Generalized Hierarchical Mixed Model Association Analysis

Runqing Yang[1, 2], Yuxin Song[2, 3], Zhiyu Hao[1] and Zhonghua Liu[1]

1. College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

2. Research Centre for Aquatic Biotechnology, Chinese Academy of Fishery Sciences, Beijing 100141, People's Republic of China

3. Wuxi Fisheries College, Nanjing Agricultural University, Wuxi 214081, China

Corresponding Authors: Runqing Yang (runqingyang@cafs.ac.cn)
Zhonghua Liu (liuzhonghua@neau.edu.cn)

Abstract

In genome-wide association analysis for complex diseases, we partitioned the genomic generalized linear mixed model (GLMM) into two hierarchies—the GLMM regarding genomic breeding values (GBVs) and a generalized linear regression of the GBVs to the tested marker effects. In the first hierarchy, the GBVs were predicted by solving for the genomic best linear unbiased prediction for GLMM, and in the second hierarchy, association tests were performed using the generalized least square (GLS) method. The so-called Hi-GLMM method exhibited advantages over existing methods in terms of both genomic control for complex population structure and statistical power to detect quantitative trait nucleotides (QTNs), especially when the GBVs were estimated precisely, and using joint association analysis for QTN candidates obtained from a test at once.

Introduction

Using random polygenic effects excluding the tested marker to correct confounders such as population stratification and cryptic relatedness, the linear mixed model (LMM) can efficiently control the false positive rate and improve the power to detect quantitative trait nucleotides (QTNs). However, because of the requirement of the high computing intensity for LMM, researchers were motivated to develop simplified algorithms [1-8]. LMM is a widely used and powerful approach for genome-wide association studies (GWAS). In general, LMM, which assumes the phenotype is distributed normally, is appropriate for continuous quantitative traits [9]. For complex

disease traits expressed as a binary phenotype, however, the genome-wide mixed model association does not provide interpretable and predictable mapping results [10].

Because of the quantitative traits, complex diseases are thought to be controlled by a number of loci each having a small effect on the phenotype [11,12]. Instead of the linear regression model, logistic regression in the generalized linear model (GLM) [13,14] has also been applied to analyze the association between markers of interest in binary disease phenotypes. Despite the correction for fixed-effect covariates [15-17], logistic regression still produces inflation of association test statistics. Therefore, the generalized linear mixed model (GLMM) [18] was proposed which considers random polygenic effects to increase the power to map QTNs for disease traits. However, genome-wide GLMM association requires much more computation time than the mixed model association with either restricted maximum likelihood estimation (REML) [19] or Markov chain Monte Carlo (MCMC) iteration [20]. In addition, when the maximum likelihood for estimation and approximations are used to avoid numerical integration, GLMM leads to serious bias induced by the approximations [21], and solutions tend to the positive/negative infinity.

In ascertained case-control studies when the proportions of cases and controls are not a random sample from the population, GLMM gives biased estimates of genomic hereditability for disease traits [22] and therefore suffers from a loss in power to detect QTNs. Based on the calibrated genomic hereditability for case-control ascertainment, a Chi-squared score statistic for GWAS of disease traits was computed from posterior mean liabilities (PMLs) under the liability-threshold model [23]. To simplify the GLMM-based association analysis, the GMMAT [24] and SAIGE [25] separately extend the EMMAX [7] and BOLT-LMM [8] from normally distributed traits to binary diseases. In this study, we partitioned the genomic GLMM into two hierarchies by genomic breeding values (GBVs) for liability. The GBVs were predicted by solving for the genomic best linear unbiased prediction (GBLUP) for GLMM in the first hierarchy, and in the second hierarchy, association tests were performed by generalized linear regression of the correlated GBVs to the SNP effect tested.

Results

Statistical properties of the Hi-GLMM

For the two genomic datasets, we simulated 40, 200, and 1000 QTNs at the low (0.2), moderate (0.5), and high (0.8) genomic heritability, respectively, to control the

phenotypes. The Hi-GLMM, a test at once was comparable with the four competing methods—GRAMMAR, GMMAT, LTMLM, and SAIGE. The Q-Q and ROC profiles are displayed in Figure 1 and Figure 2 for the simulated phenotypes controlled by 200 QTNs, respectively, and are provided in Figure 1S and Figure 2S for all the simulated phenotypes. We also recorded genomic control values for each method in Table 1S. Under ideal genomic control very close to 1, Hi-GLMM behaves stably and exhibits a slightly higher statistical power than the GMMAT which approximates the exact GLMM, irrespective of how much QTNs control the quantitative traits and how complex the population structures are. Genomic controls are the lowest among Hi-GLMM and the four competing methods, and the population structure is more complex, while the false negative rate is larger for GRAMMAR. For the simulated phenotypes in maize, LTMLM was found to be superior to the other methods in terms of statistical power, but it produced strong false positive errors. Similarly, SAIGE also exhibited higher statistical power with false positive errors for those controlled by 1,000 QTNs at the genomic heritability of 0.2. For the human database, there was no distinct difference in statistical properties between Hi-GLMM and the four competing methods. GRAMMAR yielded some false negative errors.

Furthermore, we jointly analyzed multiple QTN candidates chosen from the Hi-GLMM, a test at once at the significance level of 0.05. For convenience of comparison, the statistical powers obtained with joint association analyses were depicted along with those using a test at once. Through backward multiple regression analysis, joint association analyses exhibited significantly the improved statistical power in almost the same genomic control as a test at once. Even though it provided the highest false positive rates, LTMLM, a test at once was inferior to Hi-GLMM, joint analysis in the terms of statistical power.

Sensitivity to estimate genomic heritability or GBVs

When genomic heritability or GBVs were precisely estimated, the Hi-LMM achieved much higher statistical power to detect QTNs, than EMMAX and BOLT-LMM. To test whether this finding fits Hi-GLMM, GMMAT, and SAIGE, which are extensions of Hi-LMM, EMMAX, and BOLT-LMM, respectively, for binary traits, we analyzed the simulated phenotypes controlled by 200 QTNs at the different heritabilities by taking the simulated genomic heritability or GBVs as the estimated polygenic effects. As shown in Figure 3 and Figure 3S, for the Hi-GLMM, one test at a time exhibited

higher statistical power with the more ideal genomic controls than joint association analysis, if genomic heritability or breeding values were full estimated accurately. In contrast, GMMAT showed a somewhat reduced both statistical power and genomic control. Especially, SAIGE did not find any QTNs that were removed from the residual phenotypes. In addition, we adopted a Lasso technique implemented in R/glmnet (Friedman, et al., 2010) to precisely estimate the GBVs and found that Hi-LMM produced higher statistical power than GMMAT using the improved the GBVs. This suggested that Hi-GLMM can improve statistical power by more accurately estimating genomic heritability and GBVs, which is in contrast to GMMAT and SAIGE.

Calculation of GRMs with the sampling markers

Using sampling markers to estimate the GRMs, we improved computational efficiency of the Hi-LMM for normally distributed traits. To test whether this finding fits the binary traits, we randomly took 3 K, 5 K, 10 K, 20 K, and 25 K SNPs from the entire genomic markers and analyzed the phenotypes controlled by different numbers of QTNs at the heritability of 0.5 with all the methods, except LTMLM in which SNPs can be not selected to estimate heritability. Figure 4 shows changes in the genomic controls with varied sampling levels of SNPs. Hi-GLMM gradually controlled positive false errors and improved statistical power to detect QTNs (Figure 4S and Figure 5S) with the sampling markers increased, similar to the result of GMMAT. In addition, Hi-GLMM needed more sampling markers than GMMAT to yield the ideal genomic control and statistical power using all genomic markers, but GMMAT often made genomic control unstable at larger sampling markers. In contrast, SAIGE produced serious false negative errors in the complex maize population irrespective of how many SNPs were used.

Application to WTCCC study

We were authorized to re-analyze the data obtained from the Wellcome trust case control consortium (WTCCC) study 1 [26]. There were the 11,985 cases from six common diseases and 3,004 shared controls, genotyped at a total of 490,032 SNPs. For each dataset comprising one type of case and all shared controls, a standard quality control (QC) procedure was performed: SNPs with MAFs < 0.01 and HWE > 0.05 were excluded, and individuals with missing rates > 0.01 were also excluded. After the QC process, the number of samples and SNPs used for the generalized

mixed model association analyses were 5002 individuals (1998 cases and 3004 controls) and 409,642 SNPs for bipolar disorder (BD), 4992 individuals (1988 cases and 3004 controls) and 409,516 SNPs for coronary artery disease (CAD), 5003 individuals (1999 cases and 3004 controls) and 409,924 SNPs for rheumatoid arthritis (RA), 5005 individuals (2001 cases and 3004 controls) and 409,742 SNPs for hypertension (HT), 5004 individuals (2000 cases and 3004 controls) and 40,9674 SNPs for type I diabetes (T1D), and 5003 individuals (1999 cases and 3004 controls) and 409,805 SNPs for type II diabetes (T2D). All data analyses were performed in a CentOS Linux sever with 2.60 GHz Intel(R) Xeon(R) 40 CPUs E5-2660 v3, and 512 GB memory.

For the six common diseases, we implemented Hi in the two ways: to estimate genomic heritability and GBVs together using the entire genomic marker set and to estimate only GBVs by randomly sampling 5,000 SNPs given a heritability of 0.5. The Q-Q and Manhattan profiles for the six common diseases are shown in Figure 6S and Figure 7S which indicate results obtained by Hi-GLMM and the four competing methods used in simulations. We concluded that (1) under perfect genomic control, Hi-GLMM found the QTNs for each disease on each chromosome, and the numbers of the detected QTNs were not less than those obtained by all the competing methods; and (2) in Hi-GLMM, joint association analyses detected more QTNs than one test at once. As compared to GRAMMAR-Lambda, GRAMMAR detected less QTNs with the lowest genomic control among all the methods, while GMMAT yielded more SNPs whose -log(p) exceeded the Bonferroni corrected thresholds for CAD, T1D, T2D, and HT, but it produced the highest genomic control. Additionally, LTMLM estimated abnormal genomic heritabilities for CAD, BD, T2D and HT, producing unstable genomic control.

Next, we conducted strict QC for each dataset, as performed earlier in [22], for estimating the genomic heritability. Despite this, the missing heritabilities could not be normally estimated for BD and HT. As shown in Figure 8S and Figure 9S, all methods provided clear and comparable association results, except for GRAMMAR. Interestingly, both LTMLM and GMMAT seriously underestimated genomic heritability for each disease after strict QC. In summary, Hi-GLMM could efficiently and robustly map QTNs for binary diseases, and did not depend on the estimation of genomic heritability and QC for genomic datasets. For each dataset with standard QC, we recorded the running times from input of genotypes and phenotypes to the output

of mapped QTNs for all methods. Table 2S showed that Hi-GLMM reduced the computing time by dozens of times with the lowest memory footprint among the competing methods.

Discussion

The GLMM formulates the linear relationship between binary response variables and the normal distributed predictors for liability through link function [13,14], so the GBVs are estimated as a normal variable by the GBLUP equations for GLMM [18]. In association tests in the second hierarchy, Hi-GLMM can statistically infer the QTNs for complex disease with the GLS as continuous quantitative traits, once the GBVs are successfully obtained with the GLMM in the first hierarchy. In general, linear models have a between computing efficiency than generalized linear models for samples with the same independent variables; thus, while genomic heritability and GBVs can be estimated, the Hi-GLMM will achieve genome-wide GLMM-based association results faster as compared to the GMMAT and SAIGE when the generalized linear model is applied [24,25].

Similar to Hi-GLMM, LTMLM first estimated PMLs rather than the GBVs by Bayesian sampling under the liability-threshold model [23], and then, statistically inferred the QTNs with the LMM for the PMLs. However, both Bayesian sampling and LMM association analysis shows showed that the LTMLM had a high computing cost and serious false positive errors under complex structured populations, as shown in simulations. Like the GMMAT and SAIGE, the Hi-GLMM had to estimate the genomic heritability and GBVs using genomic markers. In contrast, the famers over-estimated polygenic heritabilities and effects by genomic heritability and GBVs, potentially increasing false negative rates, while the later unbiasedly and best estimated polygenic effects with generalized linear regression at the second hierarchy, ensuring high statistical power to detect QTNs in good genomic control.

Genome-wide GLMM association analysis is being devoted to improve statistical power and handle a large-scale population. Within the framework of Hi-GLMM, precise estimation for GBVs will contribute to attain high statistical power to detect QTNs, but this depends on the development of genomic selection for binary traits [27]. Furthermore, joint association analysis for multiple QTN candidates obtained with a test at once significantly improved the statistical power because possible linkage disequilibrium among candidate markers is considered in stepwise regression.

To speed up the Hi-GLMM, the GRM can be re-constructed with markers selected randomly across the whole genome [28]. Theoretically, the GRM may be better estimated with less and relatively independent markers separately drawn from all linkage groups than those randomly drawn from the whole genome, but construction of linkage groups may additionally increase the computational cost. In addition, transformation of human GRM to be the sparse [29] can also greatly simplify the Hi-GLMM to analyze a large-scale population.

Code availability

Hi-GLMM software can be available at https://github.com/RunKingProgram/Hi-GLMM.

Acknowledgements

The research is financially supported by the National Natural Science Foundations of China (32072726) and the Special Scientific Research Funds for Central Non-profit Institutes, Chinese Academy of Fishery Sciences (2019A001).

References

1 Aulchenko, Y. S., de Koning, D. J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577-585 (2007).

2 Kang, H. M. *et al.* Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709-1723, doi:10.1534/genetics.107.080101 (2008).

3 Zhang, Z. *et al*. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355, doi:10.1038/ng.546

4 Lippert, C. *et al*. FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833-835 (2011).

5 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-824, doi:10.1038/ng.2310 (2012).

6 Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., Duijn, C. M., Van & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics* **44**, 1166-1170 (2012).

7 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-354, doi:10.1038/ng.548 (2010).

8 Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284-290, doi:10.1038/ng.3190 (2015).

9 Henderson, C. R. Applications of linear models in animal breeding. *Applications of Linear Models in Animal Breeding* (1984).

10 Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in

the application of mixed-model association methods. *Nature Genetics* **46**, 100-106 (2014).

11 Bulmer, M. G. The Effect of Selection on Genetic Variability. *American Naturalist* **105**, 201-211 (1971).

12 Falconer, D. S. *Introduction to Quantitative Genetic*. 2ed edn, (Longman, 1981).

13 Wedderburn, R. W. M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439-447 (1974).

14 McCullagh, P. & Nelder, J. A. *Generalized linear models, 2nd ed.*, (Chapman and Hall, 1989).

15 Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet* **8**, e1003032, doi:10.1371/journal.pgen.1003032 (2012).

16 Zaitlen, N. *et al.* Analysis of case-control association studies with known risk variants. *Bioinformatics* **28**, 1729-1737, doi:10.1093/bioinformatics/bts259 (2012).

17 Joel, M. & Witte, J. S. The Covariate's Dilemma. *Plos Genetics* **8**, e1003096 (2012).

18 Breslow, N. E. & Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9-25 (1993).

19 Schall, R. Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727 (1991).

20 Sorenrsen, D. & Gianola, D. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. (2002).

21 Gilmour, A. R., Anderson, R. D. & Rae, A. L. The Analysis of Binomial Data by a Generalized Linear Mixed Model. *Biometrika* **72**, 593-599 (1985).

22 Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics* **88**, 294-305, doi:10.1016/j.ajhg.2011.02.002 (2011).

23 Hayeck, T. *et al.* Mixed Model with Correction for Case-Control Ascertainment Increases Association Power. *American Journal of Human Genetics* **96**, 720 (2015).

24 Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *American journal of human genetics* **98**, 653-666, doi:10.1016/j.ajhg.2016.02.012 (2016).

25 Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics* **50**, 1335-1341, doi:10.1038/s41588-018-0184-y (2018).

26 Consortium, W. T. C. C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).

27 Gianola, D. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics* **194**, 573-596, doi:10.1534/genetics.113.151753 (2013).

28 Jennifer, L. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525-526 (2012).

29 Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics* **51**, 1749-1755, doi:10.1038/s41588-019-0530-8 (2019).

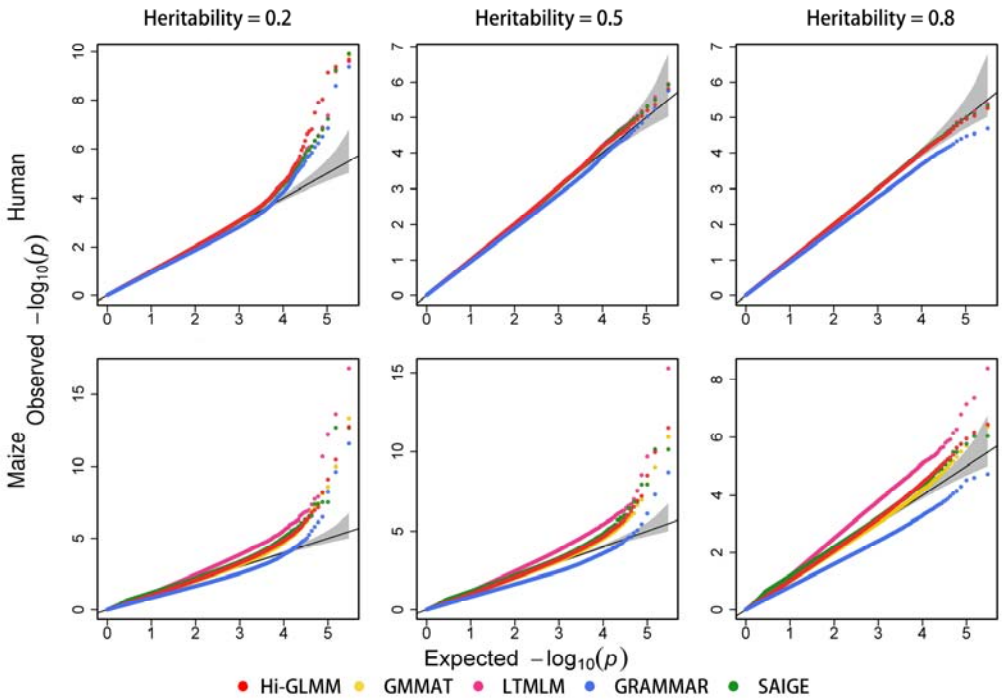

Figure 1: Comparison of Hi-GLMM with the four competing methods in the Q-Q profiles. The simulated phenotypes are controlled by 200 QTNs with the low, moderate and high heritabilities in human and maize. The Q-Q profiles for all simulated phenotypes are reported in Supplementary Figure 1S.

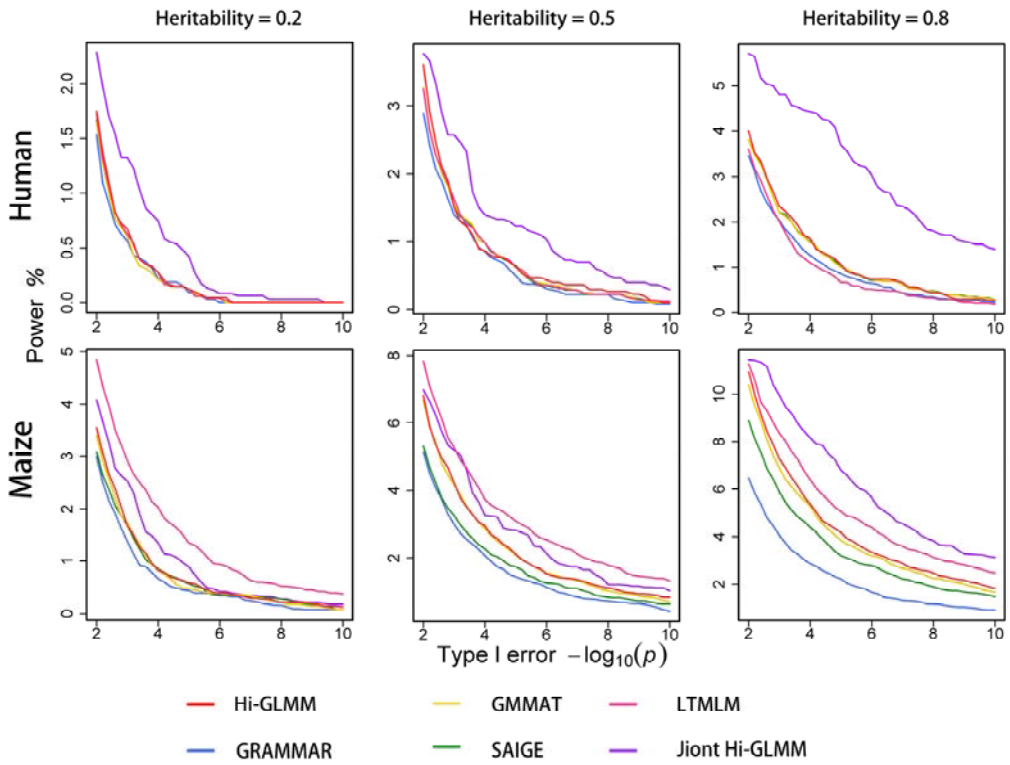

Figure 2: Comparison of Hi-GLMM with the four competing methods in the ROC profiles. The ROC profiles are plotted using the statistical powers to detect QTNs relative to the given series of Type I errors. Here, the simulated phenotypes are controlled by 200 QTNs with the low, moderate and high heritabilities in human and maize. The ROC profiles for all simulated phenotypes are reported in Supplementary Figure 2S.

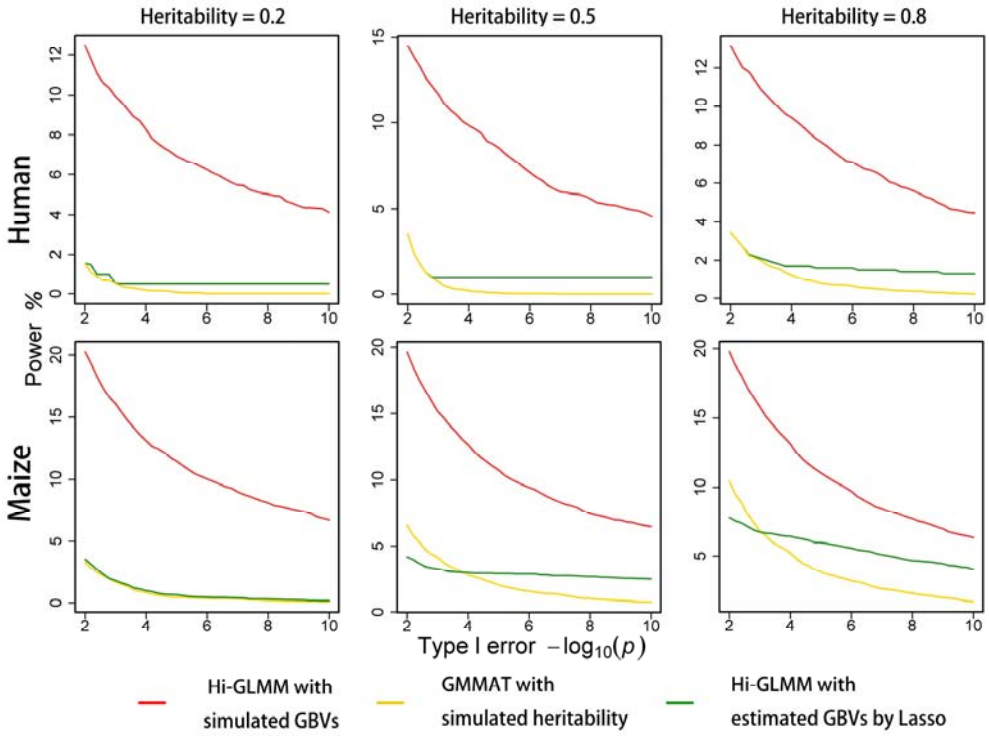

Figure 3: Sensitivity of statistical powers to estimate heritabilities or GBVs for Hi-GLMM. Statistical powers are dynamically evaluated with the ROC profiles., Both GRAMMAR and SAIGE do not detect any QTN with the simulated GBVs. The simulated phenotypes are controlled by 200 QTNs with the low, moderate and high heritabilities in human and maize.

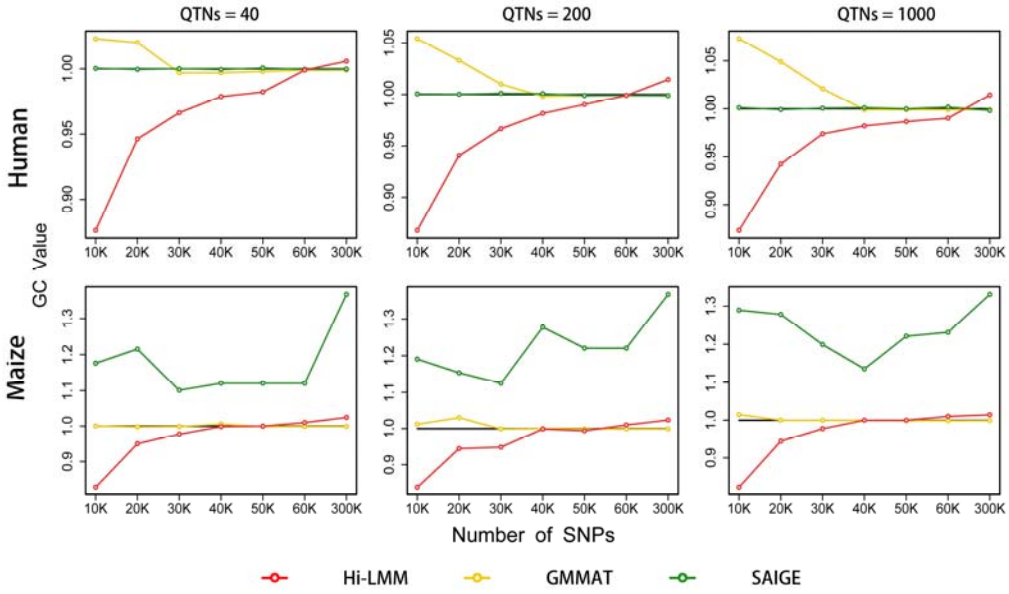

Figure 4: Changes in genomic controls with the number of sampling SNPs for Hi-GLMM and the five competing methods. Genomic control is calculated by averaging genome-wide test statistics. The simulated phenotypes are controlled by 40, 200 and 1000 QTNs with the moderate heritability in human and maize.

Online Methods

For binary disease traits, a logit regression model based on binomial distribution defines the linear relationship between phenotypes of the traits and the genetic effect of the SNPs tested. To reduce the false positive rate for mapping QTNs, polygenic effects as the confounding variables are considered as additional predictors. Thus, the genomic logit mixed model [1] was constructed as follows:

$$\ln\left(\frac{\boldsymbol{\mu}}{\mathbf{1}-\boldsymbol{\mu}}\right)=\mathbf{X}\boldsymbol{\beta}+\mathbf{u} \tag{1}$$

where $\boldsymbol{\mu}$ is the mean of binary disease traits, $\boldsymbol{\beta}$ is a vector of fixed effects including genetic effect of the tested SNPs; $\mathbf{X}$ is the incidence matrix for $\boldsymbol{\beta}$; and $\mathbf{u}$ is a vector of n random polygenic effects excluding the SNP tested, which is assumed to have normal distribution $N(\mathbf{0},\mathbf{K}\sigma_u^2)$ with a genomic relationship matrix (GRM) $\mathbf{K}$ [2] which is generally calculated from the entire genomic markers and an unknown polygenic variance σ_u^2.

Using $\mathbf{X}_{\mathrm{SNP}}\boldsymbol{\beta}_{\mathbf{SNP}}$ to denote regression terms of the tested SNP, we define GBVs as

$$\mathbf{g}=\mathbf{X}_{\mathrm{SNP}}\boldsymbol{\beta}_{\mathrm{SNP}}+\mathbf{u} \tag{2}$$

Then, model (1) is divided into two hierarchies:

$$\begin{cases}\ln\left(\frac{\boldsymbol{\mu}}{\mathbf{1}-\boldsymbol{\mu}}\right)=\mathbf{X}_{\text{-SNP}}\boldsymbol{\beta}_{\text{-SNP}}+\mathbf{g}\\ \mathbf{g}=\mathbf{X}_{\mathrm{SNP}}\boldsymbol{\beta}_{\mathrm{SNP}}+\mathbf{u}\end{cases} \tag{3}$$

Where $\mathbf{X}_{\text{-SNP}}\boldsymbol{\beta}_{\text{-SNP}}$ represents the fixed effects excluding the SNP tested. At this moment, $\mathbf{u}$ is regarded as the residuals in the model at the second hierarchy.

The model at the first hierarchy is a GLMM. Both genomic heritabilities and continuous GBVs for liability can be estimated with penalized quasi-likelihood [1,3]. Moreover, the model at the second hierarchy is a linear regression model for the continuous GBVs, but the model residuals are heterogeneous. Thus, the genetic effect of the tested SNPs should be statistically inferred by the generalized least square (GLS) method [4].

Such a hierarchical solution for genome-wide generalized mixed model association analysis is summarized as follows:

(1) Estimate GBVs with GBLUP for the Logit mixed model.

Based on the GLMM at the first hierarchy, the GBLUP equations are constructed as follows:

$$\begin{bmatrix} \mathbf{X}_{\text{-SNP}}^{\text{T}}\mathbf{W}\mathbf{X}_{\text{-SNP}} & \mathbf{X}_{\text{-SNP}}^{\text{T}}\mathbf{W}\mathbf{Z} \\ \mathbf{Z}^{\mathbf{T}}\mathbf{W}\mathbf{X}_{\text{-SNP}} & \mathbf{Z}^{\mathbf{T}}\mathbf{W}\mathbf{Z}+\frac{1-h^2}{h^2}\mathbf{K}^{\mathbf{-1}} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\text{-SNP}} \\ \hat{\mathbf{g}} \end{bmatrix}=\begin{bmatrix} \mathbf{X}_{\text{-SNP}}^{\text{T}}\mathbf{W}\mathbf{y}^* \\ \mathbf{W}\mathbf{y}^* \end{bmatrix} \tag{4}$$

with $\mathbf{W}=\boldsymbol{\mu}(\mathbf{1}-\boldsymbol{\mu})$ and $\mathbf{y}^*=\boldsymbol{\eta}+\frac{\mathbf{y}-\boldsymbol{\mu}}{\boldsymbol{\mu}(\mathbf{1}-\boldsymbol{\mu})}$.

where **y** is a binary phenotype, $h^2=\frac{\sigma_g^2}{\sigma_g^2+1}$ is the genomic heritability with a GBV variance of σ_g^2 and residual variance of 1 which is assumed in GLMM.

(2) Estimate genetic effect of each SNP

By Cholesky decomposition of the **K**' inverse, i.e. $\mathbf{K}^{-1}=\mathbf{K}^{-\frac{1}{2}}\mathbf{K}^{-\frac{1}{2}}$, $\hat{\mathbf{g}}$, $\mathbf{X}_{\text{SNP}}$ and $\mathbf{u}$ are transformed to $\hat{\mathbf{g}}^*=\mathbf{K}^{-\frac{1}{2}}\hat{\mathbf{g}}$, $\mathbf{X}_{\text{SNP}}^*=\mathbf{K}^{-\frac{1}{2}}\mathbf{X}_{\text{SNP}}$ and $\mathbf{u}^*=\mathbf{K}^{-\frac{1}{2}}\mathbf{u}$. The model at the second hierarchy becomes

$$\hat{\mathbf{g}}^*=\mathbf{X}_{\text{SNP}}^*\boldsymbol{\beta}_{\text{SNP}}+\mathbf{u}^* \tag{5}$$

Because $\mathbf{u}^*$: $N(\mathbf{0},\mathbf{I}\sigma_u^2)$, the least square estimate for the genetic effect can be obtained:

$$\hat{\boldsymbol{\beta}}_{\text{SNP}}=[(\mathbf{X}_{\text{SNP}}^*)^T\mathbf{X}_{\text{SNP}}^*]^{-1}(\mathbf{X}_{\text{SNP}}^*)^T\hat{\mathbf{g}}^* \tag{6}$$

At the same time,

$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{SNP}})=[(\mathbf{X}_{\text{SNP}}^*)^{\text{T}}\mathbf{X}_{\text{SNP}}^*]^{-1}\sigma_u^2$ with $\sigma_u^2=\frac{1}{n-2}(\hat{\mathbf{g}}^*-\mathbf{X}_{\text{SNP}}^*\hat{\boldsymbol{\beta}}_{\text{SNP}})^{\text{T}}(\hat{\mathbf{g}}^*-\mathbf{X}_{\text{SNP}}^*\hat{\boldsymbol{\beta}}_{\text{SNP}})$.

(4) Statistically infer QTN by the statistic:

$$\chi^2=\frac{\hat{\boldsymbol{\beta}}_{\text{SNP}}^2}{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{SNP}})} \tag{7}$$

where χ^2 is subject to Chi-squared distribution with 1 degree of freedom.

(5) Joint association analysis

Through one test at a time, some SNPs are chosen as QTN candidates, whose numbers are generally limited to less than the population size for stepwise regression. For the GBVs obtained with the GBLUP, we jointly analyzed multiple QTN candidates to improve the statistical power to detect QTNs. Backward regression was adopted to optimize the multiple linear model:

$$\hat{\mathbf{g}}^*=\mathbf{X}_C^*\boldsymbol{\beta}_{\text{C}}+\mathbf{u}^* \tag{8}$$

where $\mathbf{X}_C^*\boldsymbol{\beta}_{\text{C}}$ are the regression terms of QTN candidates. Given the Bonferroni corrected level [5] of significance test, the genetic effects are selected stepwise and the corresponding QTNs can be identified according to the corrected statistic (7).

Simulations

Two genomic datasets of maize [6] and human [7] were used to simulate the adaptability of GRAMMAR for binary traits to population structure. The maize population has more complex structure than the human population. We extracted 300,000 SNPs for both 2640 maize and 12000 human samples through higher quality control. In whole simulations, controls and cases were generated by 1:1 for the maize population, and 2000 human samples were drawn from the human population with a low incidence rate of 5% simulated in advance. QTNs were distributed randomly over the entire SNPs, and the additive effects were sampled from a gamma distribution with shape = 1.66 and scale = 0.4. Given the genomic heritability of liability, phenotypes of the control (0) and case (1) can be generated from the genomic logit model (1).

In addition to the population structure, the number of QTNs, genome heritability, and sampling number of SNPs were considered as experimental factors in the simulations. Under almost perfect genomic control, the ROC profiles were plotted by statistical powers to detect QTNs relative to a given series of Type I errors. Statistical powers are defined as the percentage of identified QTNs that have the maximum test statistic among their 20 closest neighbors over the total number of simulated QTNs. Simulations were repeated 50 times; in each simulation, the positions and effects of QTNs simulated varied, and the average results were recorded.

Method References

1 Breslow, N. E. & Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9-25 (1993).

2 Vanraden, P. M. Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414-4423 (2008).

3 Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *American journal of human genetics* **98**, 653-666, doi:10.1016/j.ajhg.2016.02.012 (2016).

4 Kariya, T. & Kurata, H. *Generalized Least Squares*. (John Wiley & Sons, 2004).

5 Hochberg, Y. & Tamhane, A. C. *Multiple Comparison Procedures*. (John Wiley & Sons, Inc., 1987).

6 Romay, M. C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. *Genome biology* **14**, R55 (2013).

7 Consortium, W. T. C. C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).