bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Estimating genetic inheritance in case-control studies

Na Li[1,2], Jiayan Zhu[3*], Zhengbang Li[4], Qizhai Li[1]
[1]LSC, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of information engineering, Hubei University of Chinese Medicine, Wuhan, 430065, China
[4]School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China

*Corresponding author

Abstract

Case-control genetic association study is an efficient tool to search for the deleterious genetic variants predispose to human complex diseases, where the additive mode of inheritance is commonly assumed. However, how the genetic variants influence the occurrence of a certain disease is impossible to know beforehand. We show numerically that the existing procedures using the Hardy-Weinberg equilibrium test to choose the genetic model might be inconsistent. Then we propose a new consistent method to choose the genetic model. The extensive computer simulation results demonstrate the superiorities of the new method. Applications to six single nucleotide polymorphisms(SNPs) for breast cancer and eight SNPs for Type 2 Diabetes further show the performances of the proposed method.

Key words: Case-control study, Genetic model, Consistency,Genetic inheritance, Parameter estimating.

1 Introduction

The case-control design is an efficient tool for collecting the information of covariate being assumed and tested whether they are associated with human complex diseases in epidemiologic studies. Although it is a retrospective design, the logistic regression model by taking the data as enrolled prospectively is still valid to estimate the coefficients for covariate (Prentice & Pyke, 1979) since the maximum likelihood estimator is consistent and has the asymptotic normality property. Among many covariate such as body mass index, age and blood pressure etc, the genetic variant is one of the most important one. Comparing with insertion, deletion and copy number variation etc, the single-nucleotide polymorphism (SNP), which is a genetic variation that occurs at a specific position in the genome, is a

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

more common form. By now, more than ten thousand SNPs have been identified to be associated with hundreds of human diseases.

When doing an association study, a genetic model needs to be assumed in advance, which refers to a genetic mode of inheritance. Specifying a genetic model means specifying an alternative hypothesis. There are often three commonly used genetic models including recessive, additive and dominant ones. In reality, it is rarely to know the real genetic models. The additive mode of inheritance has been used in many genetic studies (Klein et al., 2005; Hunter et al., 2007; Zheng, Li and Yuan, 2014). However, there are also some other SNPs conferring risk to disease at other modes. For example, Moltke et al. (2014) found a genetic variant p.Arg684Ter associated with 2-h plasma glucose levels and type 2 diabetes at a recessive mode; Nik-Zainal et al. (2016) reported five genes with MED23, FOXP1, MLLT4, XBP1, and ZFP36L1, acting in breast cancer also in a recessive fashion. Mis-specifying the genetic model will result in loss of statistical power. Especially, the genetic model is recessive and the dominant model is adopted and vice visa. So, the Wellcome Trust Case Control Consortium (2007) used the minimum of p-values for score test under the additive model and Pearson Chi-square with 2 degrees of freedom to search for the genetic variant associated with seven common diseases including bipolar disorder, coronary artery disease, crohn disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. Sladek et al. (2007) proposed to use MAX (the maximum values of score tests derived under three genetic models) to search for the evidence of SNPs associated with type 2 diabetes. Identifying the genetic model is an challenge problem. As far as we know, there is only work on the basis of Hardy-Weinberg equilibrium test (HWET) to choose the genetic model (Ng and Zheng, 2008; Zheng et al., 2016; Hu et al., 2017). The Hardy-Weinberg equilibrium principle is an important law in population genetics and holds in health individuals of human populations. We will show numerically later that using HWET to discovery the genetic model is inconsistent when the genetic model is additive. This motivates us to develop a new statistical methodology to fill this gap.

In this work, we develop a general framework to infer the genetic model. Our procedure has four merits. Firstly, the methodology is developed in a general setting where a parameter $\theta \in [0, 1]$ is used to represent a genetic model, which not only includes the recessive, additive and dominant models, but also contains other models. It is very suitable for the situation where the imperfect surrogate of the causal SNP is genotyped and the genetic model of the surrogate is derived from the true model of the causal SNP. Secondly, the proposed estimate of genetic model is consistent, which makes up for the shortcoming of using HWET to choose the model. Thirdly, the proposed procedure allows for handling the confounder factors, however, the existing method can not deal with it. Fourthly, we use a binary variable to replace the original genotype values. It can segregate the coefficient of

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genotype out and make the estimation of parameters feasible.

2 Methods

2.1 Notations and genetic models

Denote by Y the disease status with $Y = 1$ being an individual having a disease and $Y = 0$ being healthy control. Let $\mathbf{X}$ and G be the m-dimensional covariate and genotype value, respectively. To detect the relationship between Y and G with adjusting for the effect of $\mathbf{X}$, a typical model is the logistic regression as

$$P(Y = 1|\mathbf{X}, G) = \frac{\exp(\alpha + \mathbf{X}^\top\gamma + G\beta)}{1 + \exp(\alpha + \mathbf{X}^\top\gamma + G\beta)},$$

where α, γ and β are the parameters, and $\top$ denotes the transpose of a matrix or a vector. We consider a diallelic SNP locus and let the two alleles at a SNP locus be A and a, and A is assumed to be the risk allele or minor allele. Then there are three genotypes as aa, Aa and AA and the corresponding genotype values are 0, θ and 1, respectively, $0 \leq \theta \leq 1$. Here θ denotes the genetic model. For example, we can set θ be 0, 0.5 and 1 for the commonly used recessive, additive and dominant models, respectively.

Suppose that there are n subjects including r cases and s controls randomly drawn from case population and control population, respectively. Let $(y_i, \mathbf{x}_i^\top, g_i)^\top$ be the observation of the ith subject for $(Y, \mathbf{X}^\top, G)^\top$, $i = 1, 2, \cdots, n$. For the sake of simplicity, we assume that the first r subjects are cases and the last s subjects are controls.

2.2 Using HWE Test to choose the model

Denote the genotype frequencies of (aa, Aa, AA) in cases and controls by (p_0, p_1, p_2) and (q_0, q_1, q_2), respectively, and let the number of subjects with genotypes (aa, Aa, AA) be (r_0, r_1, r_2) in r cases and that in s controls be (s_0, s_1, s_2). Denote $(\hat{p}_0, \hat{p}_1, \hat{p}_2) = (r_0/r, r_1/r, r_2/r)$, $(\hat{q}_0, \hat{q}_1, \hat{q}_2) = (s_0/s, s_1/s, s_2)$, $\hat{\Delta}_1 = \hat{p}_2 - (\hat{p}_2 + \hat{p}_1/2)^2$, $\hat{\Delta}_2 = \hat{q}_2 - (\hat{q}_2 + \hat{q}_1/2)^2$, the HWET derived in the whole sample and only in cases can be written as, respectively

$$\text{HWET-CC} = \frac{\sqrt{rs/n}(\hat{\Delta}_1 - \hat{\Delta}_2)}{[1 - n_2/n - n_1/(2n)][n_2/n + n_1/(2n)]}$$

and

$$\text{HWET-C} = \frac{\sqrt{r}\hat{\Delta}_1}{[1 - n_2/n - n_1/(2n)][n_2/n + n_1/(2n)]},$$

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $n_1 = r_1 + s_1$, and $n_2 = r_2 + s_2$. Using the Hardy-Weinberg Equilibrium test to select the genetic model (song et al., 2006; Ng and Zheng, 2008; Zheng et al.,2016) can be summarized as follows: set a positive threshold c, for example, $c = 1.645$, the genetic model is determined as: if $Z > c$, the recessive model is selected; if $Z < -c$, the dominant model is determined; otherwise, we choose the additive model, Z can be HWET-CC or HWET-C. We let the estimate of θ be $\tilde{\theta}_{CC}$ and $\tilde{\theta}_C$ using HWET-CC and HWET-C, repectively. We find numerically that both $\tilde{\theta}_{CC}$ and $\tilde{\theta}_C$ might not be consistent to θ.

2.3 The proposed procedure

To develop a new method, we decompose the genotype data as

$$(G_1, G_2) = \begin{cases} (0,0), & \text{if } G = 0 \\ (1,0), & \text{if } G = \theta \\ (0,1), & \text{if } G = 1 \end{cases}$$

Then, the logistic regression model becomes as follows,

$$P(Y = 1|\mathbf{X}, G) = \frac{\exp(\alpha + \mathbf{X}^\top\gamma + G_1\beta_1 + G_2\beta_2)}{1 + \exp(\alpha + \mathbf{X}^\top\gamma + G_1\beta_1 + G_2\beta_2)}.$$

Under rare disease assumption with $\alpha << 0$, we can obtain that,

$$1 + \exp(\alpha + \mathbf{X}^\top\gamma) \simeq 1,$$

$$1 + \exp(\alpha + \mathbf{X}^\top\gamma + \theta\beta_1) \simeq 1,$$

and

$$1 + \exp(\alpha + \mathbf{X}^\top\gamma + \beta_2) \simeq 1.$$

By some algebras, we can obtain that,

$$P(Y = 1|\mathbf{X}, G = 0) = \frac{\exp(\alpha + \mathbf{X}^\top\gamma)}{1 + \exp(\alpha + \mathbf{X}^\top\gamma)} \simeq \exp(\alpha + \mathbf{X}^\top\gamma),$$

$$P(Y = 1|\mathbf{X}, G = \theta) = \frac{\exp(\alpha + \mathbf{X}^\top\gamma + \theta\beta_1)}{1 + \exp(\alpha + \mathbf{X}^\top\gamma + \theta\beta_1)} \simeq \exp(\alpha + \mathbf{X}^\top\gamma + \theta\beta_1),$$

and

$$P(Y = 1|\mathbf{X}, G = 1) = \frac{\exp(\alpha + \mathbf{X}^\top\gamma + \beta_2)}{1 + \exp(\alpha + \mathbf{X}^\top\gamma + \beta_2)} \simeq \exp(\alpha + \mathbf{X}^\top\gamma + \beta_2).$$

If real genetic model is additive satisfying to

$$\frac{P(Y = 1|\mathbf{X}, G = 1)}{P(Y = 1|\mathbf{X}, G = 0)} \simeq \left[\frac{P(Y = 1|\mathbf{X}, G = \theta)}{P(Y = 1|\mathbf{X}, G = 0)}\right]^2,$$

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

namely,

$$\frac{\exp(\alpha+\mathbf{X}^\top\gamma+\beta_2)}{\exp(\alpha+\mathbf{X}^\top\gamma)} \simeq \left[\frac{\exp(\alpha+\mathbf{X}^\top\gamma+\theta\beta_1)}{\exp(\alpha+\mathbf{X}^\top\gamma)}\right]^2.$$

By some algebras, we can obtain that, $\theta = \frac{\beta_2}{\beta_1} \simeq 0.5$. If real genetic model is dominant satisfying to

$$\frac{P(Y=1|\mathbf{X},G=1)}{P(Y=1|\mathbf{X},G=0)} \simeq \frac{P(Y=1|\mathbf{X},G=\theta)}{P(Y=1|\mathbf{X},G=0)},$$

namely,

$$\frac{\exp(\alpha+\mathbf{X}^\top\gamma+\beta_2)}{\exp(\alpha+\mathbf{X}^\top\gamma)} \simeq \frac{\exp(\alpha+\mathbf{X}^\top\gamma+\theta\beta_1)}{\exp(\alpha+\mathbf{X}^\top\gamma)}.$$

By some algebra, we can obtain that, $\theta = \frac{\beta_2}{\beta_1} \simeq 1$. On the basis of above derivation, we can see that, $\frac{\beta_2}{\beta_1}$ represent genetic inheritance in case-control studies under common recessive, dominant, and additive genetic model. So $\frac{\beta_2}{\beta_1}$ can measure genetic inheritance to some extent in case-control studies.

For $i = 1, \cdots, n$, denote observation $\mathbf{x}_i^\top$ for $\mathbf{X}^\top$, and (g_{i1}, g_{i1}) for G. The likelihood function is

$$L(\alpha,\beta_1,\beta_2) = \prod_{i=1}^{n}\left[\frac{\exp(\alpha+\mathbf{X}^\top\gamma+g_{i1}\beta_1+g_{i2}\beta_2)}{1+\exp(\alpha+\mathbf{X}^\top\gamma+g_{i1}\beta_1+g_{i2}\beta_2)}\right]^{y_i}\left[\frac{1}{1+\exp(\alpha+\mathbf{X}^\top\gamma+g_{i1}\beta_1+g_{i2}\beta_2)}\right]^{1-y_i}$$

and the log-likelihood function is

$$l(\alpha,\beta_1,\beta_2) = \sum_{i=1}^{n}\left\{\left[y_i(\alpha+\mathbf{x}_i^\top\gamma+g_{i1}\beta_1+g_{i2}\beta_2)\right] - \ln\left(1+\exp(\alpha+\mathbf{x}_i^\top\gamma+g_{i1}\beta_1+g_{i2}\beta_2)\right)\right\}$$

Then the estimate of β_1 and β_2 (denote them by $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively) can be obtained by solving the constrain optimization problem

$$(\hat{\beta}_1,\hat{\beta}_2) = \underset{\beta_1,\beta_2\geq 0,\beta_1\neq 0}{\operatorname{argmax}}\ l(\alpha,\beta_1,\beta_2)$$

So, the estimate of θ, denoted it by $\hat{\theta}$, is $\hat{\theta} = \hat{\beta}_2/\hat{\beta}_1$. Since $(\hat{\beta}_1,\hat{\beta}_2)^\top$ is the consistent estimate of $(\beta_1,\beta_2)^\top$ under a general logistic regression setting, so the $\hat{\theta}$ is consistent to θ. Or, we can obtain estimate of θ directly based on following log-likelihood function,

$$\tilde{l}(\alpha,\beta_1,\theta) = \sum_{i=1}^{n}\left\{\left[y_i(\alpha+\mathbf{x}_i^\top\gamma+g_{i1}\beta_1+g_{i2}\theta\beta_1)\right] - \ln\left(1+\exp(\alpha+\mathbf{x}_i^\top\gamma+g_{i1}\beta_1+g_{i2}\theta\beta_1)\right)\right\}$$

and following constrain optimization problem,

$$(\hat{\beta}_1,\hat{\theta}) = \underset{\theta\geq 0,\beta_1\neq 0}{\operatorname{argmax}}\ \tilde{l}(\alpha,\beta_1,\theta)$$

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 Results

3.1 Simulation Results

To illustrate the performance of the proposed method (denote it by PROPOSED), we conduct simulation studies and compare the mean squared error (MSE) with the existing HWET-CC (Ng and Zheng, 2008) and HWET-C (Zheng et al., 2016) selection procedures. We assume the Hardy-Weinberg equilibrium holds in the general population, that is, the genotype frequencies satisfy to $P(aa) = (1-p)^2$, $P(Aa) = 2p(1-p)$ and $P(AA) = p^2$, where $p = P(A)$ and p is chosen from $\{0.05, 0.15, 0.30, 0.45\}$. Consider two disease prevalences $K = 0.02$ and 0.05, which results in $\alpha = \ln(K/(1-K)) = -3.89$ and -2.94, respectively, when $\gamma = \mathbf{0}_m$ and $\beta = 0$, where $\mathbf{0}_m$ is a m-dimensional vector with all the elements being 0. For a fixed α, we let $m = 1$ and assume that $\mathbf{X}$, independent of G, follows the standard normal distribution, $\gamma = \ln 1.1$ and $\beta = \ln 1.5$. $r = s \in \{500, 1000, 1500, 2000\}$ and θ is chosen from $\{0, 0.25, 0.5, 0.75, 1\}$. 1000 replicates are conducted to calculate the empirical mean squared error (MSE).

Figures 1 and 2 shows the empirical MSEs of the HWET-CC, HWET-C and PROPOSED for $K = 0.05$. It can be shown that the MSEs of the PROPOSED are decreasing with the sample size increasing. For example, when $r = s = 500$ and $p = 0.15$ and the genetic model is recessive, the empirical MSEs of the PROPOSED is 0.18 and that is 0.11 for $r = s = 1000$. The empirical MSEs of the HWET-CC and HWET-C under the additive model is almost unchanged for different sample sizes. For example, when $r = s = 500$ and $p = 0.30$, the empirical MSEs of the HWET-CC and HWET-C are 0.025 and 0.066, respectively, and those are 0.028 and 0.064, respectively, for $r = s = 1500$. As expected, when the sample size is large, the MSEs of the PROPOSED is smaller than those of HWET-CC and HWET-C under the dominant model. For example, when the sample size $r = s = 800$ and $p = 0.45$, the empirical MSEs of the PROPOSED, HWET-CC and HWET-C are 0.032, 0.094 and 0.048, respectively.

Figures 3 and 4 shows the empirical MSEs of the HWET-CC, HWET-C and PROPOSED for $K = 0.02$. We have similar finds with $K = 0.05$. For instance, when the genetic model is dominant and $p = 0.30$, when $r = s = 1000$ and $r = s = 1500$, the empirical MSEs of the PROPOSED is 0.081 and 0.040, respectively, and it can be shown that the MSEs of the PROPOSED are decreasing with the sample size increasing. The empirical MSEs of the HWET-CC and HWET-C under the additive model is almost unchanged for different sample sizes. For instance, when $r = s = 500$ and $p = 0.15$, the empirical MSEs of the HWET-CC and HWET-C are 0.023 and 0.060, respectively, and those are 0.022 and 0.059, respectively, for $r = s = 2000$. Again, as expected, the MSEs of the PROPOSED is smaller than those of HWET-CC and HWET-C when the sample size is large under the

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

recessive model. For example, when the sample size $r = s = 1000$ and $p = 0.45$, the empirical MSEs of the PROPOSED, HWET-CC and HWET-C are 0.027, 0.065 and 0.035, respectively.

3.2 Two Real Applications

Breast cancer is the common cancer for women. Almost 15 percent of women with breast cancer have family members diagnosed with it, which means that the genetic variants might confer some risk of developing breast cancer. Hunter et al. (2007) conducted a genome-wide association study and have identified 6 SNPs including rs10510126, rs12505080, rs17157903, rs1219648, rs7696175, and rs2420946, associated with breast cancer. The genotype value summaries are shown in Table 1. Type 2 diabetes is a lifelong disease. Typically, the genetic factor confer risk to the type 2 diabetes. Sladek et al. (2007) conducted a genome-wide association study and identified 8 SNPS associated with Type 2 diabetes. The summarized data are given also in Table 1.

We apply the HWET-CC and HWET-C and PROPOSED to these 14 SNPs to search for their genetic models. The results are given in Table 1. We find that, for the breast cancer, half of them are dominant model and the others are dominant models if the HWET-C or HWET-CC is used. Using the PROPOSED can give the detailed value of the genetic model. For example, for SNP rs10510126, the estimated genetic model is $\hat{\theta} = 2.36 \times 10^{-11}$, which is recessive model, while using HWET-C gives the additive model. Otherwise, the PROPOSED can give some other model beyond three commonly employed three genetic models. For instance, for SNP rs2420946, using the PROPOSED give a genetic model of 0.38.

4 Discussion

Case-control genetic association study has been proved to be an efficient tool to identify the deleterious variants by scanning the human genome. There are several genetic variants including insertion, deletion, copy number variation, and SNP etc. Among them, the SNP is the most common one. There are 2.96 billion base pairs in human genome and the number of SNPs is about 30 million. By now, more than ten thousand of SNPs have been identified to be associated with hundreds of diseases or traits. To evaluate the significance of a SNP, one has to specify a genetic model. Among three genetic models including recessive, additive and dominant model, the additive model is more frequently to be assumed when conducting an association study. However, in practice, the real genetic model is impossible to know. Especially the causal SNP cannot be genotyped and the SNP locus is its surrogate. Thus the genetic model between the causal SNP and surrogate might be different (Hormozdiari et al., 2015). Misspecifying a genetic model might result

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in a loss of statistical power.

In this work, we use θ to denote the genetic model and θ varies range from 0 to 1. The existing work only make the inference for $\theta = 0$, 0.5 and 1. It cannot estimate other values of θ. By decomposing the genotype score, we proposed a new procedure to estimate θ, which is shown to have consistency, while the existing procedures are not consistent based on the numerical results. On the hand hand, we obtained the consistent estimate of the genetic model, the next step should be construct the association test based on the chosen model. There is existing the correlation between choosing the genetic model and the association test, which is the future topic.

Acknowledgements

Research of J. Zhu is partially supported by seeding project funding(No. 2019ZZX026), scientific research project funding of talent recruitment, and start up funding for scientific research of Hubei University of Chinese Medicine. Research of Z. Li is partially supported by the self-determined research funds of Central China Normal University(CCNU) from the colleges basic research of MOE (CCNU18QN031). Research of Q. Li is partially supported by Beijing Natural Science Foundation (Z180006) and National Nature Science Foundation of China (11722113). We thank L. Xun from Changchun University of Technology in China for insightful discussions and suggestions.

References

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson,S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A. et al., 2007 A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics 39:870-874. doi: 10.1038/ng2075

Hu X. N, Duan X. G, Pan D. D, Zhang S. G, Li Q, 2017 A model-embedded trend test with incorporating Hary-Weinberg equilibrium information. Journal of Systems Science and Complexity 30:101-110. doi:10.1007/s11424-017-6187-4

Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E., 2015 Identification of causal genes for complex traits. Bioinformatics 31:i206-i213.doi: 10.1093/bioinformatics/btv240

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S.,Haynes,C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T.et al., 2005 Complement factor H polymorphism in aged-related macular degeneration. Science 308:385-389. doi: 10.1126/science.1109557

Moltke I., et al., 2014 A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature 512(7513),190-193. doi: 10.1038/na-

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ture13425

Nik-Zainal, S., et al., 2016 Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534(7605),47-54. doi: 10.1038/nature17676

Prentice, R. L. and Pyke, R., 1979 Logistic disease incidence models and case-control studies. Biometrika 66:403-411

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S. et al., 2007 A genome-wide association study identifies novel risk loci for type2 diabetes. Nature 445: 881-885. doi:10.1038/nature05616

Song, K., and Elston, R. C., 2006 A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Statistics in Medicine 25:105C126. doi: 10.1002/sim.2350

The Wellcome Trust Case Control Consortium, 2007 Genomewide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature 447, 661-678.

Zheng G, Li Q, Yuan A, 2014 Some statistical properties of efficiency robust tests for genetic studies. Scandinavian Journal of Statistics 41:762-774. doi: 10.1111/sjos.12060

Zheng G. and Ng H. K. T., 2008 Genetic model selection in two-phase analysis for case-control association studies. Biostatistics 9(3): 391-399. doi:10.1093/biostatistics/kxm039

Zheng G., Zhang W., Xu J., et al., 2016 Genetic risks and genetic model specification. Journal of Theoretical Biology 403:68-74. doi: 10.1016/j.jtbi.2016.05.016

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. The estimated genetic models for 6 SNPs associated with breast cancer and 8 SNPs associated with Type 2 diabetes.

	r_0	r_1	r_2	s_0	s_1	s_2	$\hat{\theta}$	$\tilde{\theta}_C$	$\tilde{\theta}_{CC}$
6 SNPs associated with breast cancer									
rs10510126	955	180	10	854	272	14	2.36×10^{-11}	0.5	0.5
rs12505080	608	477	50	628	408	99	0.99	1	1
rs17157903	777	316	18	862	220	26	0.99	1	1
rs1219648	352	543	250	433	538	170	0.36	0.5	0.5
rs7696175	353	605	187	396	496	249	0.99	1	1
rs2420946	357	546	242	440	537	165	0.38	0.5	0.5
8 SNPs associated with Type 2 diabetes.									
rs7903146	197	348	149	335	254	65	0.62	0.5	0.5
rs13266634	54	229	411	53	293	307	1.62×10^{-8}	0	0
rs1111875	77	302	315	119	308	227	0.54	0.5	0.5
rs7923837	66	300	328	116	296	242	0.66	0.5	0.5
rs7480010	301	327	66	363	246	353	0.82	1	0.5
rs3740878	25	273	386	65	249	353	1	1	1
rs11037909	25	274	387	65	251	353	0.99	1	1
rs1113132	25	271	390	63	251	355	0.98	1	1

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

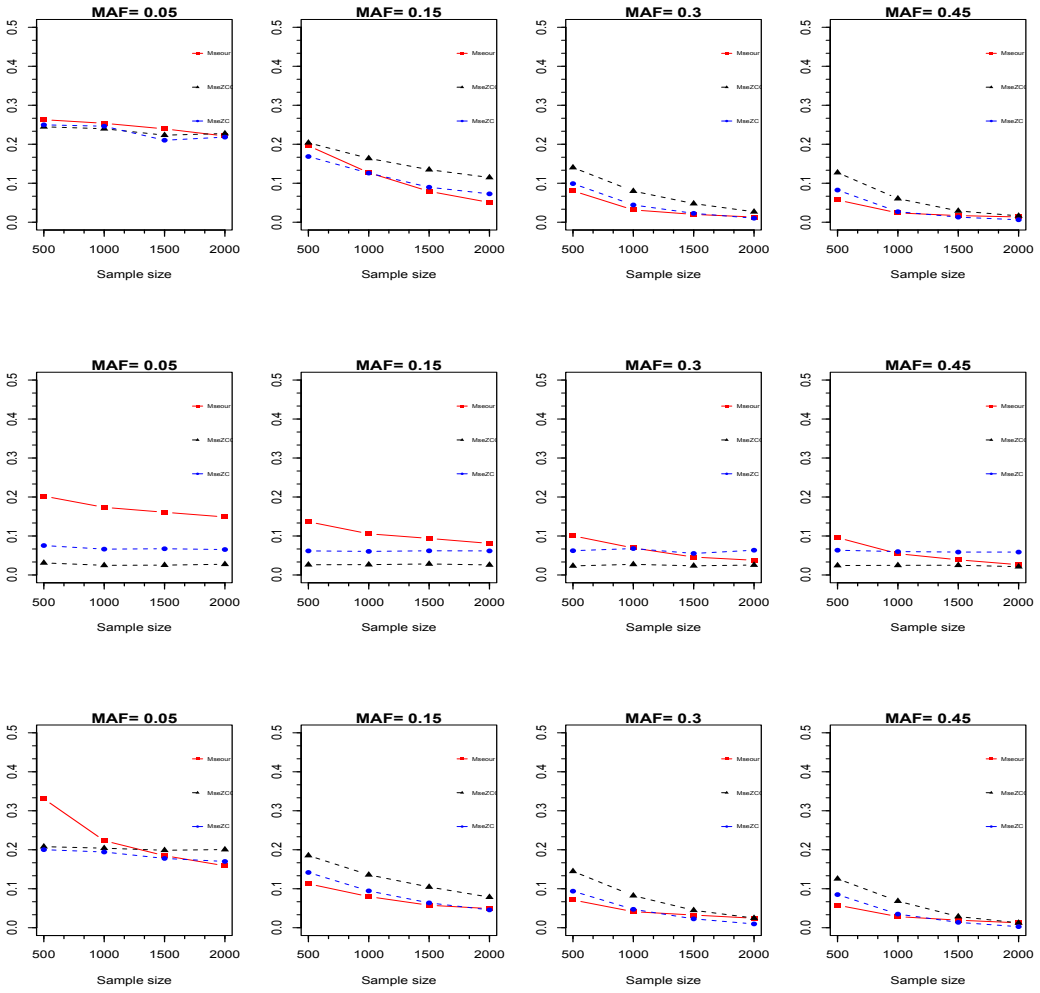

Figure 1: The MSEs of the PROPOSED (square), HWET-CC (triangle) and HWET-C (dot) when $k = 0.05$, $p \in \{0.05, 0.15, 0.30, 0.45\}$ and $\theta \in \{0, 0.5, 1\}$, where the first row is for $\theta = 0$, the second row is for $\theta = 0.5$ and the third row is for $\theta = 1$. The horizontal axis is the sample size, and the vertical axis is the value of MSEs.

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

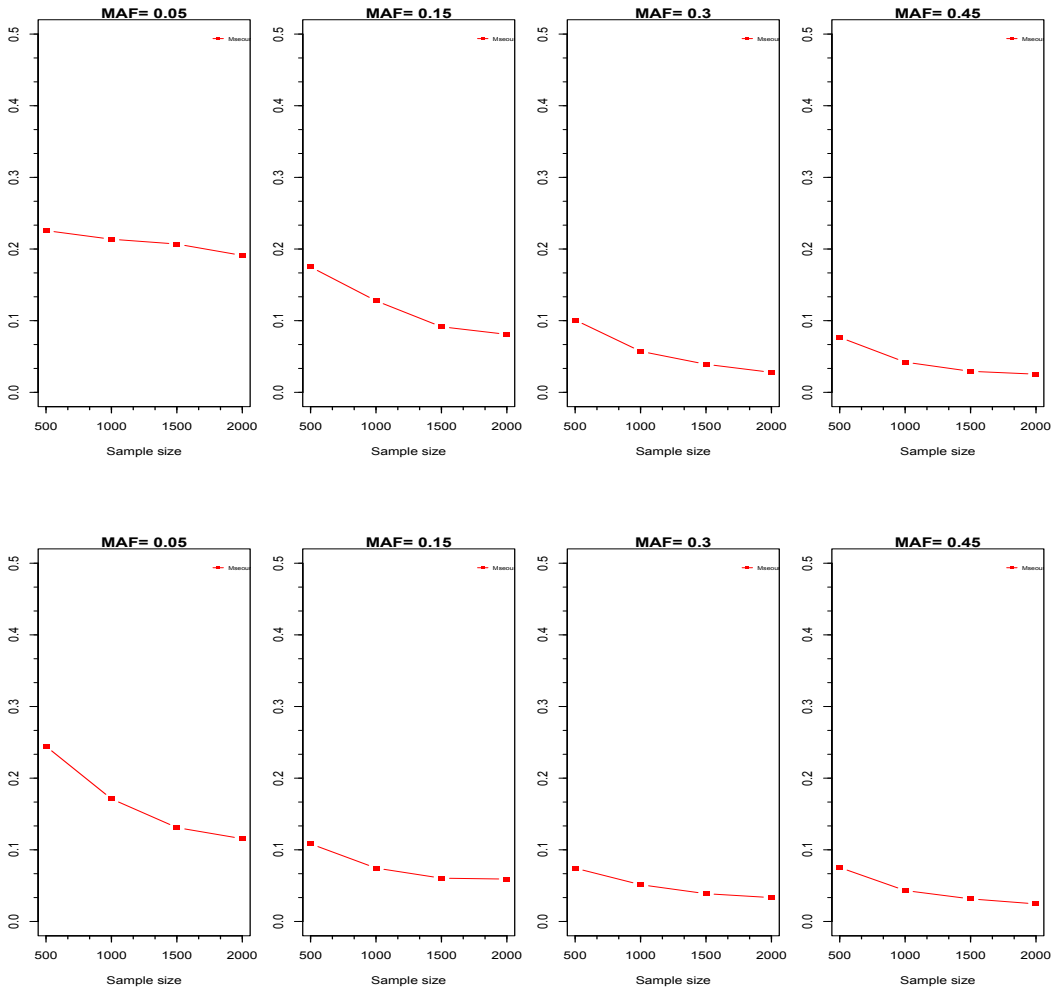

Figure 2: The MSEs of the PROPOSED when $k = 0.05$, $p \in \{0.05, 0.15, 0.30, 0.45\}$ and $\theta \in \{0.25, 0.75\}$, where the first row is for $\theta = 0.25$ and the second row is for $\theta = 0.75$. The horizontal axis is the sample size, and the vertical axis is the value of MSEs.

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

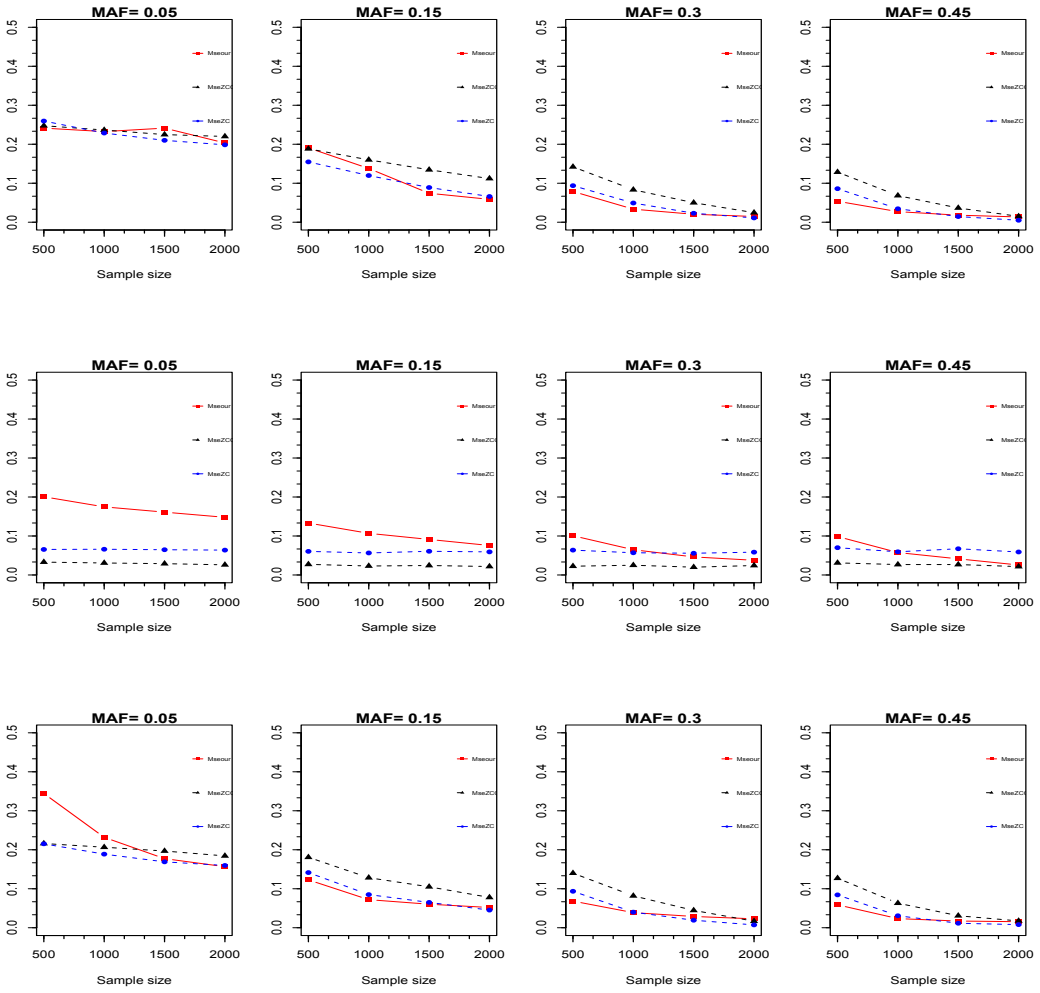

Figure 3: The MSEs of the PROPOSED (square), HWET-CC (triangle) and HWET-C (dot) when $k = 0.02$, $p \in \{0.05, 0.15, 0.30, 0.45\}$ and $\theta \in \{0, 0.5, 1\}$, where the first row is for $\theta = 0$, the second row is for $\theta = 0.5$ and the third row is for $\theta = 1$. The horizontal axis is the sample size, and the vertical axis is the value of MSEs.

bioRxiv preprint doi: https://doi.org/10.1101/847269; this version posted December 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

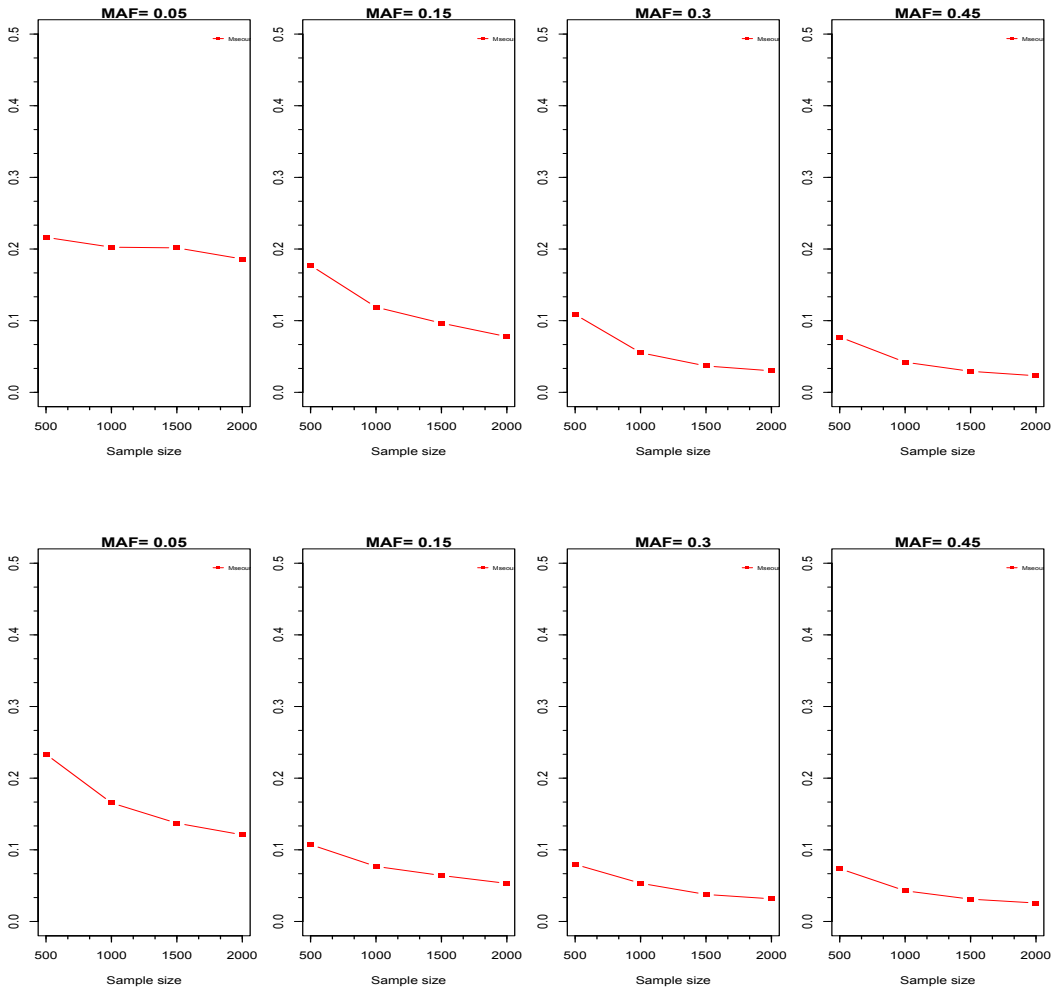

Figure 4: The MSEs of the PROPOSED when $k = 0.02$, $p \in \{0.05, 0.15, 0.30, 0.45\}$ and $\theta \in \{0.25, 0.75\}$, where the first row is for $\theta = 0.25$ and the second row is for $\theta = 0.75$. The horizontal axis is the sample size, and the vertical axis is the value of MSEs.