Technical Note: Efficient and accurate estimation of genotype odds ratios in biobank-based unbalanced case-control studies

Rounak Dey*& Seunggeun Lee†

Abstract

In genome-wide association studies (GWASs), genotype log-odds ratios (LORs) quantify the effects of the variants on the binary phenotypes, and calculating the genotype LORs for all of the markers is required for several downstream analyses. Calculating genotype LORs at a genome-wide scale is computationally challenging, especially when analyzing large-scale biobank data, which involves performing thousands of GWASs phenome-wide. Since most of the binary phenotypes in biobank-based studies have unbalanced ($case : control = 1 : 10$) or often extremely unbalanced ($case : control = 1 : 100$) case-control ratios, the existing methods cannot provide a scalable and accurate way to estimate the genotype LORs. The traditional logistic regression provides biased LOR estimates in such situations. Although the Firth bias correction method can provide unbiased LOR estimates, it is not scalable for genome-wide or phenome-wide scale association analyses typically used in biobank-based studies, especially when the number of non-genetic covariates is large. On the other hand, the saddlepoint approximation-based test (fastSPA), which can provide accurate p values and is scalable to analyse large-scale biobank data, does not provide the genotype LOR estimates as it is a score-based test. Here, we propose a scalable method based on score statistics, to accurately estimate the genotype LORs, adjusting for non-genetic covariates. Comparing to the Firth method, our proposed method reduces the computational complexity from $O(nK^2 + K^3)$ to $O(n)$, where n is the sample-size, and K is the number of non-genetic covariates. Our method is $\sim$ 10x faster than the Firth method when 15 covariates are being adjusted for. Through extensive numerical simulations, we show that the proposed method is both scalable and accurate in estimating the genotype ORs in genome-wide or phenome-wide scale.

1 Introduction

Recent developments in genotyping and imputation technologies [Marchini and Howie, 2010, Das et al., 2016], and the availability of electronic health records (EHR)-based phenotypic information in different biobanks [Bycroft et al., 2017, Krokstad et al., 2013, Fritsche et al.,

*Department of Biostatistics, Harvard T.H. Chan School of Public Health
†Department of Biostatistics, University of Michigan

2018], are allowing the researchers to perform Phenome-wide scale genome-wide association studies (GWASs) to investigate the pleiotropic effect [Pendergrass et al., 2013, Hebbring, 2014, Verma et al., 2018] of different variants on multiple phenotypes. Since, subjects in biobanks are usually recruited in a cohort-based design, most binary phenotypes based on biobank data are unbalanced ($case : control = 1 : 10$), or often extremely unbalanced ($case : control = 1 : 100$).

In case-control GWASs, genotype odds ratios (ORs) quantify the effect of the variants on the phenotypes. Genotype ORs are required for several downstream analyses. For example, in Mendelian randomization studies [Burgess et al., 2013, Bowden et al., 2015, Burgess et al., 2015], they are used to control for unknown confounders to establish causal relationships between different risk factors and biomedical outcomes. Genotype ORs are also used in the inverse variance-weighted meta-analysis method to combine results from multiple studies [Willer et al., 2010, Evangelou and Ioannidis, 2013, Tsoi et al., 2017]. Usually, it is estimated in the logarithmic scale (hence called log-odds ratios or LORs) by maximizing the likelihood of appropriate logistic regression models.

Maximizing the logistic regression likelihood involves updating the estimates iteratively, which requires $O(nK^2 + K^3)$ computation per iteration, where n is the sample-size, and K is the number of covariates to adjust for in the model. Since, the samples in modern large-scale biobank-based GWASs are usually heterogeneous in nature (ancestry, gender, age etc.), the number of covariates, K can become substantially large in these studies, and the overall computation time can increase at a fast pace. Moreover, when the case-control ratios are unbalanced and the minor allele counts (MACs) are low, the logistic regression likelihood, and thus the estimated genotype LORs, can be biased [Firth, 1993]. Even though, the maximum penalized likelihood estimation [Ma et al., 2013, Firth, 1993] can provide bias-adjusted LORs, the method is also not scalable to handle biobank-scale datasets. Recently, several score test-based methods were proposed [Dey et al., 2017, Zhou et al., 2017, Dey et al., 2019] to efficiently and accurately test for genotype-phenotype associations for such unbalanced phenotypes. However, as the score tests only fit the model under the null hypothesis of no association, they do not provide any estimate of the genotype LORs.

Here, we propose a fast computation method for genotype LORs in case-control studies using the score statistics of the fastSPA test, which improves the computation time from $O(nK^2 + K^3)$ to $O(n)$. Through applications on simulated data, we demonstrate the performance of our proposed method under different case-control ratios, and for variants with different rarity of alleles.

2 Methods

We consider a case-control study with n samples, where the phenotype $Y_i = 1, 0$ denotes the case-control status of the i-th subject. Let $G_i = 0, 1, 2$ denote the MAC for the variant of interest, and X_i denote other covariates or observed confounders, for the i-th subject. We denote the outcome vector by $Y = (Y_1, \ldots, Y_n)$, the genotype vector by $G = (G_1, \ldots, G_n)$, and the covariate matrix by $X = (X_1 \ldots X_n)^\top$. To model the phenotypes on the genotypes and the covariates, we use the following logistic regression model,

$$logit\,(\mu_i) = X_i^\top \beta + G_i \gamma, \tag{1}$$

where $\mu_i = P(Y_i = 1|X_i, G_i)$, and β, γ are the regression parameters. For simplicity of notations, we augmented the intercept term in the covariate matrix X. Let $\mu = (\mu_1, \ldots, \mu_n)$, and $W = diag(\mu_i(1-\mu_i))$. Then, the score function and information matrix under model 1 are given by,

$$S(\beta, \gamma) = \begin{pmatrix} X^\top \\ G^\top \end{pmatrix} (Y - \mu); \quad I(\beta, \gamma) = \begin{bmatrix} X^\top W X & X^\top W G \\ G^\top W X & G^\top W G \end{bmatrix}$$

respectively. The standard Newton-Raphson algorithm to obtain the maximum likelihood estimates (mle) $\left(\hat{\beta}, \hat{\gamma}\right)$ under this model iterates over the following updates,

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}_{new} = \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}_{old} + \widehat{I}\left(\hat{\beta}_{old}, \hat{\gamma}_{old}\right)^{-1} \widehat{S}\left(\hat{\beta}_{old}, \hat{\gamma}_{old}\right),$$

where $\widehat{S}$ and $\widehat{I}$ are the estimated score function and the information matrix where the parameters are replaced by their estimates at the current iteration. The computations required to calculate the information matrices and to invert them at each iteration are $O(nK^2)$ and $O(K^3)$ respectively, where K is the number of covariates. When analyzing millions of variants across thousands of phenotypes in models including large numbers of covariates, the computation time can become substantially large.

Our method is motivated by the observations that for relatively small effect sizes in the log-odds scale (LORs), the logistic model behaves closely to a linear model [Pirinen et al., 2013], and the score functions for linear and logistic regressions are algebraically of similar forms. We thus propose to estimate the genotype LORs by fitting a genotype-only model, using analogous estimation techniques used in linear models. For linear regression, the mle for the genotype LOR in the full model $E(Y_i|X_i, G_i) = X_i^\top \beta + G_i\gamma$ is given by $\widehat{\gamma}_{lin} = \left(G^\top(I - P_X)G\right)^{-1} G^\top \tilde{Y}$, where $P_X = X\left(X^\top X\right)^{-1} X^\top$ is the projection matrix of the covariates, and $\tilde{Y} = (I - P_X)Y$ is the residual from the null model $E(Y_i|X_i) = X_i^\top \beta$. If we use $\widetilde{Y}_i$s as outcomes in the genotype-only model $E\left(\widetilde{Y}_i|G_i\right) = \alpha + G_i\gamma$, the estimate of γ is given by $\widetilde{\gamma}_{lin} = \left(G^\top(I - J)G\right)^{-1} G^\top \widetilde{Y}$, where nJ is the $n \times n$ matrix of all elements equal to unity. Therefore, $\widetilde{\gamma}_{lin} = f^2 \widehat{\gamma}_{lin}$, where the scaling factor $f = \left\{\left(G^\top(I - P_X)G\right) / \left(G^\top(I - J)G\right)\right\}^{1/2}$. It is clear from this linear regression setup, that the mles of γ from the genotype-only model $E\left(\widetilde{Y}_i/f|G_i\right) = \alpha + fG_i\gamma$ will be identical to $\widehat{\gamma}_{lin}$. The score function corresponding to this model is,

$$S(\alpha, \gamma) = \begin{pmatrix} 1^\top \\ fG^\top \end{pmatrix} \left(\widetilde{Y}/f - E\left(\widetilde{Y}/f|G\right)\right). \tag{2}$$

Since, the score functions of the logistic and linear regressions are of the same form, we propose to regress the binary outcomes Y on the scaled genotype vector fG in a logistic regression, using the same score function as 2. Assuming G to be mean-centered, in order to keep the prevalence the same, we use a modified score function,

$$S^*(\alpha, \gamma) = \begin{pmatrix} 1^\top \\ fG^\top \end{pmatrix} \left(\widetilde{Y}/f + \bar{Y} - \mu^*\right),$$

where $\mu_i^* = [1 + \exp\{\alpha + G_i\gamma\}]^{-1}$. Notice that, this modification does not change the component corresponding to G in the score function as G is mean-centered. We solve the score equation $S^*(\alpha, \gamma) = 0$ by iterating through the following Newton-Raphson updates,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix}_{new} = \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix}_{old} + \widehat{I}^* \left(\hat{\alpha}_{old}, \hat{\gamma}_{old}\right)^{-1} \widehat{S}^* \left(\hat{\alpha}_{old}, \hat{\gamma}_{old}\right),$$

where $\widehat{S}^*$ and $\widehat{I}^*$ are the estimated score function and the information matrix where the parameters are replaced by their estimates at the current iteration. The information matrix is given by,

$$I^*(\alpha, \gamma) = \begin{bmatrix} 1^\top W^* 1 & 1^\top W^* G \\ G^\top W^* 1 & G^\top W^* G \end{bmatrix},$$

where $W^* = diag(\mu_i^*(1 - \mu_i^*))$. The computations required to calculate the information matrices and to invert them at each iteration are $O(n)$ and $O(1)$ respectively. Therefore, the computation do not depend on K, as the information matrices are always of order 2×2. Therefore, computationally this method provides substantial improvement over the logistic regression on the full model.

For unbalanced case-control studies, we can further correct the bias in our score function by adjusting it based on the Firth's bias correction method [Firth, 1993],

$$S_F^*(\alpha, \gamma) = S^*(\alpha, \gamma) + \frac{1}{2} \begin{pmatrix} tr\left[I^{*-1}\left\{\partial I^*/\partial\alpha\right\}\right] \\ tr\left[I^{*-1}\left\{\partial I^*/\partial\gamma\right\}\right] \end{pmatrix}.$$

2.1 Simulation Studies

To evaluate the proposed method using numerical simulations, we considered three different case-control ratios: balanced ($case : control = 1 : 1$), moderately unbalanced ($case : control = 1 : 9$), and extremely unbalanced ($case : control = 1 : 99$). For each choice of case-control ratio, the phenotypes were simulated based on the following logistic regression model,

$$logit\left[P(Y_i = 1 | X_{1i}, X_{2i}, G_i)\right] = \beta_0 + 0.5\left(X_{1i} + X_{2i}\right) + \gamma G_i, \quad i = 1, \ldots, n,$$

where the two covariates X_{1i} and X_{2i} were simulated from $X_{1i} \sim Bernoulli(0.5)$ and $X_{2i} \sim N(0, 1)$, and the intercept β_0 corresponds to the prevalence rate of 1%. We considered two settings to generate the genotypes: (i) uncorrelated with the covariates, where $G_i \sim Binomial(2, p)$, and (ii) correlated with the covariate X_{2i}s, where $G_i \sim Binomial\left(2, p^*\right)$, $p^* = p(1 - 1/2r + X_{2i}/r)$. The minor allele frequencies (MAF) were chosen to be $p = 0.001, 0.01, 0.05$, and the correlation factors (r) was chosen to represent allele frequency differences of $0.5p$ (highly correlated) and $0.2p$ (moderately correlated) between the two groups denoted by $X_{2i} = 0$ and 1. The uncorrelated setting corresponds to $r \to \infty$. Under each case-control ratio, MAF, and r, we simulated 200 datasets of sample size $n = 20000$ each using different values of the genotype LOR γ (see Figures 8–10). Then, we applied our proposed method (Logistic-Reduced) and Firth's bias-controlled logistic regression [Ma et al., 2013] under the full model (Logistic-Full), and compared the estimated genotype LORs from these two methods.

To compare the computation times under different numbers of covariates, we simulated datasets with sample size $n = 20000$ from the following logistic regression model,

$$logit\left[P(Y_i = 1|X_i^\top, G_i)\right] = \beta_0 + 0.5\sum_{j=1}^{K} X_{ij} + \gamma G_i, \quad i = 1, \ldots, n,$$

where K is the number of covariates. The model is similar to the previously discussed simulation setting, except all the covariates X_{ij}s follow i.i.d. $N(0, 1)$ distribution. The genotype LORs (γ) are also selected randomly from $U(-2, 2)$ for each dataset. For each of the three case-control ratios, and $K = 3, 6, 10, 15, 20, 25$, we simulated 500 datasets, and compared the computation times for the Logistic-Full and Logistic-Reduced methods.

3 Results

We compare the estimated genotype LORs under different case-control ratios and computation times under different numbers of covariates based on extensive numerical simulations.

3.1 Simulation Studies

We compare the estimated genotype LORs based on Logistic-Full and Logistic-Reduced methods under different case-control ratios, MAFs, and genotype-covariate correlations. Figures 2,3 and 4 present the scatter plot of the estimated LORs from these two methods for the simulations where genotypes and the covariates are highly correlated, moderately correlated and uncorrelated, respectively. The results show that the estimated LORs are almost identical between these two methods, which suggests that the Logistic-Reduced method provides almost as accurate estimates as the Logistic-Full method. When the case-control ratio is moderately or extremely unbalanced, and the true LOR is very small, there is very little underestimation in Logistic-Reduced compared to Logistic-Full. The comparison of the estimated LORs with the true parameter values for these two methods are presented in Figures 5–10. We notice that, for the rare variants (MAF = 0.01 or 0.001) in the extremely unbalanced case-control setting, both Logistic-Full and Logistic-Reduced methods have extremely large sampling variances, especially when the true LOR is negative. Moreover, the estimates seem to have discrete levels, as the expected number of minor alleles in cases becomes smaller than one. This observation is clearer when MAF = 0.001, where both the Logistic-Full and Logistic-Reduced estimates cluster around zero for negative values of the true LORs.

Figure 1 presents the projected computation times for analyzing 10 million variants across 1500 phenotypes on a dataset of 20000 samples. The projected computation times clearly show that Logistic-Reduced is faster than Logistic-Full. For example, 197 CPU-years is required to analyze a moderately unbalanced dataset using Logistic-Full compared to only 16 CPU-years using Logistic-Reduced, when the number of covariates is $K = 20$. Moreover, the computation times of the Logistic-Full method increase at a faster than linear rate with the number of covariates, whereas the computation times of the Logistic-Reduced method remains almost the same.

4 Conclusion

We proposed a faster computation method for genotype LORs in case-control studies, which reduces the computation complexity from $O(nK^2 + K^3)$ to $O(n)$ compared to the traditional logistic regression with Firth bias-correction (Logistic-Full), where n is the sample-size, and K is the number of covariates. Our proposed method requires $\sim$ 8–10x less computation time when analyzing a logistic regression model with 15 covariates, and $\sim$ 16–20x less computation time when there are 25 covariates. Using simulation studies, we showed that for relatively smaller genotype LORs, the estimates from our method are almost identical to the estimates from the Logistic-Full method across different case-control ratios and rarity of alleles.

Acknowledgments

The research was supported by NIH R01 HG008773.

References

Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, 06 2015. ISSN 0300-5771. doi: 10.1093/ije/dyv080. URL `https://doi.org/10.1093/ije/dyv080`.

Stephen Burgess, Adam Butterworth, and Simon G. Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665, 2013. doi: 10.1002/gepi.21758. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21758`.

Stephen Burgess, Robert A. Scott, Nicholas J. Timpson, George Davey Smith, Simon G. Thompson, and EPIC- InterAct Consortium. Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology*, 30(7):543–552, Jul 2015. ISSN 1573-7284. doi: 10.1007/s10654-015-0011-z. URL `https://doi.org/10.1007/s10654-015-0011-z`.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Gil McVean, Stephen Leslie, Peter Donnelly, and Jonathan Marchini. Genome-wide genetic data on 500,000 uk biobank participants. *bioRxiv 166298 doi:10.1101/166298*, 2017.

Sayantan Das, Lukas Forer, Sebastian Schonherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, Emily Y. Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G. Iacono, Anand Swaroop, Laura J. Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Goncalo R. Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods.

Nat Genet, 48(10):1284–1287, 2016. ISSN 1061-4036. doi: 10.1038/ng.3656. URL http://dx.doi.org/10.1038/ng.3656.

Rounak Dey, Ellen M. Schmidt, Goncalo R. Abecasis, and Seunggeun Lee. A fast and accurate algorithm to test for binary phenotypes and its application to phewas. *The American Journal of Human Genetics*, 101(1):37 – 49, 2017. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2017.05.014. URL http://www.sciencedirect.com/science/article/pii/S000292971730201X.

Rounak Dey, Jonas B. Nielsen, Lars G. Fritsche, Wei Zhou, Huanhuan Zhu, Cristen J. Willer, and Seunggeun Lee. Robust meta-analysis of biobank-based genome-wide association studies with unbalanced binary phenotypes. *Genetic Epidemiology*, 0(0), 2019. doi: 10.1002/gepi.22197. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22197.

E. Evangelou and J. P. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet*, 14(6):379–89, 2013. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: 10.1038/nrg3472. URL https://www.ncbi.nlm.nih.gov/pubmed/23657481.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993. ISSN 00063444. URL http://www.jstor.org/stable/2336755.

Lars G. Fritsche, Stephen B. Gruber, Zhenke Wu, Ellen M. Schmidt, Matthew Zawistowski, Stephanie E. Moser, Victoria M. Blanc, Chad M. Brummett, Sachin Kheterpal, Gonçalo R. Abecasis, and Bhramar Mukherjee. Association of polygenic risk scores for multiple cancers in a phenome-wide study: Results from the michigan genomics initiative. *The American Journal of Human Genetics*, 102(6):1048 – 1061, 2018. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2018.04.001. URL http://www.sciencedirect.com/science/article/pii/S0002929718301332.

S. J. Hebbring. The challenges, advantages and future of phenome-wide association studies. *Immunology*, 141(2):157–65, 2014. ISSN 1365-2567 (Electronic) 0019-2805 (Linking). doi: 10.1111/imm.12195. URL https://www.ncbi.nlm.nih.gov/pubmed/24147732.

S. Krokstad, A. Langhammer, K. Hveem, T. L. Holmen, K. Midthjell, T. R. Stene, G. Bratberg, J. Heggland, and J. Holmen. Cohort profile: the hunt study, norway. *Int J Epidemiol*, 42(4):968–77, 2013. ISSN 1464-3685 (Electronic) 0300-5771 (Linking). doi: 10.1093/ije/dys095. URL https://www.ncbi.nlm.nih.gov/pubmed/22879362.

Clement Ma, Tom Blackwell, Michael Boehnke, and Laura J. Scott. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology*, 37(6):539–550, 2013. doi: 10.1002/gepi.21742. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21742.

Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, 2010. ISSN 1471-0056. doi: 10.1038/nrg2796. URL http://dx.doi.org/10.1038/nrg2796.

S. A. Pendergrass, K. Brown-Gentry, S. Dudek, A. Frase, E. S. Torstenson, R. Goodloe, J. L. Ambite, C. L. Avery, S. Buyske, P. Buzkova, E. Deelman, M. D. Fesinmeyer, C. A. Haiman, G. Heiss, L. A. Hindorff, C. N. Hsu, R. D. Jackson, C. Kooperberg, L. Le Marchand, Y. Lin, T. C. Matise, K. R. Monroe, L. Moreland, S. L. Park, A. Reiner, R. Wallace, L. R. Wilkens, D. C. Crawford, and M. D. Ritchie. Phenome-wide association study (phewas) for detection of pleiotropy within the population architecture using genomics and epidemiology (page) network. *PLoS Genet*, 9(1):e1003087, 2013. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1003087. URL `https://www.ncbi.nlm.nih.gov/pubmed/23382687`.

Matti Pirinen, Peter Donnelly, and Chris C. A. Spencer. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390, 2013. ISSN 19326157. URL `http://www.jstor.org/stable/23566515`.

Lam C. Tsoi, Philip E. Stuart, Chao Tian, Johann E. Gudjonsson, Sayantan Das, Matthew Zawistowski, Eva Ellinghaus, Jonathan N. Barker, Vinod Chandran, Nick Dand, Kristina Callis Duffin, Charlotta Enerbäck, Tõnu Esko, Andre Franke, Dafna D. Gladman, Per Hoffmann, Külli Kingo, Sulev Kõks, Gerald G. Krueger, Henry W. Lim, Andres Metspalu, Ulrich Mrowietz, Sören Mucha, Proton Rahman, Andre Reis, Trilokraj Tejasvi, Richard Trembath, John J. Voorhees, Stephan Weidinger, Michael Weichenthal, Xiaoquan Wen, Nicholas Eriksson, Hyun M. Kang, David A. Hinds, Rajan P. Nair, Gonçalo R. Abecasis, and James T. Elder. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nature Communications*, 8:15382, 2017. URL `https://doi.org/10.1038/ncomms15382`.

A. Verma, A. Lucas, S. S. Verma, Y. Zhang, N. Josyula, A. Khan, D. N. Hartzel, D. R. Lavage, J. Leader, M. D. Ritchie, and S. A. Pendergrass. Phewas and beyond: The landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from geisinger. *Am J Hum Genet*, 102(4):592–608, 2018. ISSN 1537-6605 (Electronic) 0002-9297 (Linking). doi: 10.1016/j.ajhg.2018.02.017. URL `https://www.ncbi.nlm.nih.gov/pubmed/29606303`.

Cristen J. Willer, Yun Li, and Gonçalo R. Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 07 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq340. URL `https://doi.org/10.1093/bioinformatics/btq340`.

Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken B. Elvestad, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *bioRxiv*, 2017. doi: 10.1101/212357. URL `https://www.biorxiv.org/content/biorxiv/early/2017/11/24/212357.full.pdf`.

Figures

Figure 1: The projected computation times for estimating the genotype LORs using different methods for 10 million variants across 1500 Phenotypes with different number of covariates. The computation times are based on testing 500 simulated variants on an Intel i7 2.70GHz processor and then projecting them onto a PheWAS with 10 million variants and 1500 phenotypes. The solid lines represent the computation times required for Logistic-Full, and the dashed lines represent the computation times required for Logistic-Reduced.

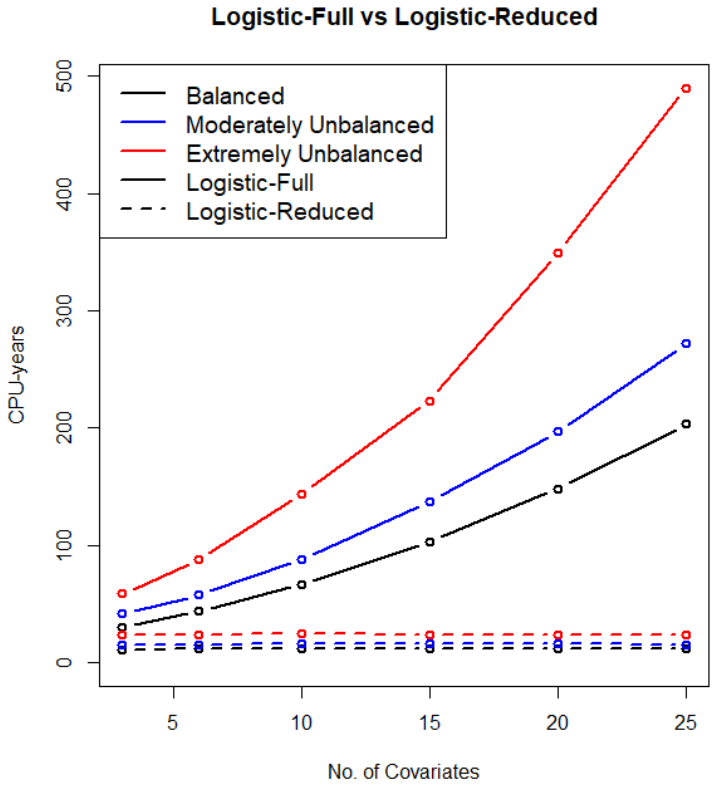

Figure 2: Scatter plots of estimated genotype LORs from Logistic-Full (x-axis) and Logistic-Reduced (y-axis) methods when the genotypes and covariates are highly correlated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

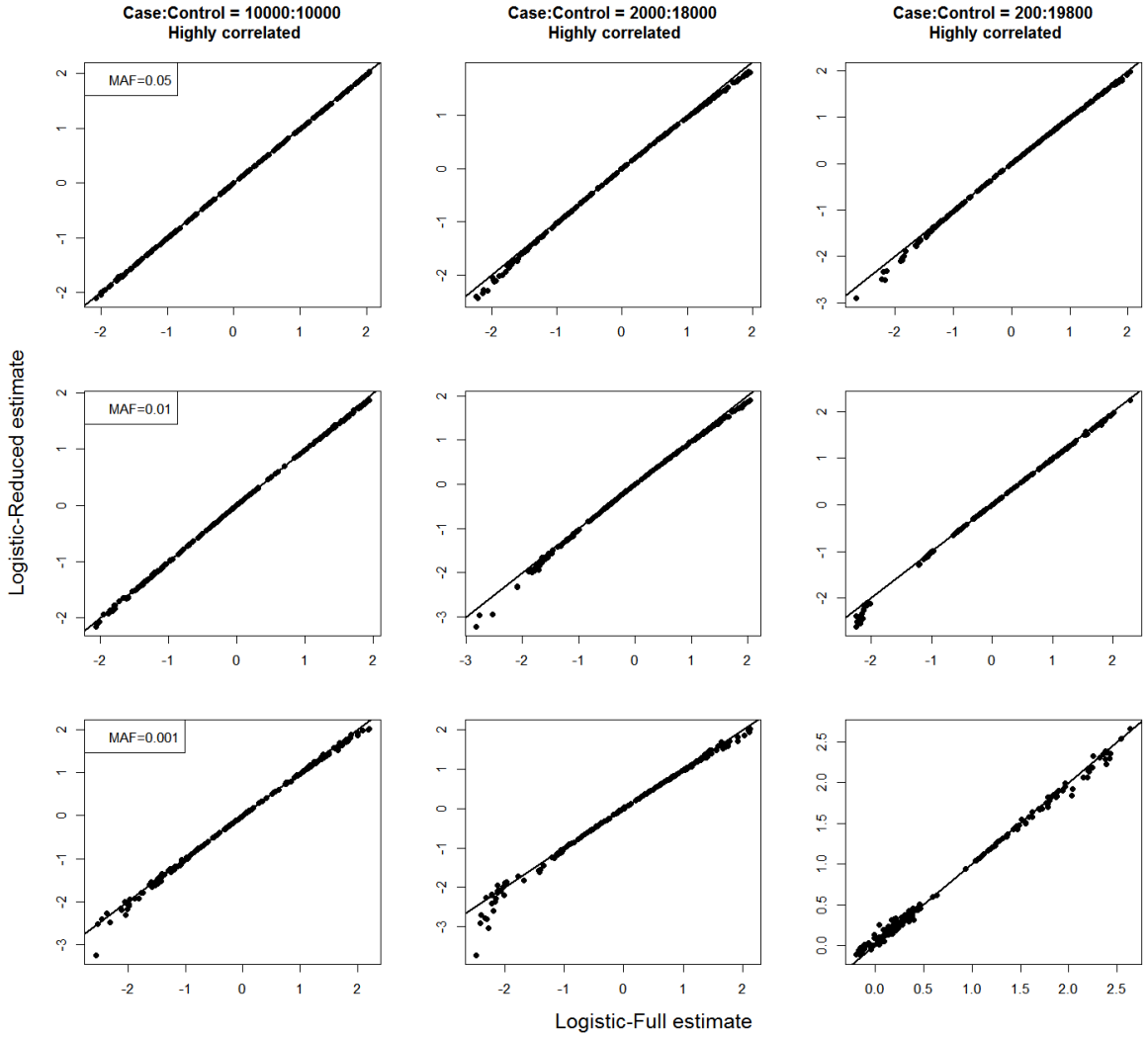

Figure 3: Scatter plots of estimated genotype LORs from Logistic-Full (x-axis) and Logistic-Reduced (y-axis) methods when the genotypes and covariates are moderately correlated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

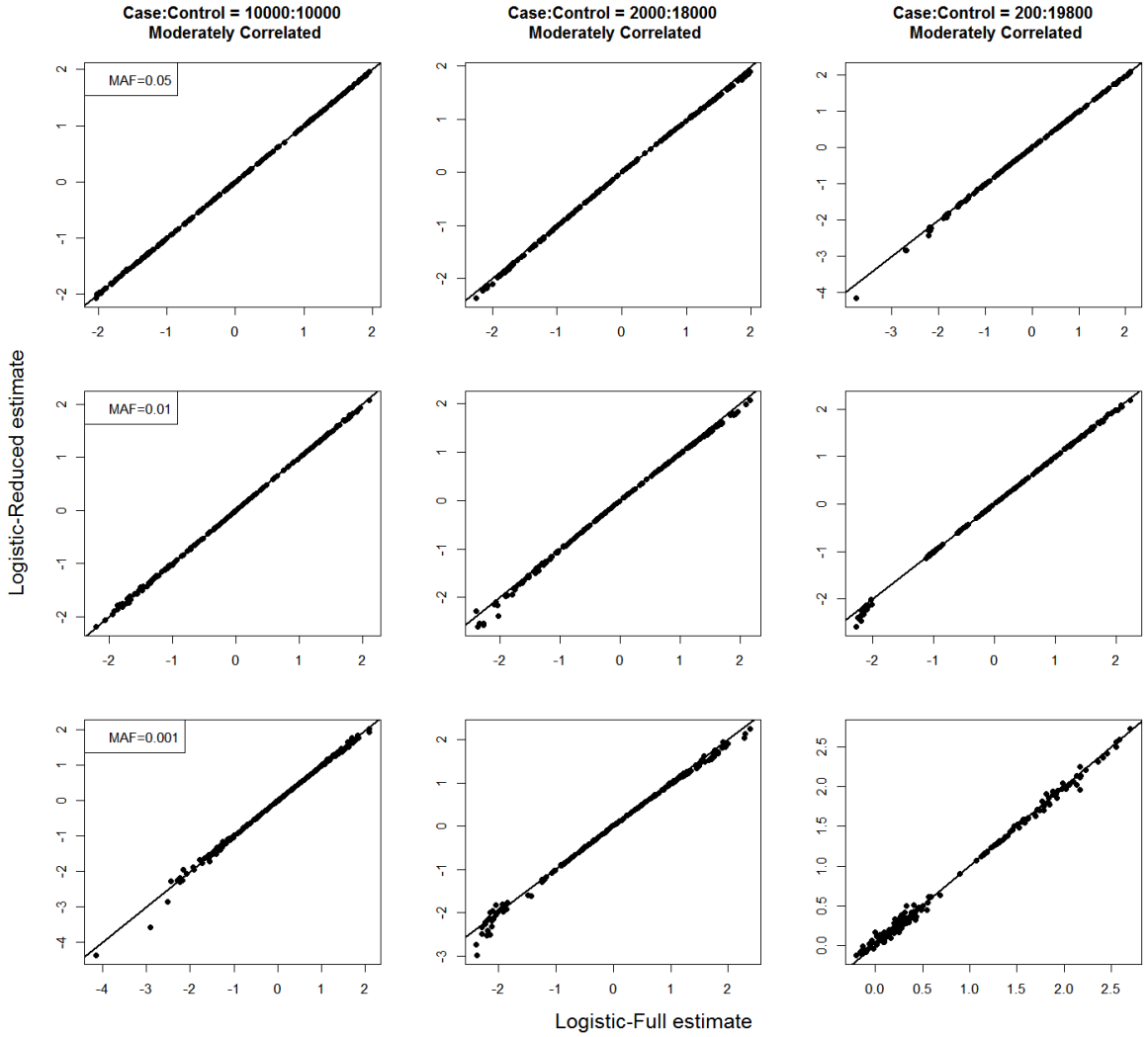

Figure 4: Scatter plots of estimated genotype LORs from Logistic-Full (x-axis) and Logistic-Reduced (y-axis) methods when the genotypes and covariates are uncorrelated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

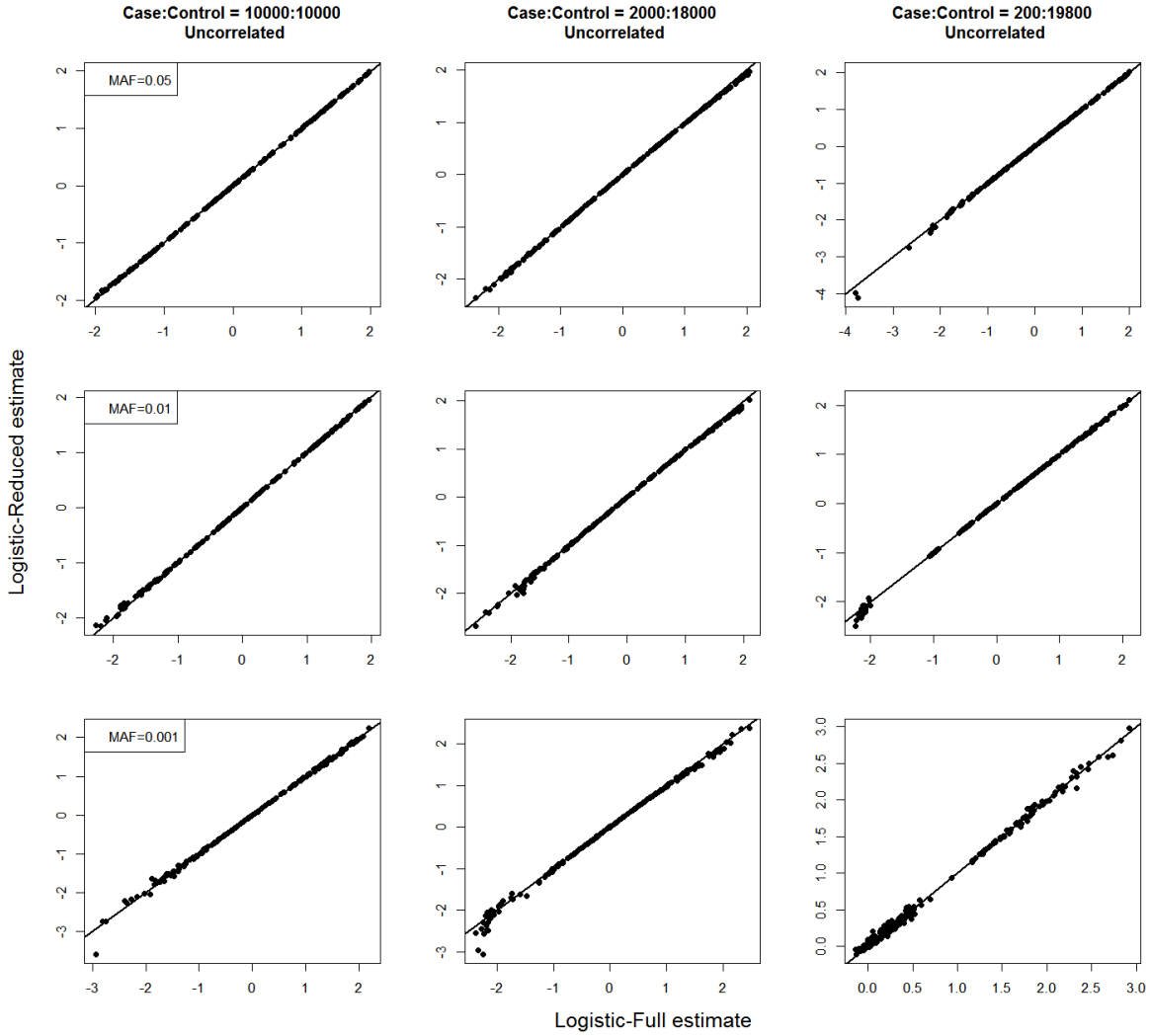

Figure 5: Scatter plots of the true parameter values (x-axis) and estimated genotype LORs from the Logistic-Full (y-axis) methods when the genotypes and covariates are highly correlated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

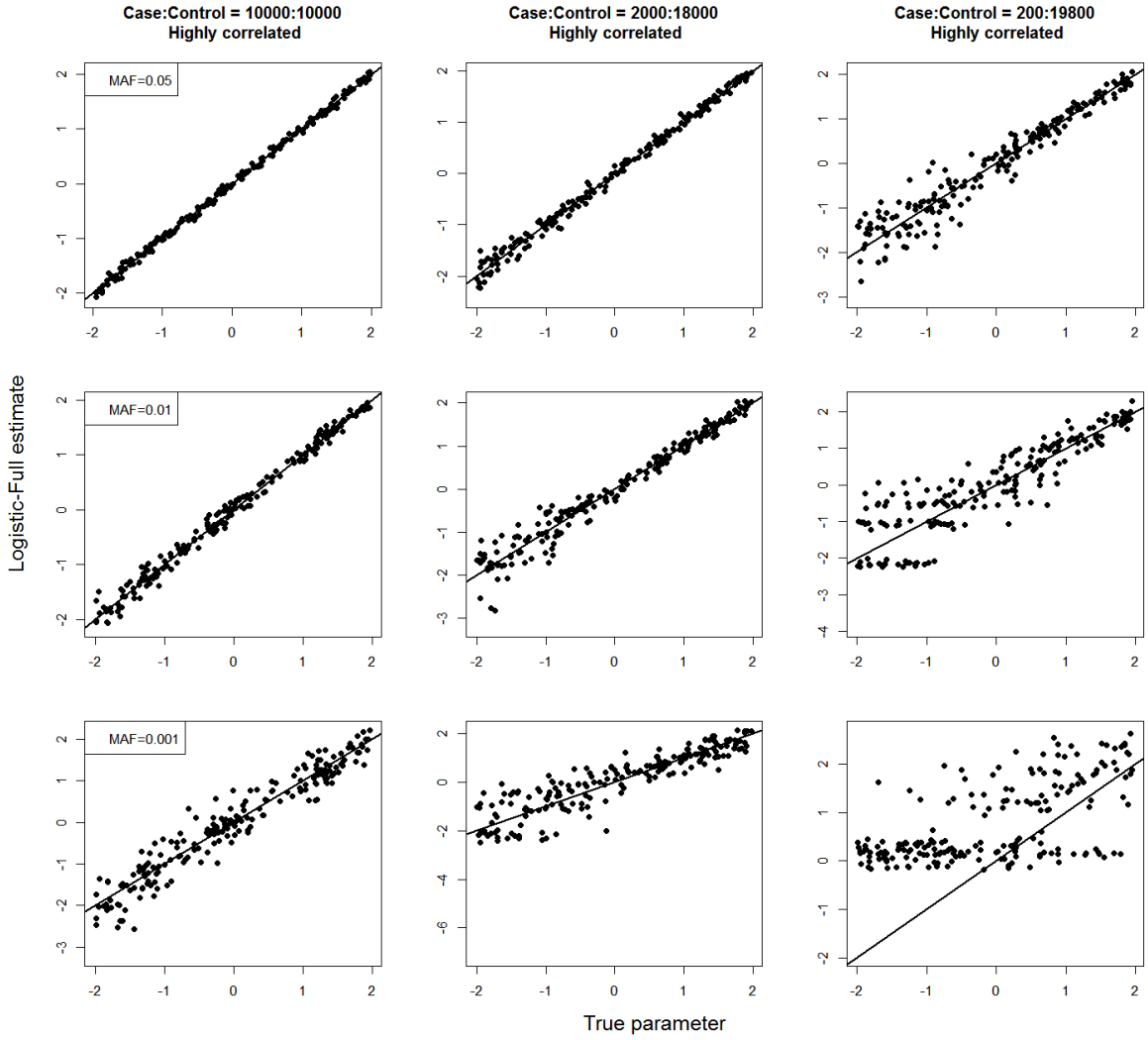

Figure 6: Scatter plots of the true parameter values (x-axis) and estimated genotype LORs from the Logistic-Full (y-axis) methods when the genotypes and covariates are moderately correlated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

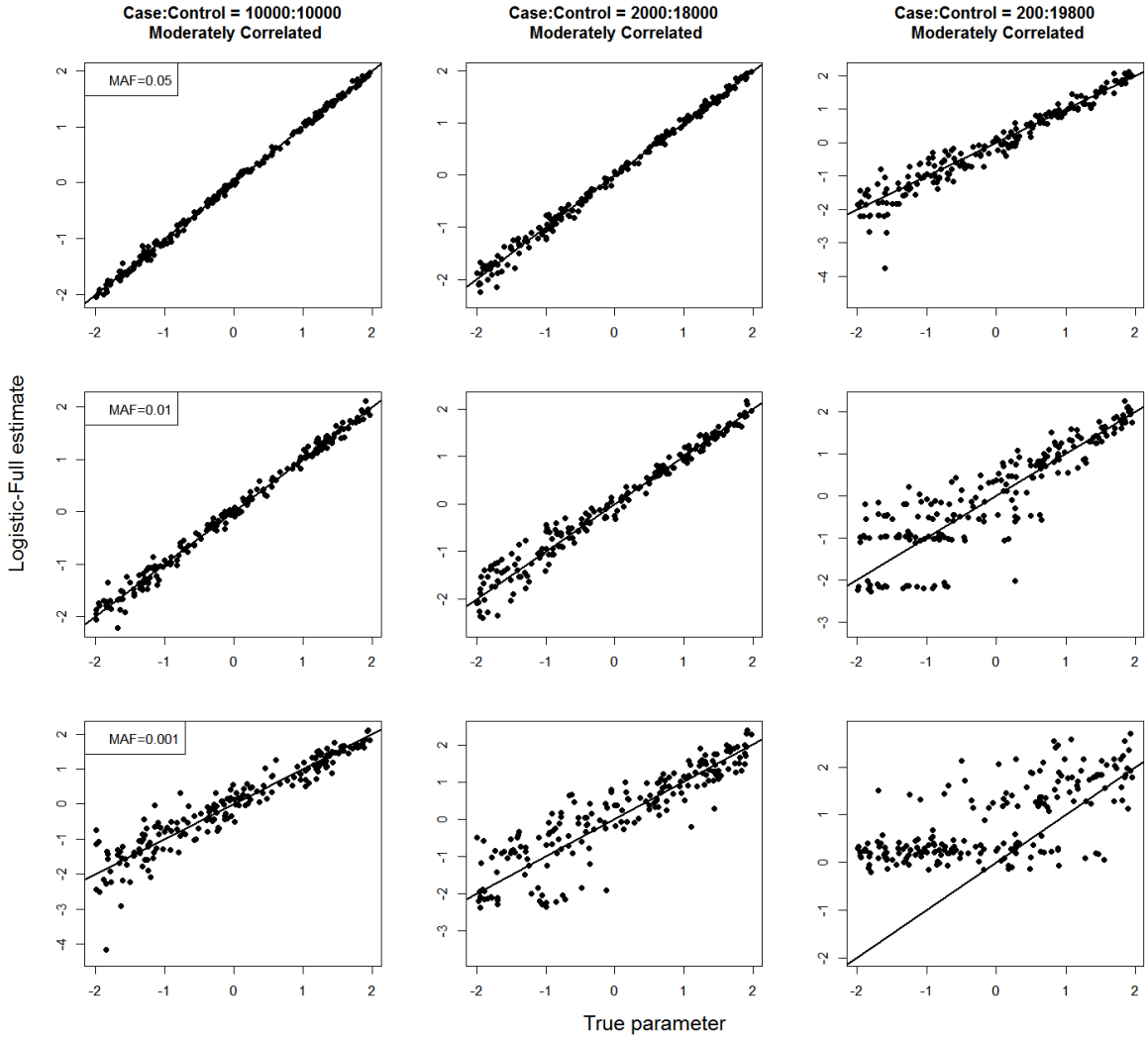

Figure 7: Scatter plots of the true parameter values (x-axis) and estimated genotype LORs from the Logistic-Full (y-axis) methods when the genotypes and covariates are uncorrelated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

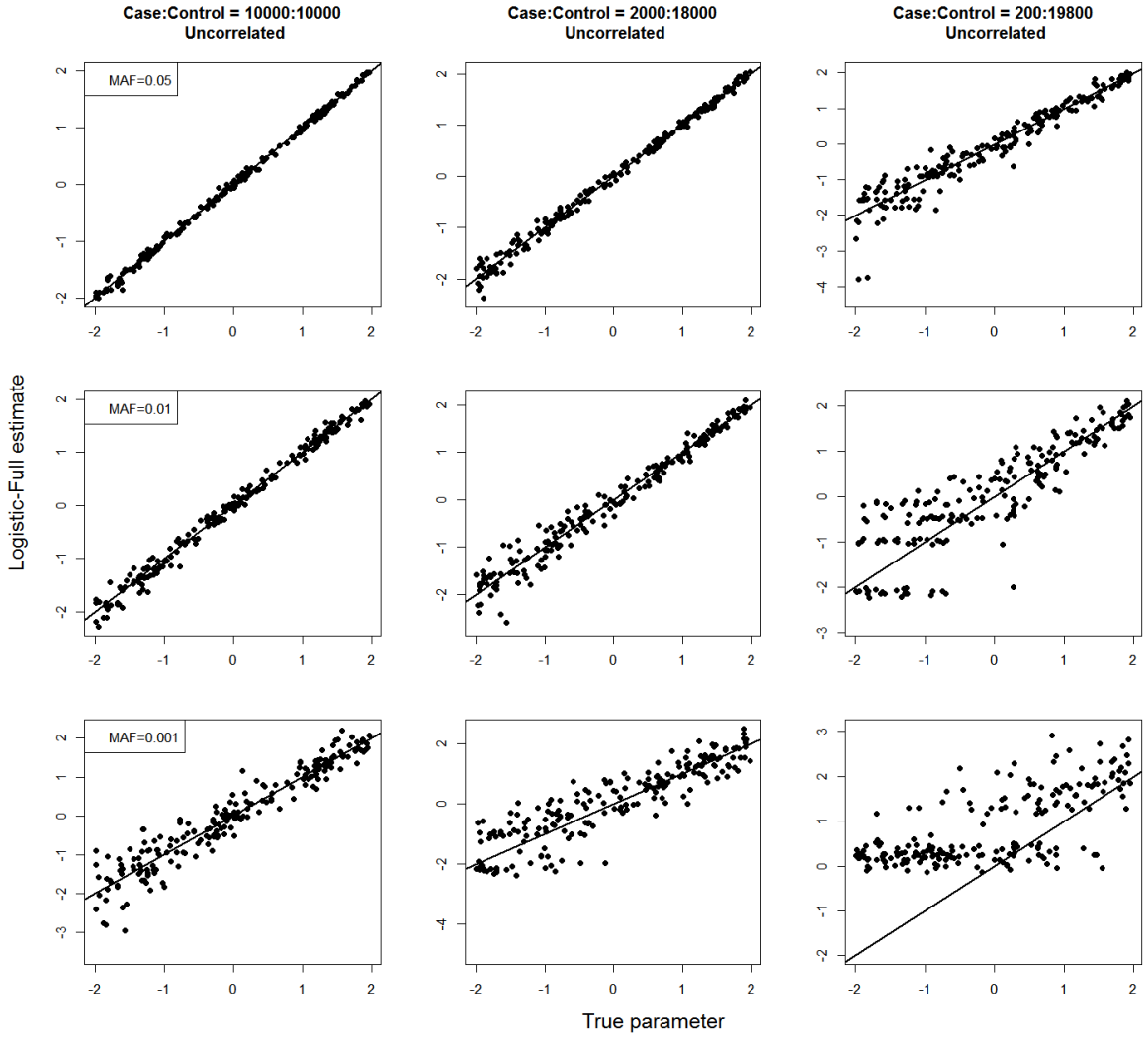

Figure 8: Scatter plots of the true parameter values (x-axis) and estimated genotype LORs from the Logistic-Reduced (y-axis) methods when the genotypes and covariates are highly correlated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

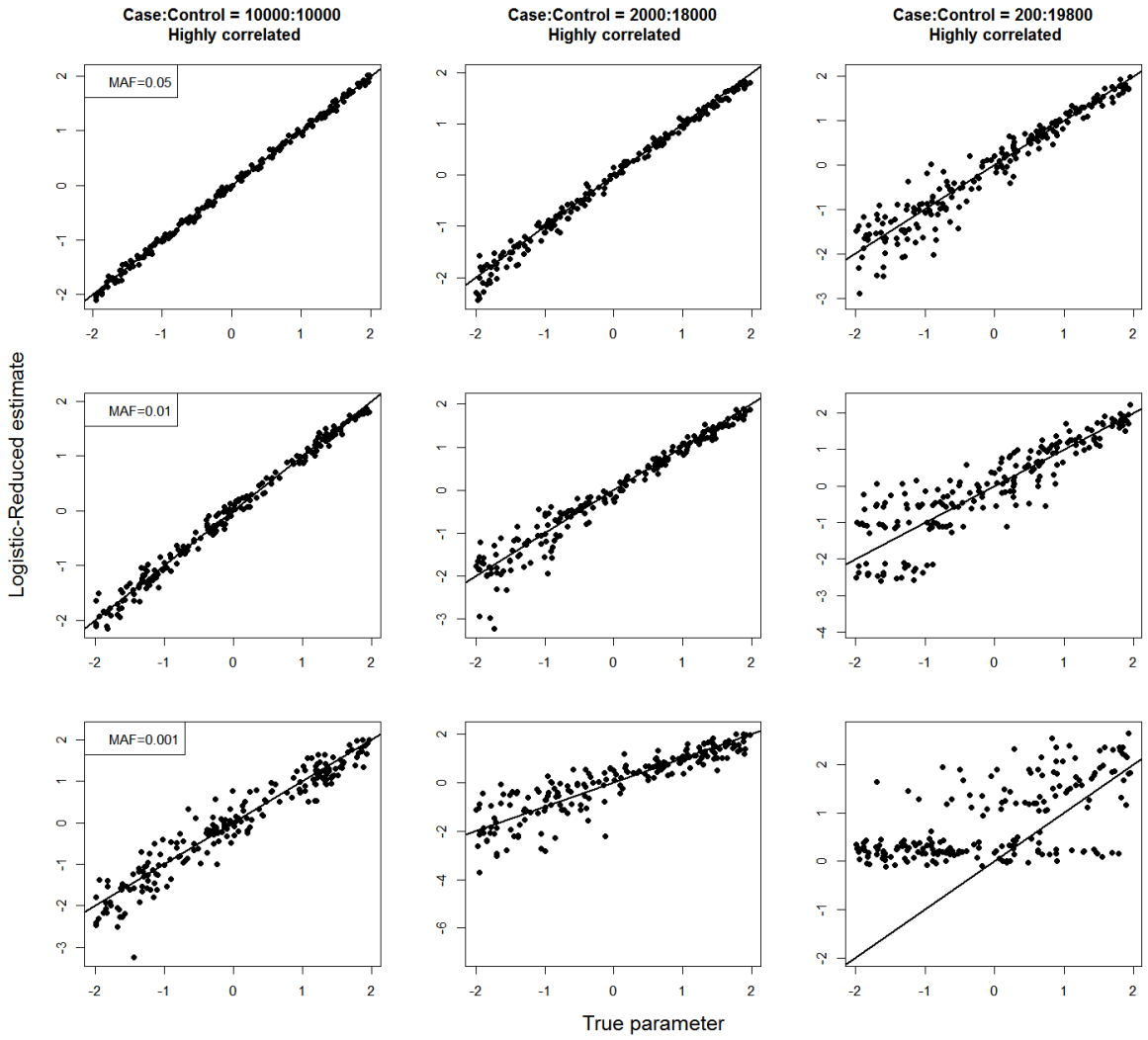

Figure 9: Scatter plots of the true parameter values (x-axis) and estimated genotype LORs from the Logistic-Reduced (y-axis) methods when the genotypes and covariates are moderately correlated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.

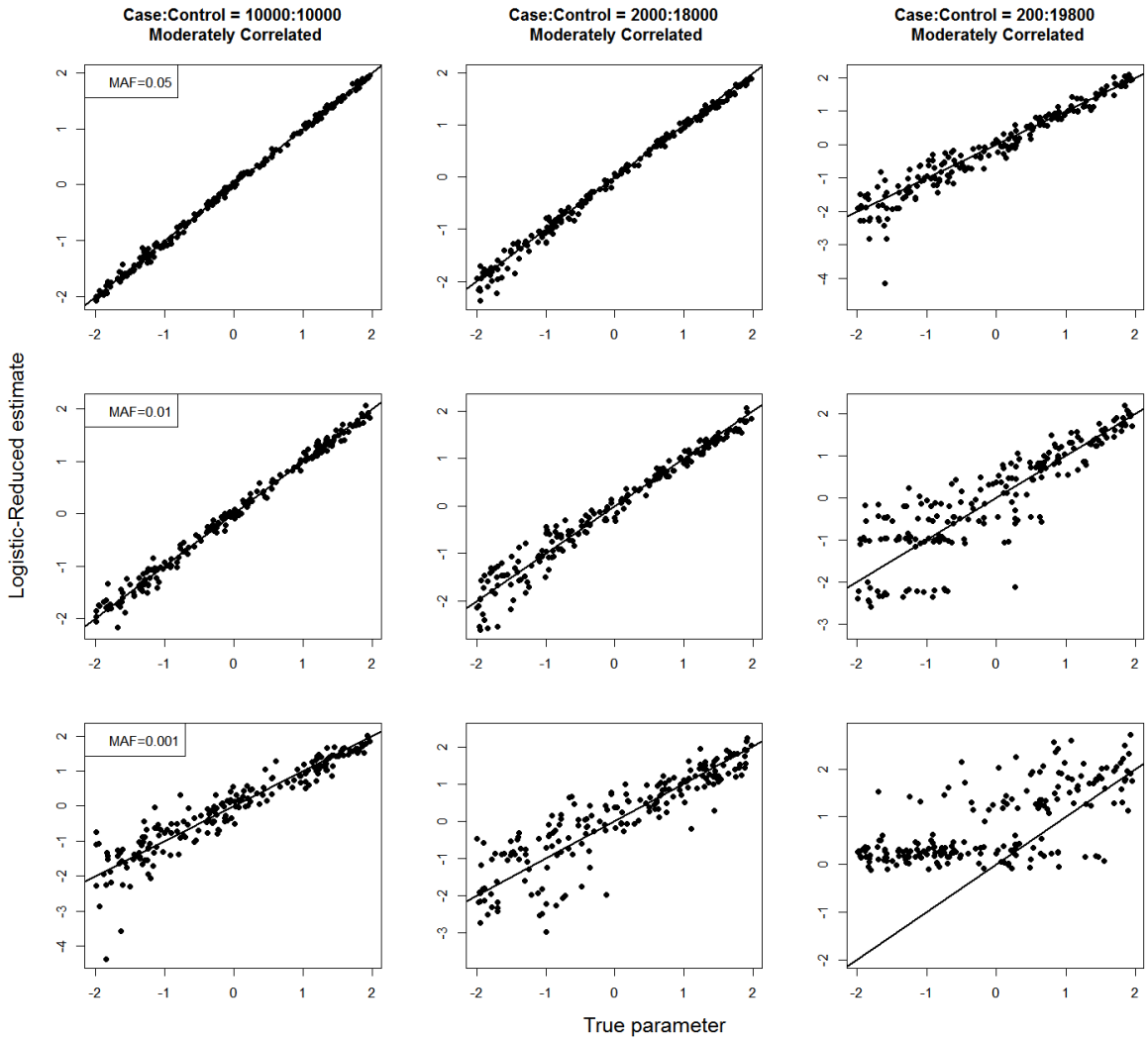

Figure 10: Scatter plots of the true parameter values (x-axis) and estimated genotype LORs from the Logistic-Reduced (y-axis) methods when the genotypes and covariates are uncorrelated. From left to right, the panels represent case-control ratios $1:1$, $1:9$ and $1:99$, respectively. From top to bottom, the panels represent MAFs $0.001, 0.01$ and 0.05, respectively.