim F^{(j)}$ under the null hypothesis of no association. Traditional Z-score-based meta-analysis is a special case of this framework, where the reference distributions are normal distributions with means zero and variances given by the effective sample sizes of the individual studies. The effective sample size (Han & Eskin, 2011) is calculated as $n_j^* = 4n_{j1}n_{j0}/n_j$, where n_{j1} and n_{j0} are the number of cases and controls in the jth study, respectively. This meta-analysis method first inverts the P values using a standard normal distribution to obtain the signed Z scores $Z^{(j)} = \pm\Phi^{-1}(p^{(j)}/2)$, where the signs depend on the directions of associations. Then, the scores $R^{(j)}$s are calculated as $R^{(j)} = \sqrt{n_j^*}Z^{(j)}$, for $j = 1,...,J$, and the meta-analysis score is given by $R_{\text{meta}} = \sum_{j=1}^{J} R^{(j)} \sim N\left(0, \sum_{j=1}^{J} n_j^*\right)$ under the null hypothesis. We can test the null hypothesis of no association between the phenotype and the variant by testing $Z_{\text{meta}} = R_{\text{meta}}/\sqrt{\sum_{j=1}^{J} n_j^*}$, which follows $N(0, 1)$ under the null hypothesis.

This meta-analysis strategy can control for Type I error rates when each study-specific P value follows the uniform distribution. When the case–control is unbalanced and variants are rare, however, each study-specific test statistic $S^{(j)}$ can have a discrete and often very skewed null distribution, which can result in the set of possible study-specific P values to be discrete, and the two-sided probabilities that constitute those P values, to be asymmetric. In such situations, although SPA can be applied to control Type I error rates within each individual study, inverting such SPA-based P values to normally distributed Z scores might not be appropriate, and can introduce systematic bias.

We notice that the best possible reference distribution $F^{(j)}$ would be the null distribution of the score statistic $S^{(j)}$ under Model (2.1) (let it be $\tilde{F}^{(j)}$). In that case, $R^{(j)}$s will be the same as $S^{(j)}$s. Within each individual study, $\tilde{F}^{(j)}$ can be approximated based on the CGF of the score statistic, using the SPA method. However, it is difficult to share the CGFs as summary level statistics. In our first method, we suggest sharing the overall genotype counts (GCs) from the individual studies to construct our reference distributions. For the second approach, we propose a simpler technique to approximate $\tilde{F}^{(j)}$s using summary level statistics and suggest sharing $S^{(j)}$s instead of the P values so that we can directly use $R^{(j)} = S^{(j)}$. This is equivalent to a P value-based meta-analysis using the approximations of $\tilde{F}^{(j)}$s as the reference distributions $F^{(j)}$s, because $R^{(j)}$s will closely approximate $S^{(j)}$s when $F^{(j)}$s closely approximate $\tilde{F}^{(j)}$. Although our approaches require more information than just the P values, case–control sample sizes and directions of associations, the additional information is also summary level information and hence does not need individual level data.

2.3 | GC-based method

Here we propose a practical approach to approximate the CGFs using the GCs (number of 0, 1, and 2 genotypes) at different markers. For rare variants where homozygous minor genotypes are usually not present in the data, or for variants that follow Hardy–Weinberg equilibrium, sharing only the MACs will be sufficient, as the GCs can be easily calculated based on the MACs.

Suppose, for the jth study, the GCs for the variant to be tested are m_{j0}, m_{j1}, and m_{j2} ($m_{j0} + m_{j1} + m_{j2} = n_j$) corresponding to the genotypes 0, 1, and 2, respectively.

Then, we can construct the genotype vector $G^{(j)*}$ of length n_j where the first m_{j2} elements are 2s, next m_{j1} elements are 1s, and the rest are 0s. We propose using the null distribution (let it be $F^{(j)*}$) of the score statistic in the following genotype-only Model (2.2) as our reference distribution,

$$\text{logit } [\Pr(Y_i^{(j)} = 1 | G_i^{(j)*})] = \alpha^{(j)*} + \gamma^{(j)*} G_i^{(j)*}, \quad (2.2)$$

where $G_i^{(j)*}$ is the ith elements of $G^{(j)*}$, $\alpha^{(j)*}$ is the intercept, and $\gamma^{(j)*}$ is the genotype log odds ratio. Intuitively, when the nongenetic covariates are relatively balanced across cases and controls, the discreteness and asymmetry in the null distribution of the score statistic mainly depend on the imbalance or the rarity of the phenotype and the genotype. Therefore, the null distribution of the score statistic under the genotype-only model can be a reasonable alternative to the traditionally used normal distribution, as a reference distribution. To apply this method, we first need to calculate the CGF of the score statistic and its first and second derivatives in the genotype-only Model (2.2) given by

$$K^{(j)*}(t) = \sum_{i=1}^{n_j} \log\left(1 - \hat{\mu}^{(j)*} + \hat{\mu}^{(j)*} e^{\overline{G_i^{(j)*}} t}\right),$$

$$K'^{(j)*}(t) = \sum_{i=1}^{n_j} \frac{\hat{\mu}^{(j)*} \overline{G_i^{(j)*}}}{(1 - \hat{\mu}^{(j)*}) e^{-\overline{G_i^{(j)*}} t} + \hat{\mu}^{(j)*}},$$

$$K''^{(j)*}(t) = \sum_{i=1}^{n_j} \frac{(1 - \hat{\mu}^{(j)*}) \hat{\mu}^{(j)*} \overline{G_i^{(j)*}}^2 e^{-\overline{G_i^{(j)*}} t}}{\left[(1 - \hat{\mu}^{(j)*}) e^{-\overline{G_i^{(j)*}} t} + \hat{\mu}^{(j)*}\right]^2},$$

where $\overline{G_i^{(j)*}} = G_i^{(j)*} - \bar{G}^{(j)*}$ is the mean-centred genotypes, and $\hat{\mu}^{(j)*}$= the proportion of cases, is the maximum likelihood estimator of $\mu^{(j)*} = \Pr(Y_i^{(j)} = 1)$ under the null hypothesis H_0^*: $\gamma^{(j)*} = 0$. Based on this CGF, we can approximate the distribution $F^{(j)*}$ and calculate the score $R^{(j)}$ by inverting $F^{(j)*}$ at the signed fastSPA P value, $\pm p^{(j)}$, which is calculated from the Model (1.1) with all covariates. Since the signed P values have one-to-one relationships with the score values, the inversion of $\pm p^{(j)}$ to obtain the score $R^{(j)}$ can be performed using root-finding algorithms such as Newton–Raphson (Press, Flannery, Teukolsky, & Vetterling, 1992), Brent (Brent, 1973), bisection (Press et al., 1992), and so forth. In our implementation, we applied Brent's method for this purpose. The meta-analysis score $R_{\text{meta}} = \sum_{j=1}^{J} R^{(j)}$ will then have the CGF $K_{\text{meta}} = \sum_{j=1}^{J} K^{(j)*}$, and we can apply the SPA test on R_{meta} to obtain the meta-analysis P value.

2.4 | CGF sharing-based method

The aforementioned GC sharing-based method assumes relatively balanced covariates, which have little effect on the discreteness and asymmetry of the null distribution of the score statistics. A more general and mathematically appropriate approach would be to share and utilise the whole CGFs of the within-study score statistics for constructing the reference distribution. Since sharing a complicated function like a CGF using only summary statistics is very difficult, we propose to share the function only at some node-points, and reconstruct the function at the meta-analysis stage using spline approximations. Detailed methodology for this approach is provided in Appendix A in Supporting Information Materials A.

2.5 | Software implementation

We implemented all our proposed methods and the Z-score-based method in our R package SPAtest (available on CRAN). The software can be used to perform fastSPA or Score test and prepare summary level information relevant to the different meta-analysis methods, as well as to perform the final meta-analysis. The software can also perform a hybridised meta-analysis based on the availability of different kinds of summary level information. For example, suppose one study provides only the P value and direction of association, a second study additionally provides the GCs or MAC (if it is a rare variant), and a third study provides the score statistic and spline-based information. Then, a hybrid meta-analysis approach will be to use a normal reference distribution for the P value from the first study, and a reference distribution-based on the genotype-only model for the P value from the second study to calculate the converted scores and their corresponding CGFs. The CGF of the score statistic in the third study can be reconstructed based on spline approximation. Then, the final meta-analysis score will be the sum of those individual scores, and the corresponding CGF will be the sum of those individual CGFs. The meta-analysis P value can then be obtained using the SPA method.

2.6 | Numerical simulations

We evaluated the Type I error rates and empirical powers of the Z-score-based and proposed methods through extensive simulation studies. We considered three different simulation study settings. For the first setting, we meta-analysed seven studies coming from the same population where the genotypes and the nongenetic covariates are simulated independently. For the second setting, we considered a meta-analysis of seven studies where the genotypes and the nongenetic covariates were simulated based on the MAF and principal component (PC) scores in different ethnic groups in the UK Biobank data. In the third setting, we assessed the performance of the methods when a smaller but balanced case–control study is meta-analysed along with a small number of larger but unbalanced biobank-based studies.

2.6.1 | Simulation Study 1

Our first simulation study was designed to represent a meta-analysis of multiple studies from the same population. We considered seven studies with sample sizes $n_j = 2{,}000$ for all $j = 1,...,7$. We further considered three case–control ratios: balanced with the case–control ratio of 1:1 within each study, moderately unbalanced with the case–control ratio of 1:9 within each study, and extremely unbalanced with the case–control ratio of 1:49 within each study. For each choice of case–control ratio, the phenotypes in the j^{th} study were simulated using the following logistic model:

$$\text{logit} \quad [\Pr(Y_i^{(j)} = 1)] = \alpha^{(j)} + 0.5 \times (X_1^{(j)} + X_2^{(j)}) + G_i^{(j)}\gamma^{(j)} \quad \text{for} \quad i = 1,\ 2,\ \dots,\ n_j, \tag{3.1}$$

where $X_1^{(j)} \sim N(0,\ 1)$ and $X_2^{(j)} \sim \text{Bernoulli}(0.5)$ were the nongenetic covariates, and the genotypes ($G_i^{(j)}$s) were generated from a Binomial $(2,\ p)$ distribution where p (same across the seven studies) was the MAF. The intercepts ($\alpha^{(j)}$s) were selected such that the prevalence within each study would become 0.01. The parameters $\gamma^{(j)}$s represent the within-study log odds ratios. For the Type I error comparisons, all $\gamma^{(j)}$s were set to be 0. A wide range of $\gamma^{(j)}$ values were used for the power calculations (see Section 3).

To compare the Type I error rates of different methods under different MAFs, we considered five different MAFs, $p = 0.001,\ 0.005,\ 0.01,\ 0.05,\ 0.1,$ and simulated 5×10^8 variants for each of the MAFs and the three case–control ratios. We recorded the number of rejections at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8} genome-wide significance levels. We further performed a power comparison with 5,000 simulated variants for each of the three case–control ratios and two choices of the MAF, $p = 0.01$ and 0.05, at different values of $\gamma^{(j)}$. As the genome-wide significance threshold for power calculations, we used both a nominal $\alpha = 5 \times 10^{-8}$, and a Type I error adjusted empirical α where the corresponding method has Type I error 5×10^{-8}. The empirical α level was calculated based on 5×10^8 simulated data sets from the simulation setting described above (seven studies, each with 2,000 samples) where the MAFs were sampled from the MAF spectrum (Figure S2 in Supporting Information Material A) of the white British ancestry group (~117 k samples) in the UK Biobank interim release data.

2.6.2 | Simulation Study 2

Our second simulation study was designed to represent a trans-ethnic meta-analysis, where contrary to the first simulation study setting, we not only allow the MAFs to be different across the studies, but also simulate the genotypes in a way such that they are correlated with the covariates to adjust for. We considered seven studies with sample sizes $n_j = 2{,}000$ for all $j = 1,...,6$, and $n_7 = 1{,}500$. To simulate the genotypes and the nongenetic covariates from a realistic meta-analysis of GWAS, we used genotype data from the UK Biobank interim release data (UK Biobank, 2015). The first five studies included first four PC scores as covariates and genotypes simulated from the MAF spectrum of the white ancestry group in the UK Biobank samples. To maintain the correlated nature of the genotypes and the PC scores, genotypes were simulated using PC scores. We further added a binary covariate generated from a Bernoulli (0.5) distribution independent of the PC scores and the genotypes. Covariates and genotypes were simulated in a similar way for Studies 6 and 7 based on the south Asian and black ancestry groups, respectively. The model to simulate the phenotypes was similar to the one used in the first simulation study, except for different nongenetic covariates. Detailed explanation of the simulation procedure is provided in Appendix B in Supporting Information Material A.

In Transethnic studies, variants have different MAFs across different ancestry groups. To calculate the Type I error rates for diverse scenarios of MAFs, we first considered three MAF bins for the alleles of the simulated variants: rare variants with $\text{MAF} < 0.01$, low frequency variants with $0.01 < \text{MAF} < 0.05$ and common variants with $\text{MAF} > 0.05$. We then categorised the simulated variants in the following four categories based on their allele frequencies (AF): (a) all rare, when the variant has the same minor rare allele in all seven studies, (b) all low frequency, when the variant has the same low frequency allele in all seven studies, (c) all common, when the variant has the same common allele in all seven studies, and (d) different AF, when the variant falls in different MAF bins in at least two different studies. The different AF category also includes variants which have different alleles as the minor alleles in different studies. For each variant category and case–control ratio, we simulated 5×10^8 data sets under the null hypothesis and recorded the number of rejections at the genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}.

2.6.3 | Simulation Study 3

We investigated the performance of different meta-analysis strategies when a balanced case–control study, which is smaller in sample size, is meta-analysed along with two larger biobank-based unbalanced studies. This simulation study represents the real-world meta-analyses where the researchers collect balanced case–control data on rare traits/diseases, and attempt to meta-analyse them with association results from a small number of larger cohort-based studies. To simulate the genotypes, nongenetic covariates and the phenotypes, we used the same simulation and logistic regression models as in our first simulation study setting. The sample size for the balanced case–control study was 2,000 with 1,000 cases and 1,000 controls, and the unbalanced studies had sample size 10,000 each. We considered two case–control ratios for these unbalanced studies: moderately unbalanced with case:control = 1:9 within each study, and extremely unbalanced with case: control = 1:49 within each study. For each of the case–control ratio, we compared the Type I error rates of different methods under five different MAFs, $p = 0.001,\ 0.005,\ 0.01,\ 0.05,$ and 0.1 based on 5×10^8 simulated variants each.

For the first two simulation settings and the unbalanced studies in the third simulation setting, the within-study *P* values were calculated using the traditional score test (score), fastSPA test with 2 *SD*s threshold (fastSPA-2), and fastSPA test with 0.1 *SD*s threshold (fastSPA-0.1). Since score test is relatively well-calibrated for balanced case–control studies (Dey et al., 2017), only score *P* values were calculated for the balanced study in the third simulation setting. We then considered the following meta-analysis methods to compare their Type I error rates and empirical powers: *Z*-score-based meta-analysis (*Z* score), GC sharing-based meta-analysis (GC), and CGF sharing-based meta-analysis (CGF-Spline). Score *P* values were meta-analysed using the *Z*-score method, fastSPA-2 and -0.1 *P* values were meta-analysed using both the *Z*-score and GC methods, and the within-study observed score statistics were meta-analysed using the CGF-Spline method. For the balanced case–control study in the third simulation setting, the *Z* scores obtained from the score *P* values were used in the GC method, and the corresponding normal distribution-based CGFs were used in the CGF-Spline method. We also compared the Type I error rates and the empirical powers of a joint analysis (Joint) using the fastSPA-2 test on the pooled data as the gold standard. We further provided a computation time comparison of our proposed methods in Appendix C in Supporting Information Material A.

2.7 | UK biobank data analysis

We demonstrated the performance of our proposed methods by analysing two phenotypes based on the UK Biobank interim release data (UK Biobank, 2015). The UK Biobank (Bycroft et al., 2017) contains detailed phenotypic information based on EHRs for ~500 k individuals in the United Kingdom. In the interim release (May 2015), information on ~150 k individuals were released to the public. Details about the data and preprocessing are provided in Appendix D in Supporting Information Materials A. A histogram of the case–control ratios (Figure S1 in Supporting Information Materials A) of different binary phenotypes shows that the ratios are heavily skewed towards zero, which means the binary phenotypes are mostly unbalanced.

To compare our proposed methods with the *Z*-score-based meta-analysis method, we analysed two phenotypes, ulcerative colitis (PheWAS code: 555.2, case:control ≈ 1:100), and psoriasis (PheWAS code: 696.4, case:control ≈ 1:165) based on 117,494 unrelated samples from the White British ancestry group of the interim release data. The samples were then divided into 22 groups based on the assessment centre where they first consented to be included in the biobank. We selected 19 centres (Table S1 in Supporting Information Material A) with at least five cases for each of the two phenotypes, and treated these centres as our individual studies to perform association analyses of the phenotypes on the autosomal variants within each of them. For the within-study association analyses, we applied fastSPA-2, fastSPA-0.1 and Score tests, adjusting for age, sex, genotyping array, and first four principal components. Individuals which had phenotype or at least one covariate information missing, were removed from the analysis of that corresponding phenotype. We only applied the within-study tests for variants with within-study MAC at least three. Because the GC-based meta-analysis requires the overall GCs, we applied our within-study tests on the best called genotypes instead of dosages in the imputed data. We then meta-analysed the results using the *Z*-score-based meta-analysis (*Z* score), GC sharing-based meta-analysis (GC), and CGF sharing-based meta-analysis (CGF-Spline). The meta-analysis methods were only applied for variants that were tested in at least two different studies, and the overall MACs were at least ten. For each phenotype, ~29 million variants were meta-analysed. We further performed a joint analysis (Joint) with the pooled samples using the fastSPA-2 test, adjusting for the assessment centre. Due to the computational burden of performing a pooled joint analysis, we only performed it for the variants with GC–fastSPA-2 *P* values smaller than 5×10^{-3}. Otherwise, we recorded the *P* values from GC–fastSPA-2 method as the joint analysis *P* values.

3 | RESULTS

In this section, we evaluate the performance of the proposed methods against the *Z*-score-based meta-analysis based on the numerical simulations and the UK Biobank data application described above.

3.1 | Numerical simulations

3.1.1 | Type I error comparison

The Type I error comparison based on simulation Study 1 (Figure 1) clearly shows that the proposed CGF-Spline and GC methods provided well-controlled Type I error rates across all the MAFs and all the case–control ratios. Expectedly, the joint analysis also controlled the Type I error rates. In contrast, the *Z*-score method resulted in inflated Type I error rates in moderately unbalanced and extremely unbalanced settings, especially for the rarer minor AF. Interestingly, the *Z*-score method with fastSPA-0.1 performed worse than that with fastSPA-2, although fastSPA-0.1 used the SPA to more variants. This further verifies our assertion that using normal distributions to invert the study-specific *P* values which are possibly discrete, asymmetric, and originally calculated using the SPA, can result in failure to control Type I error in the meta-analysis process. In contrast, the GC method shows similar performance using fastSPA-0.1 and fastSPA-2 *P* values, which shows its robustness in meta-analysing *P* values regardless of whether they were originally calculated using the normal approximation or the SPA. For MAF = 0.001 under the extremely unbalanced setting, there is conservative behaviour shown by the *Z*-score method when using fastSPA-0.1 or fastSPA-2 *P* values at $\alpha = 5 \times 10^{-5}$ level. All methods provided well-controlled Type I error rates for the balanced case–control ratio. We further simulated 5×10^{8} data sets under the settings of simulation Study 1 with a much more extreme case–control ratio (1:99), and even under such extreme case–control imbalance, our proposed methods showed well-controlled Type I errors, whereas the *Z*-score method overall resulted in Type I error inflation (Figure S3 in Supporting Information Material A).

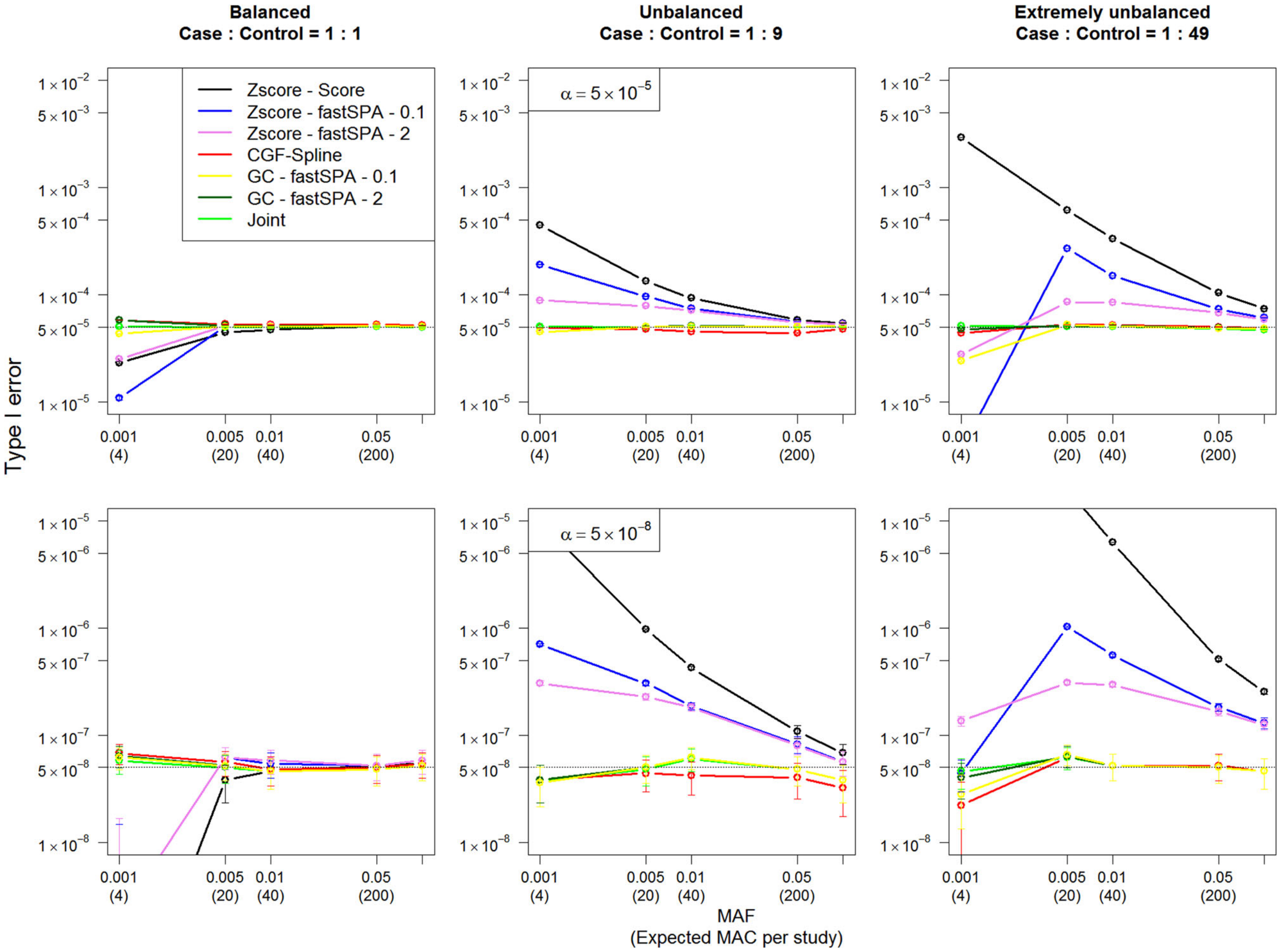

FIGURE 1 Type I error comparison between the *Z*-score-based meta-analysis and our proposed CGF-Spline and GC methods where the phenotypes, nongenetic covariates, and the genotypes are simulated as described in simulation Study 1. Joint represents the joint analysis with the pooled data. The top and the bottom panels show empirical Type I error rates at genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. From left to right, the plots consider the within-study case–control ratios 1:1, 1:9, and 1:49, respectively. In each plot, the *X*-axis represents MAFs with expected MACs per study in parenthesis, and the *Y*-axis (in logarithmic scale) represents the empirical Type I error rates. 95% confidence intervals at different MAFs are also presented. CGF: cumulant generating function; GC: genotype count; MAC: minor allele count; MAF: minor allele frequency

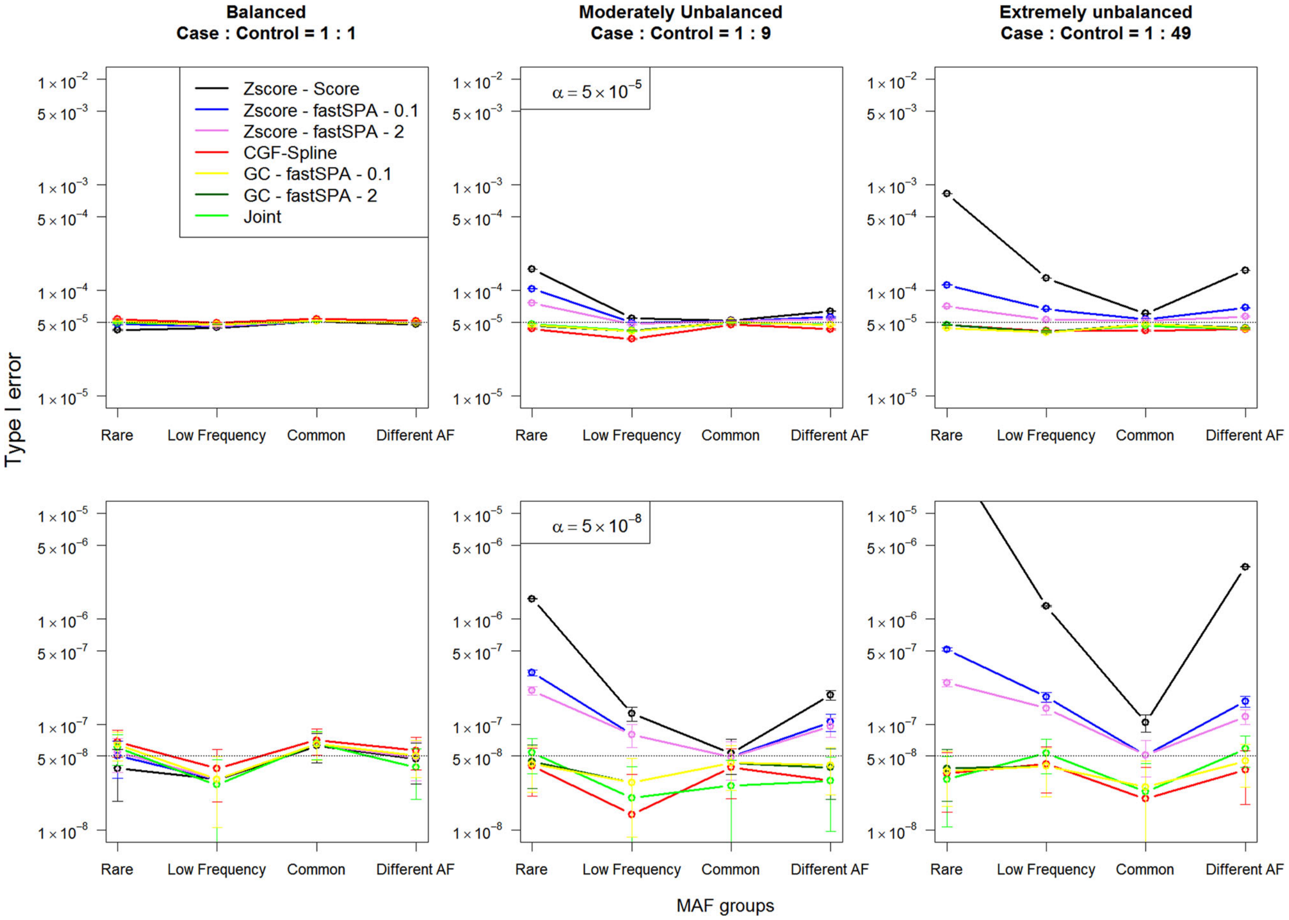

FIGURE 2 Type I error comparison between the *Z*-score-based meta-analysis and our proposed CGF-Spline and GC methods where the phenotypes, nongenetic covariates, and the genotypes are simulated as described in simulation Study 2. Joint represents the joint analysis with the pooled data. The top and the bottom panels show empirical Type I error rates at genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. From left to right, the plots consider the within-study case–control ratios 1:1, 1:9, and 1:49, respectively. In each plot, the *X*-axis represents different MAF groups: rare (variant is rare in all studies), low frequency (variant is low frequency in all studies), common (variant is common in all studies), and different AF (variant is in different allele frequency group in at least two different studies). The *Y*-axis (in logarithmic scale) represents the empirical Type I error rates. 95% confidence intervals at different MAFs are also presented. AF: allele frequencies; CGF: cumulant generating function; GC: genotype count; MAF: minor allele frequency

Similar observation follows for simulation Study 2. The Type I error comparison (Figure 2) suggests that our proposed methods showed no sign of Type I error inflation across different MAFs and case–control ratios, whereas the *Z*-score method resulted in inflated Type I error rates for the moderately unbalanced and extremely unbalanced settings, especially for the all rare, all low frequency and different MAF categories. *Z*-score method using Score *P* values had the maximum inflation across all categories.

In simulation Study 3, we also have similar results (Figure 3) for our proposed methods. However, the *Z*-score method using the fastSPA-0.1 or fastSPA-2 *P* values showed no sign of significant Type I error inflation in the extremely unbalanced case–control setting, and only slight inflation in the moderately unbalanced setting. This suggests that the *Z*-score-based method can be adequate for controlling the Type I error rates when only a small number of biobank-based studies are included in the meta-analysis. However, as seen from the other two simulation studies, the *Z*-score method may fail to control Type I error rates when large number of unbalanced studies are involved.

3.1.2 | Power comparison

Next, we compare the empirical powers of different meta-analysis strategies along-with the joint analysis as the gold standard under the first simulation setting. Because the *Z*-score-based meta-analysis method provided inflated Type I error rates as seen in the Type I error comparisons, we used empirical α levels calculated from Type I error simulations for each method where the empirical Type I error rate becomes 5×10^{-8}. The power curves (Figure 4) show that the *Z*-score method has slightly lower power (lowest when using score test *P* values) in the moderately

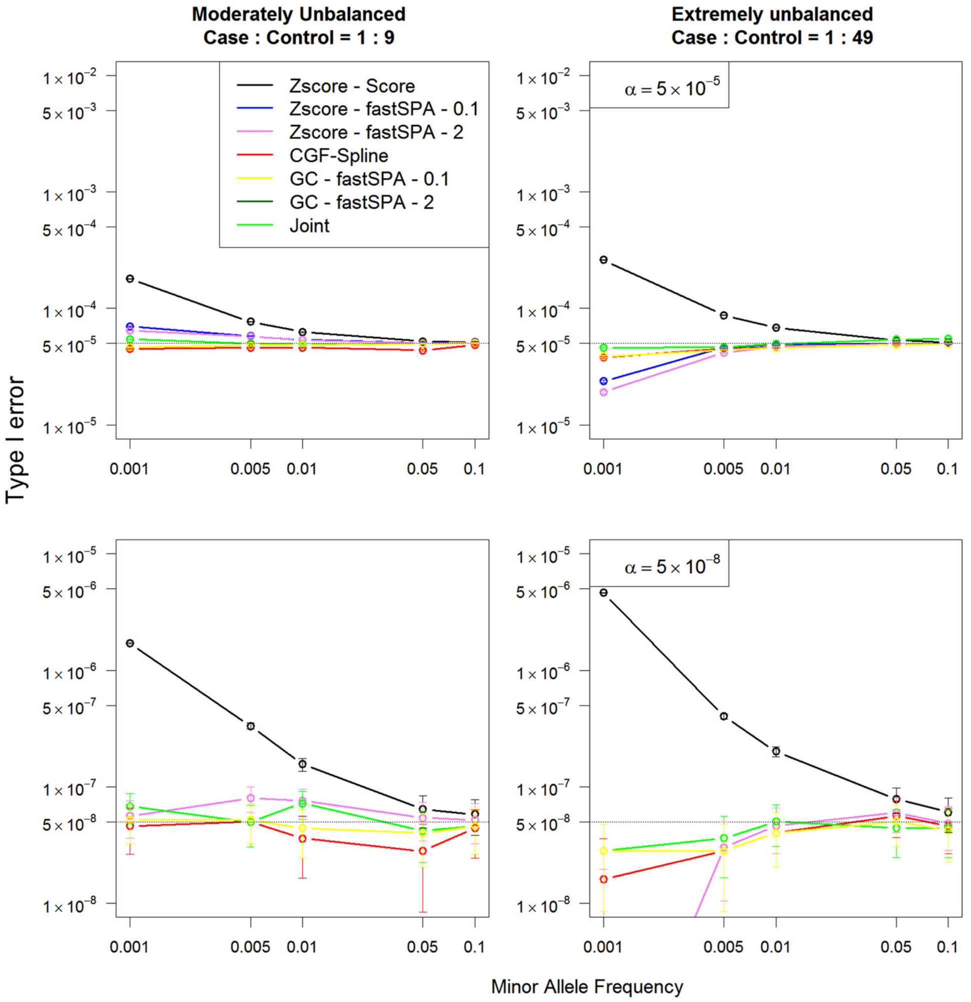

FIGURE 3 Type I error comparison between the *Z*-score-based meta-analysis and our proposed CGF-Spline and GC methods where the phenotypes, nongenetic covariates, and the genotypes are simulated as described in simulation Study 3. Joint represents the joint analysis with the pooled data. The top and the bottom panels show empirical Type I error rates at genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. The left and right panels consider the within-study case–control ratios 1:9 and 1:49, respectively for the unbalanced studies. In each plot, the *X*-axis represents MAFs with expected MACs in parenthesis, and the *Y*-axis (in logarithmic scale) represents the empirical Type I error rates. 95% confidence intervals at different MAFs are also presented. The empirical Type I error rates were almost identical between *Z* score–fastSPA-2 and *Z* score–fastSPA-0.1, and between GC–fastSPA-2 and GC–fastSPA-0.1, and hence the lines are sometimes overlapped in this plot. CGF: cumulant generating function; GC: genotype count; MAF: minor allele frequency; SPA: saddlepoint approximation

and extremely unbalanced case–control ratios. Our proposed methods provide very similar power to the joint analysis, and all methods provide similar power in the balanced case–control setting. When nominal $\alpha = 5 \times 10^{-8}$ level was used (Figure S4 in Supporting Information Material A), the *Z*-score method expectedly showed higher powers in the unbalanced settings as it is not calibrated for its Type I errors.

3.2 | UK biobank data analysis

We meta-analysed the results from 19 individual studies (assessment centres) for the phenotypes ulcerative colitis and psoriasis, using the *Z*-score-based meta-analysis (*Z* score), GC sharing-based meta-analysis (GC), and CGF sharing-based meta-analysis (CGF-Spline). The quantile–quantile (QQ) plots presented in Figures 5 and 6 show that the meta-analysis *P* values from our proposed methods closely follow the uniform distribution, whereas those from the *Z*-score method are either much smaller (*Z*-score method using Score or fastSPA-0.1 *P* values) or larger (*Z*-score method using fastSPA-2 *P* values) than expected for rare variants (MAF < 0.01). This suggests conservative behaviour of the *Z*-score method when using the fastSPA-2 *P* values, and extremely anticonservative behaviour when using fastSPA-0.1 or Score *P* values. In

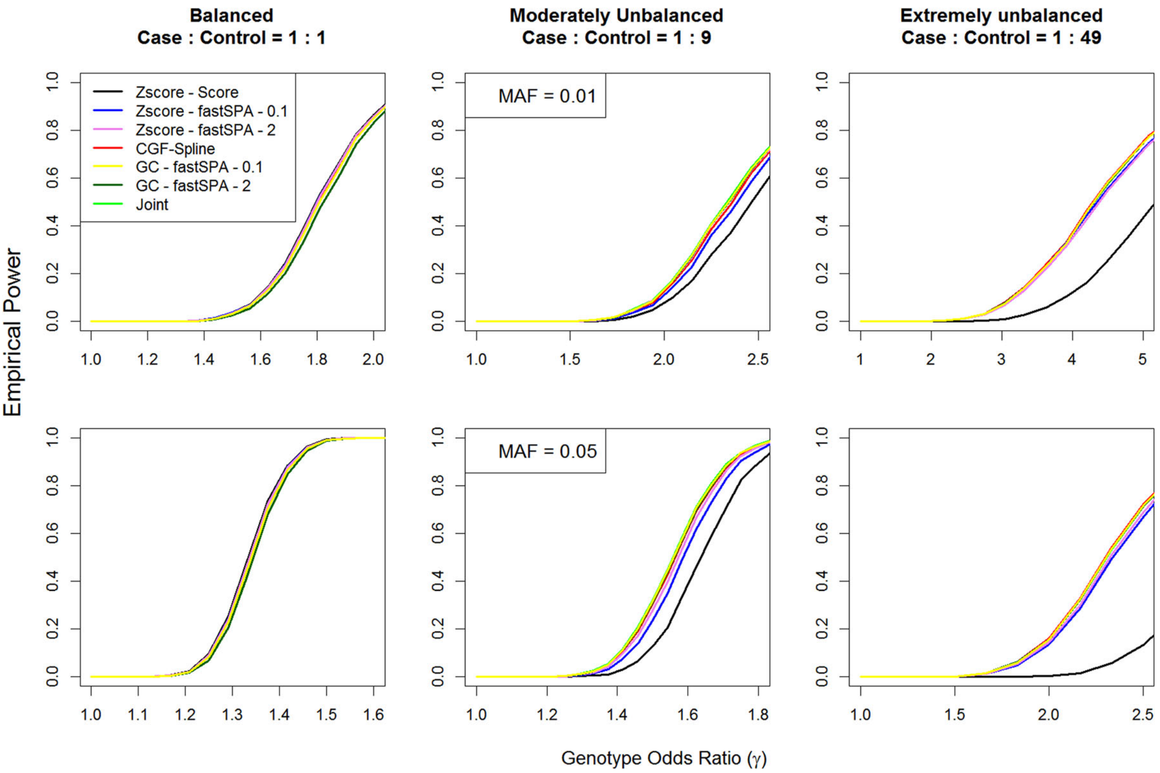

FIGURE 4 Power curves for the *Z*-score, CGF-Spline, and GC methods. Top panel considers MAF = 0.01 and bottom panel considers MAF = 0.05. From left to right, the plots consider case–control ratios 1:1, 1:9, and 1:49, respectively. In each plot the *X*-axis represents genotype odds ratios and the *Y*-axis represents the empirical power. Empirical power was estimated from 5,000 simulated data sets at their Type I error adjusted empirical α levels where their empirical Type I errors are equal to 5×10^{-8}. CGF: cumulant generating function; GC: genotype count; MAF: minor allele frequency; SPA: saddlepoint approximation

contrast, both the GC and CGF-Spline methods improve the accuracy of the meta-analysis *P* values and provide well-calibrated QQ plots. Further, the QQ plots from our proposed methods show similar behaviour to the QQ plots from the Joint analysis (Figure S5 in Supporting Information Material A). We also presented the genomic control inflation factors (λ) of different meta-analysis strategies in Table S2 in Supporting Information Material A. For ulcerative colitis, all our proposed methods showed no inflation or deflation in the genomic controls at *P* value quantiles $q = 0.01$ and 0.001, whereas the *Z*-score method showed severely inflated inflation factors when using the score (e.g., $\lambda = 1.34$ at $q = 0.01$) and fastSPA-0.1 (e.g., $\lambda = 3.16$ at $q = 0.01$) *P* values and deflated inflation factors when using the fastSPA-2 (e.g., $\lambda = 0.82$ at $q = 0.01$) *P* values at those *P* value quantiles. This result further supports the observations made from the QQ plots. When considering the inflation factors at the median *P* value quantile ($q = 0.5$), the CGF-Spline ($\lambda = 0.74$), and GC method using fastSPA-2 *P* values ($\lambda = 0.84$) showed deflated inflation factors, and GC method using fastSPA-0.1 *P* values ($\lambda = 1.40$) showed inflated inflation factor. This is expected, since the SPA test *P* values near the median are not calculated using the SPA as discussed in Dey et al. (2017). In that paper, they also found inflated genomic control factors for fastSPA-0.1 and deflated genomic control factors for fastSPA-2 *P* values at the median level for extremely unbalanced case–control ratios. The inflation factors showed similar patterns for psoriasis. However, at *P* value quantile $q = 0.001$, the GC method using fastSPA-2 *P* values, and the CGF-Spline method showed slightly larger than expected inflation factors ($\lambda = 1.10$ for both methods). This might be due to the presence of the major histocompatibility complex (MHC) in the 6p21 region which contains a large number of polymorphic variants and it is a known associated region for psoriasis (Stuart et al., 2015). After excluding the MHC region from the inflation factor calculation, the inflation factors became very close to unity.

The top genome-wide significant single-nucleotide polymorphisms (SNPs) in different regions, identified by the CGF-Spline method, are listed in Table S3 in Supporting Information Material A. The top significant SNPs were identical for the GC method. The *P* values for the top significant SNPs were similar for all the

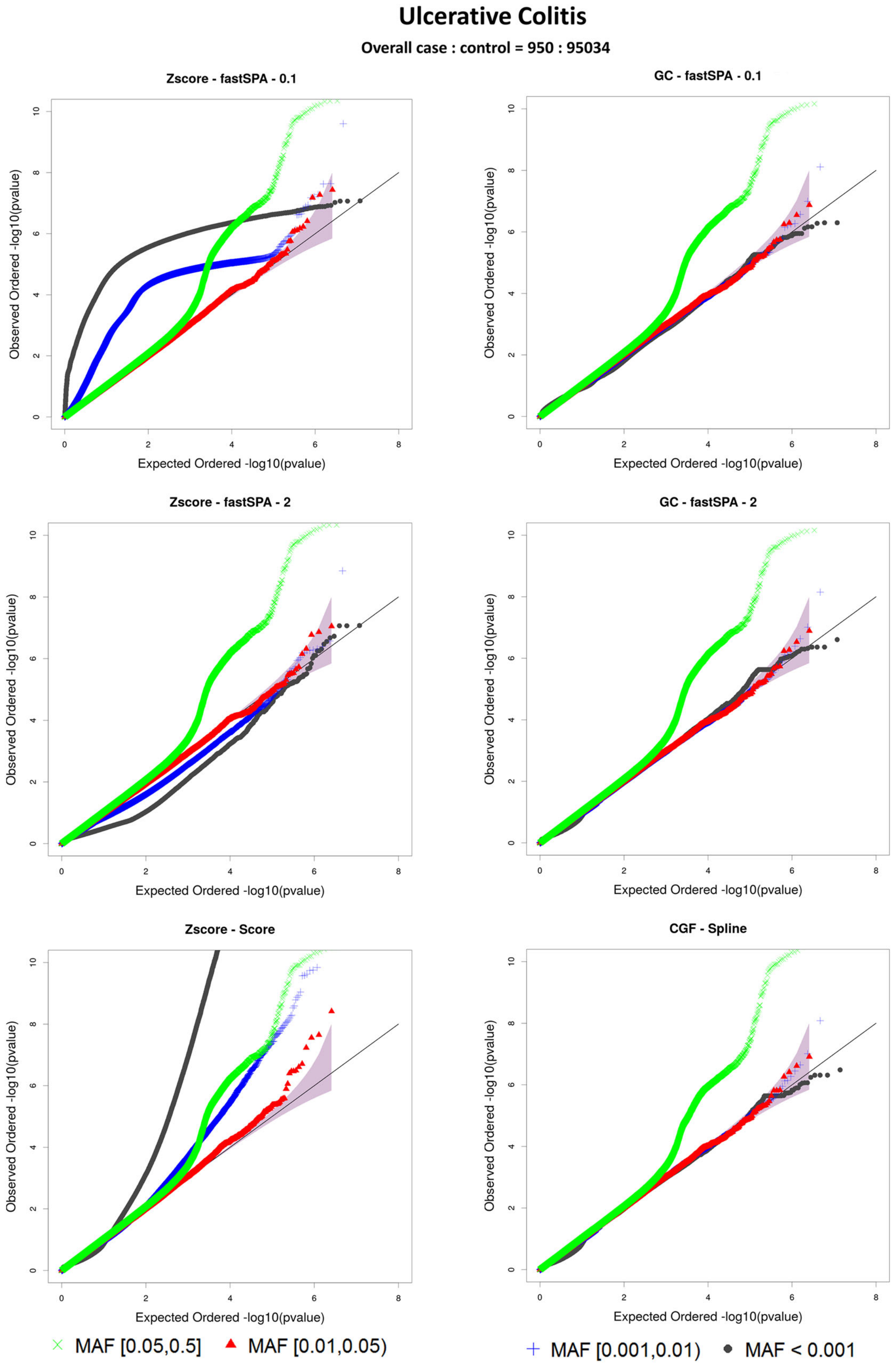

FIGURE 5 QQ plots for ulcerative colitis based on the UK Biobank interim release data. QQ plots using the *Z*-score method are provided in the left panel, and the QQ plots using our proposed methods are provided on the right panel. The plots are colour-coded based on different MAF categories. MAF: minor allele frequency; QQ: quantile–quantile; SPA: saddlepoint approximation

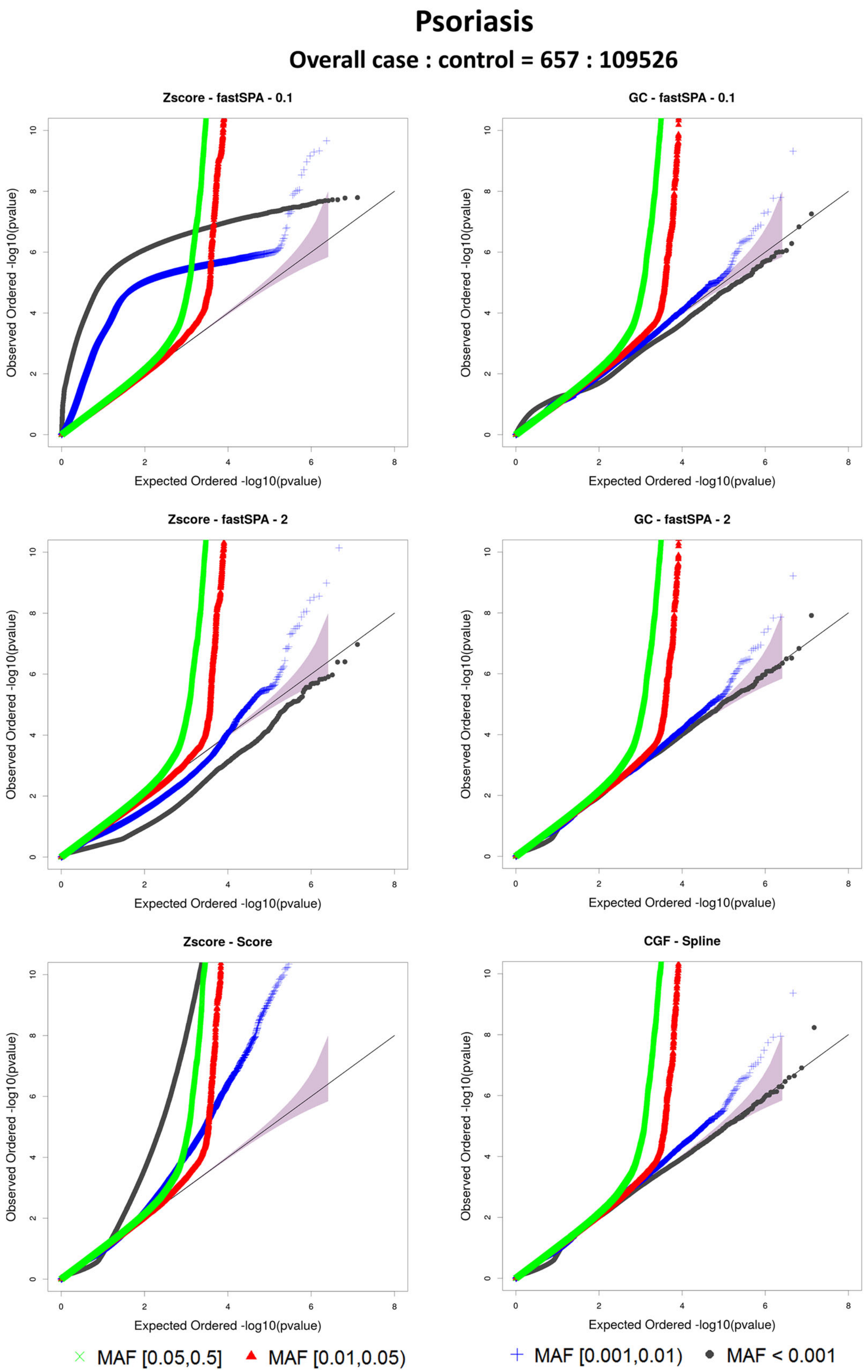

FIGURE 6 QQ plots for psoriasis based on the UK Biobank interim release data. QQ plots using the *Z*-score method are provided in the left panel, and the QQ plots using our proposed methods are provided on the right panel. The plots are colour-coded based on different MAF categories. MAF: minor allele frequency; QQ: quantile–quantile; SPA: saddlepoint approximation

methods, except Z-score-based meta-analysis using Score and fastSPA-0.1 P values. Z-score method using score P values resulted in much smaller meta-analysed P values for all of those SNPs, and Z-score method using fastSPA-0.1 P values resulted in surprisingly large P values for testing psoriasis on the two SNPs on chromosome 22 (rs549956609 and rs560106765). All other meta-analysis procedures and the joint analysis on these two SNPs resulted in P values which were close to the genome-wide significance level (GC–fastSPA-2, CGF-Spline, and Joint analysis P values smaller than, and GC–fastSPA-0.1 and Z-score–fastSPA-2 P values larger than $\alpha = 5 \times 10^{-8}$ level.).

3.2.1 | Applicability on imputed dosages

To assess the performance of our methods with genotype dosage data, we further performed our within-study tests to calculate the P values, scores and spline-based summary statistics using the dosage data, and then meta-analysed the results using our proposed methods. For the GC method, we calculated the within-study P values based on the dosages, but constructed the genotype-only model using GCs calculated using three methods: counting the best-called genotypes (BCG), rounding off the dosages to the nearest integers and counting them (rounded dosages), and GCs obtained from the MACs assuming Hardy–Weinberg equilibrium (HWE). We also compared the results with a joint analysis performed in the same way described for the genotype data analysis. The resulting QQ plots (Figure S6–S8 in Supporting Information Material A) showed no sign of inflation or deflation for the GC methods, and showed very similar behaviour to the QQ plots from the CGF-Spline method and the Joint analysis (Figure S9 in Supporting Information Material A). which suggests that the methods are robust for the analysis of dosage data.

We further generated the QQ plots for four different ranges (<0.3, 0.3–0.6, 0.6–0.9, and ≥0.9) of imputation quality Impute-INFO scores (Howie, Donnelly, & Marchini, 2009; Supporting Information Material B). Overall, our proposed methods provided close to uniform QQ plots. For variants with smaller INFO scores (INFO < 0.6), the GC method using fastSPA-0.1 P values showed small amount of inflation when the BCG or rounded dosage values (rounded dosage) were used. However, GC method using only MAC information provided the most calibrated (close to the uniform distribution) QQ plots. This is expected, because lower imputation quality is more often observed for rare variants, for which MAC information is enough to calculate the GCs, as we do not usually observe homozygous minor genotypes.

4 | DISCUSSION

In this paper, we evaluated the performance of the traditional Z-score-based meta-analysis strategy to combine association results from multiple unbalanced GWASs, and proposed two alternative strategies that can provide well-calibrated meta-analysis P values, even when the case–control ratios are extremely unbalanced and the MACs are small. Through extensive numerical studies and an application on the UK Biobank data, we showed that the Z-score-based method can result in conservative or anticonservative behaviour in the meta-analysis P values, whereas our proposed methods provided well-controlled Type I error rates. The proposed methods also showed similar empirical powers as a joint analysis on the pooled data.

When the effect sizes are not available, such as in the case of the SPA-based test, it is widely popular to use the Z-score-based meta-analysis approach and combine the individual P values into a meta-analysis P value. In our third simulation setting, we showed that the Z-score-based approach can still be appropriate when only a small number of biobank-based studies with unbalance phenotypes are included in the meta-analysis. However, we will suggest the researchers to be cautious when using the Z-score-based approach, as including more such unbalanced studies can result in a loss of calibration in the meta-analysis P values. When effect size estimates are available, for example, when using the Firth's bias-corrected likelihood ratio test (Firth, 1993), the inverse variance-weighted method is another popular meta-analysis approach used by the researchers. However, Ma et al. (2013) showed that the inverse variance-weighted meta-analysis method using the Firth's bias-corrected effect size estimates also results in Type I error inflation when meta-analysing several unbalanced studies.

In this paper, we assumed that the individual studies do not have genetically related samples. In presence of related samples, the SAIGE test (Zhou et al., 2018) can properly account for the sample relatedness and provide accurate P values in single studies with unbalanced case–control ratios. As the SAIGE P values are calculated using the SPA method based on the score statistic and its CGF, the spline-based meta-analysis method can still be applicable for combining multiple studies that are analysed using SAIGE. However, the GC-based method may not be appropriate in such scenarios as the genotype-only model does not contain any information about the sample relatedness. The applicability of our methods in studies containing genetically related samples, is left for future research.

Comparing the two proposed methods, the spline-based method (CGF-Spline) does not require any assumption on the effect of the nongenetic covariates since it reconstructs spline approximations of the null distributions of the score statistics and uses them to calculate the meta-analysis *P* values. Thus, it is more suitable to be applied regardless of of the covariate effects. On the other hand, the GC-based method (GC) assumes relatively balanced nongenetic covariates with low covariate effects. However, the numerical simulations with very strong covariate effects (Figure S10 in Supporting Information Material A) also showed no sign of Type I error inflation or deflation for this method. Another difference between the proposed methods is in their applicability on imputed dosage data. As the GC method requires the overall GCs to construct the genotype-only model, it is more suitable to be applied when the within-study analyses are performed on the BCG instead of dosages. The CGF-spline method is robust in this aspect as it can utilise the CGFs of the test statistics regardless of whether they were calculated from genotype or dosage data. However, in our UK Biobank data analysis example, both our proposed methods showed no sign of inflation or deflation of Type I errors, even when the within-study tests were performed on dosage data. Therefore, for practical application purposes, the GC-based method can be used to obtain accurate meta-analysis *P* values. One advantage of the GC-based method is that it is software-independent, and requires information which are more readily available compared to the spline-based method.

ACKNOWLEDGEMENTS

This study has been conducted using the UK Biobank Resource under application number 24460. We would like to thank the investigators of UK Biobank for providing us with access to the genotype and phenotype data. S. L. and R. D. were supported by NIH (R01 HG008773).

ORCID

Rounak Dey http://orcid.org/0000-0002-6540-8280
Wei Zhou http://orcid.org/0000-0001-7719-0859
Seunggeun Lee http://orcid.org/0000-0002-8097-3878

REFERENCES

Barndorff-Nielsen, O. E. (1990). Approximate interval probabilities. *Journal of the Royal Statistical Society*, *52*(3), 485–496.

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … Marchini, J. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*. 166298. https://www.nature.com/articles/s41586-018-0579-z

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, *25*(4), 631–650. https://doi.org/10.1214/aoms/1177728652

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *American Journal of Human Genetics*, *101*(1), 37–49. https://doi.org/10.1016/j.ajhg.2017.05.014

Evangelou, E., & Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, *14*(6), 379–389. https://doi.org/10.1038/nrg3472

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Han, B., & Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American Journal of Human Genetics*, *88*(5), 586–598. https://doi.org/10.1016/j.ajhg.2011.04.014

Hebbring, S. J. (2014). The challenges, advantages and future of phenome-wide association studies. *Immunology*, *141*(2), 157–165. https://doi.org/10.1111/imm.12195

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T., Midthjell, K., Stene, T., … Holmen, J. (2013). Cohort profile: The HUNT study, Norway. *International Journal of Epidemiology*, *42*(4), 968–977. https://doi.org/10.1093/ije/dys095

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., & investigators, G. D. (2013). Recommended joint and meta-analysis strategies for case–control association testing of single low-count variants. *Genetic Epidemiology*, *37*(6), 539–550. https://doi.org/10.1002/gepi.21742

Press, W. H., Flannery., B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). Cambridge, England: Cambridge University Press.

Stuart, P. E., Nair, R. P., Tsoi, L. C., Tejasvi, T., Das, S., Kang, H. M., … Elder, J. T. (2015). Genome-wide association analysis of psoriatic arthritis and cutaneous psoriasis reveals differences in their genetic architecture. *American Journal of Human Genetics*, *97*(6), 816–836. https://doi.org/10.1016/j.ajhg.2015.10.019

UK Biobank. (2015). Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource. Retrieved from http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Elvestad, M. B., Wolford, B. N., … Lee, S. (2018). Efficiently controlling for case–control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*, 1335–1341. https://doi.org/10.1101/212357

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Dey R, Nielsen JB, Fritsche LG, et al. Robust meta-analysis of biobank-based genome-wide association studies with unbalanced binary phenotypes. *Genet. Epidemiol.* 2019;43:462–476. https://doi.org/10.1002/gepi.22197