Optimal Genomic Control in Large-scale Genetic Associations for Binary Diseases

Yuxin Song[1], Li Jiang[1], Zhiyu Hao[2]and Runqing Yang[1, 2]

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

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

Correspondence to Runqing Yang (runqingyang@cafs.ac.cn)

Complex computation and approximate solution hinder the application of generalized linear mixed models (GLMM) into genome-wide association studies. We extended GRAMMAR to handle binary diseases by considering genomic breeding values (GBVs) estimated in advance as a known predictor in genomic logit regression, and then controlled polygenic effects by regulating downward genomic heritability. Using simulations and case analyses, we showed in optimizing GRAMMAR, polygenic effects and genomic controls could be evaluated using the fewer sampling markers, which extremely simplified GLMM-based association analysis in large-scale data. In addition, joint analysis for quantitative trait nucleotide (QTN) candidates chosen by multiple testing offered significant improved statistical power to detect QTNs over existing methods.

Introduction

Although usually expressed as binary phenotypes, many disease traits in plants and animals are thought to be controlled by a number of loci each having a small effect [1,2]. Thus, random polygenic effects excluding the tested markers should be considered in genome-wide association study (GWAS) for disease traits to correct the population stratification and cryptic relatedness, as linear mixed model (LMM) does for quantitative traits [3,4]. Logistic regression, as a kind of generalized linear model (GLM) [5,6], has been used earlier for genome-wide association analysis with the disease traits [7]. Despite correction for fixed-effect covariates [8-10], logistic regression still produces inflation of association test statistics. Therefore, it is necessary to introduce a generalized linear mixed model (GLMM) [11] that considers random polygenic effects to increase the power to map QTNs for disease traits.

However, genome-wide GLMM association consumes much more computing time than mixed model association with either restricted maximum likelihood estimation

Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

(REML) [12] or Markov Chain Monte Carlo (MCMC) iteration [13]. If using the maximum likelihood in estimation and approximations to avoid numerical integration [14], the GLMM yields a problem of serious bias induced by the approximations [15], especially solutions tending toward the positive/negative infinity. In the GWAS for case-control studies, a non-random sampling of cases from the population results in biased estimation for genomic heritability. Analyzed by the LMM for binary phenotypes, genomic heritability estimate is transformed to a liability scale by adjusting both for scale and for ascertainment of the case samples [16]. The genomic heritability of liability is estimated in a biased manner for disease traits, even though it is done by GLMM via MCMC iteration. Based on the calibrated genomic hereditability for case-control ascertainment, a Chi-squared score statistic for GWAS of disease traits is computed from posterior mean liabilities (PMLs) under the liability-threshold model [17]. The individual PMLs are estimated with a multivariate Gibbs sampler, which increases the computational demand. For simplification to GLMM-based association analysis, the GMMAT [18] and SAIGE [19] separately extend the EMMAX [20] and BOLT-LMM [21], respectively, for normally distributed traits to binary phenotypes.

Owing to the computational intensity and approximate solutions obtained, GLMM can hardly be employed in GWAS for disease traits. Moreover, genomic heritability cannot be accurately estimated for complex diseases, especially in ascertained case-control studies. Motivated by the optimal genomic control for mixed model association analysis for quantitative traits distributed normally, we extended GRAMMAR [22] to handle binary traits by considering genomic breeding values (GBVs) estimated in advance as a known predictor in genomic logit regression, and then, optimized the genomic control for GRAMMAR for binary traits by regulating the downward genomic heritability to estimate the residual phenotypes. The complicated GLMM does not need to be directly solved by the Optim-GRAMMAR for binary traits, and it only repeatedly estimates GBVs with genomic best linear unbiased prediction (GBLUP) [23] for GLMM, achieving genome-wide GLMM association analysis rapidly. Finally, we jointly analyzed the candidate quantitative trait nucleotides (QTNs) chosen by multiple testing to improve the statistical power to detect QTNs.

Results

Statistical properties of Optim-GRAMMAR for binary traits

Based on the two genomic datasets, we simulated phenotypes controlled by 40, 200,

and 1,000 QTNs at the low (0.2), moderate (0.5), and high (0.8) genomic heritability, respectively. The statistical properties of Optim-GRAMMAR using a test at once for binary traits were investigated by comparing it with GRAMMAR, GMMAT, LTMLM, and SAIGE. The Q-Q and ROC profiles are displayed selectively in Figure 1 and Figure 2, respectively, and in Supplementary Figure 1S and Figure 2S, respectively, in detail. The genomic controls are estimated in Table 1S. Making genomic control infinitely close to 1.0, Optim-GRAMMAR achieved almost the same statistical power to detect QTNs as the GMMAT which approximates the exact GLMM, irrespective of how many QTNs and heritabilities are simulated. Among Optim-GRAMMAR and the four competing methods, GRAMMAR had the lowest genomic controls and statistical power, and for GRAMMAR, the population structure was more complex and the false negative rate was larger. Although LTMLM achieved the highest statistical power to detect QTNs for all simulated phenotypes for the maize dataset, and SAIGE demonstrated a higher statistical power for the dataset controlled by 1,000 QTNs at the genomic heritability of 0.2. A strong false positive error rate was observed for SAIGE. In the human dataset, there were no distinct differences in the statistical properties between GRAMMAR-lambda and the four competing methods, although GRAMMAR provided some false negative errors.

After optimization for genomic control, Optim-GRAMMAR jointly analyzed multiple QTN candidates chosen from a test at once at a significance level of 0.05. For convenience for comparison, we analyzed the statistical powers obtained with one test at a time and joint analyses together. By backward regression analysis, Optim-GRAMMAR evidently exhibited improved statistical power. In contrast, LTMLM was inferior to joint analysis of Optim-GRAMMAR in the terms of statistical power, even with the highest false positive rates.

Calculation of GRMs and GCs with the sampling markers

To investigate the effects of sampling markers on Optim-GRAMMAR, we randomly took 3 K, 5 K, 10 K, 20 K, and 25 K SNPs from the entire genomic markers to calculate GRM. Changes in the genomic control at the varied sampling levels of SNPs are depicted in Figure 3 for Optim-GRAMMAR, GRAMMAR, GMMAT, and SAIGE. Because LTMLM cannot sample SNPs, it was not included in the comparison. No competing method stably controlled the positive/negative false errors using less than 25 K sampling SNPs, besides SAIGE for human phenotypes. Specifically, GMMAT

gradually controlled the positive false errors as the sampling markers increased; GRAMMAR controlled the negative false rate by sampling less markers, while SAIGE produced serious false negative errors in the complex maize population. In comparison, Optim-GRAMMAR still retained a high statistical power to detect QTNs through almost perfect genomic control, even using less than 3000 sampling markers (see Supplementary Figure 3S and Figure 4).

Application of Optim-GRAMMAR to WTCCC study

We were authorized to re-analyze the Wellcome trust case-control consortium (WTCCC) study 1 [24]. 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, 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 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 Optim-GRAMMAR using entire genomic markers and 5,000 sampling SNPs, respectively, to estimate the GRM. The Q-Q and Manhattan profiles for the six common diseases are depicted in Figure 4S and Figure 5S obtained with the Optim-GRAMMAR using a test at once and the four competing methods used in simulations, while in Figure 5S with the Optim-GRAMMAR using joint association analyses. The association analyses illustrated that (1) under perfect genomic control, Optim-GRAMMAR found the QTNs for each disease on each chromosome, and the numbers of detected QTNs were not less than all the competing methods; and (2) in Optim-GRAMMAR, joint association analyses detected more QTNs than a test at once. As compared to Optim-GRAMMAR, 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 obtained the highest genomic control. Additionally, LTMLM estimated the abnormal genomic heritabilities for CAD, BD, T2D, and HT, producing unstable genomic controls.

Further, we conducted strict QC for each dataset, as done in [16] for estimating genomic heritability. Despite this, the missing heritabilities could not be normally estimated for BD and HT. As shown in Figure 6S and Figure 7S, all methods exhibited clear and comparable association results, except for GRAMMAR. Interestingly, both LTMLM and GMMAT seriously underestimated the genomic heritability for each disease after strict QC. In summary, Optim-GRAMMAR could efficiently and robustly map QTNs for binary diseases and did not depend on 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 phenotype to output of mapping QTNs for all the methods. Table 2S shows that Optim-GRAMMAR reduced the computing time by dozens of times with the lowest memory footprint.

Discussion

Development of the GRAMMAR for GLMM association was an essential prerequisite for extending the Optim-GRAMMAR to rapidly optimize mixed model association analysis for binary traits. For genomic GLMM, however, no binary residuals could be produced because of the scale difference between binary phenotype and predictors. Thus, we considered the GBVs estimated in advance as a known predictor in genomic logit regression and then executed association tests for candidate markers. This ensured that GRAMMAR had the lowest computing complexity for association tests for binary traits among the existing GLMM-based association methods [17-19]. Because GRAMMAR for binary diseases produces high false negative rates for quantitative traits distributed normally, we optimized the genomic control for GRAMMAR by regulating downward genomic heritability to underestimate the GBVs with GBLUP equations for GLMM. Thus, optim-GRAMMAR solved the GBLUP equations and performed association tests with simple logit regression only for several iterations, thus improving the computational efficiency for genome-wide GLMM association analysis.

Several GLMM-based association methods such LTMLM, GMMAT, and SAIGE have simplified genome-wide mixed model association analysis for binary traits to a certain extent, but they are more appropriate to handle the less complex populations

such as human datasets [18,19]. Moreover, the heritability for binary diseases could not be robustly and precisely estimated using genomic markers [16] [14,15], which also limited the efficient application of these association methods. In contrast, because optim-GRAMMAR does not need to directly estimate genomic heritability, it can powerfully and robustly map QTNs for binary traits in complexly structured populations. Within the framework of Optim-GRAMMAR, further, joint analysis for the candidate quantitative trait nucleotides chosen by multiple testing significantly improved statistical power to detect QTNs with almost perfect genomic control.

The Optim-GRAMMAR extremely simplified genome-wide GLMM association analysis for binary traits in large-scale population. For a genomic dataset containing m SNPs genotyped on n individuals, Optim-GRAMMAR for binary traits took only the computing complexity of $O(mn^2)$ to build the relationship matrix and $O(imn)$ for association tests with i rounds to optimize genomic controls. If we solved genetic effects of the m_1 sampled markers using ridge regression [25] with given heritability and then estimated GBVs, then the computing complexity to build the information matrix would reduce to $O(m_1^2 n)$, as in FaST-LMM-Select [26]. At the same time, if we evaluated genomic control using m_1 sampled markers at each iteration, then the computing complexity for association tests would reduce to $O(im_2 n)$. For the simulated 8 million SNPs on 400,000 individuals, Optim-GRAMMAR required only 5.1hr to analyze single binary phenotype by sampling 5,000 SNPs to calculate GRM, 0.9hr of which to build the information matrix and optimize genomic controls, while SAIGE did about 534 hr [15]. A user friendly GRL-Binary software was developed, which is freely available at https://github.com/RunKingProgram/Binary Optim-GRAMMAR.

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 (2017A001).

Competing interests

The authors declare no competing financial interests.

References

1. Bulmer, M.G. The Effect of Selection on Genetic Variability. *American Naturalist* **105**, 201-211

(1971).
2. Falconer, D.S. *Introduction to Quantitative Genetics, 2nd ed.*, (Longman, London, 1981).
3. Henderson, C.R. *Applications of linear models in animal breeding*, (University of Guelph, Guelph, 1984).
4. Yu, J.M. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203-208 (2006).
5. Wedderburn, R.W.M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439-447 (1974).
6. McCullagh, P. & Nelder, J.A. *Generalized linear models, 2nd ed.*, (Chapman and Hall, New York, 1989).
7. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
8. Mefford, J. & Witte, J.S. The Covariate's Dilemma. *PLoS Genet* **8**, e1003096 (2012).
9. Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet* **8**, e1003032 (2012).
10. Zaitlen, N. *et al.* Analysis of case-control association studies with known risk variants. *Bioinformatics* **28**, 1729-1737 (2012).
11. Breslow, N.E. & Clayton, D.G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9-25 (1993).
12. Patterson, H.D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554 (1971).
13. Sorenrsen, D. & Gianola, D. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*, (Springer, New York, 2002).
14. Schall, R. Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727 (1991).
15. 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).
16. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
17. Hayeck, T.J. *et al.* Mixed Model with Correction for Case-Control Ascertainment Increases Association Power. *American Journal of Human Genetics* **96**, 720-730 (2015).
18. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* **98**, 653-66 (2016).
19. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341 (2018).
20. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-354 (2010).
21. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284-290 (2015).
22. 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).
23. Vanraden, P.M. *et al.* Invited review: reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16-24 (2009).

24. Consortium, W.T.C.C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
25. Hoerl, A.E. & Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55-67 (1970).
26. Jennifer, L. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525-526 (2012).

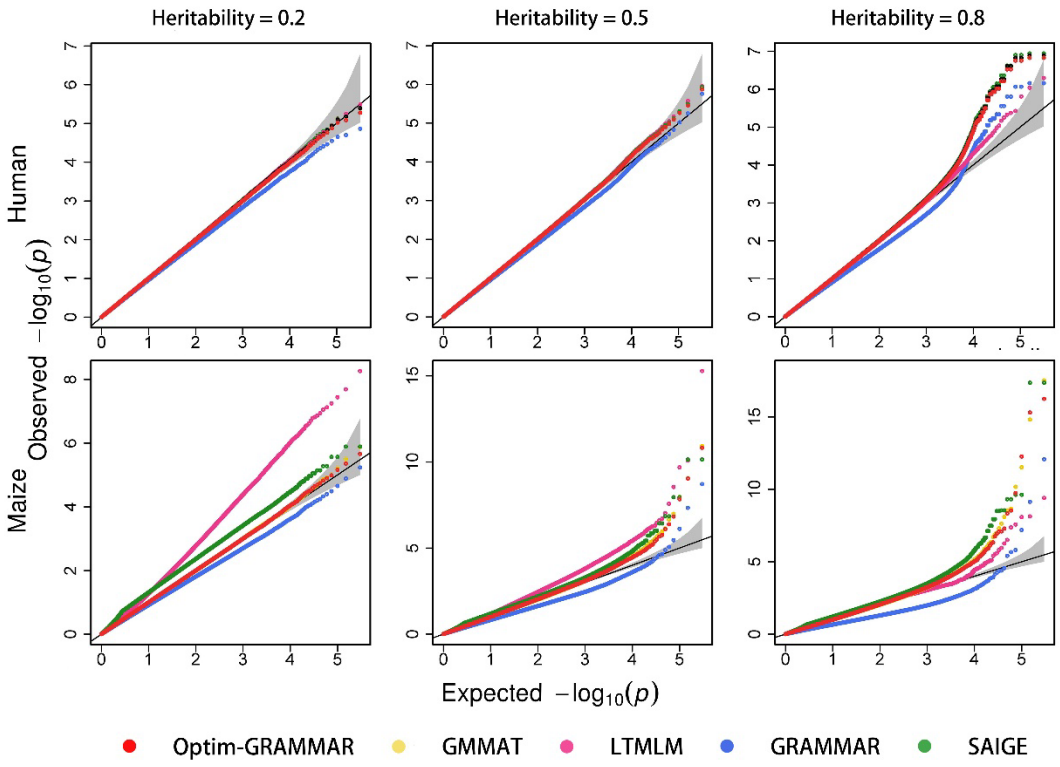

Figure 1: Comparison in the Q-Q profiles between Optim-GRAMMAR and the four competing methods. The simulated phenotypes are controlled by 200 QTNs with the low, moderate and high heritabilities in human and maize. The Q-Q profiles for all simulated phenotypes are reported in Supplementary Figure 1S.

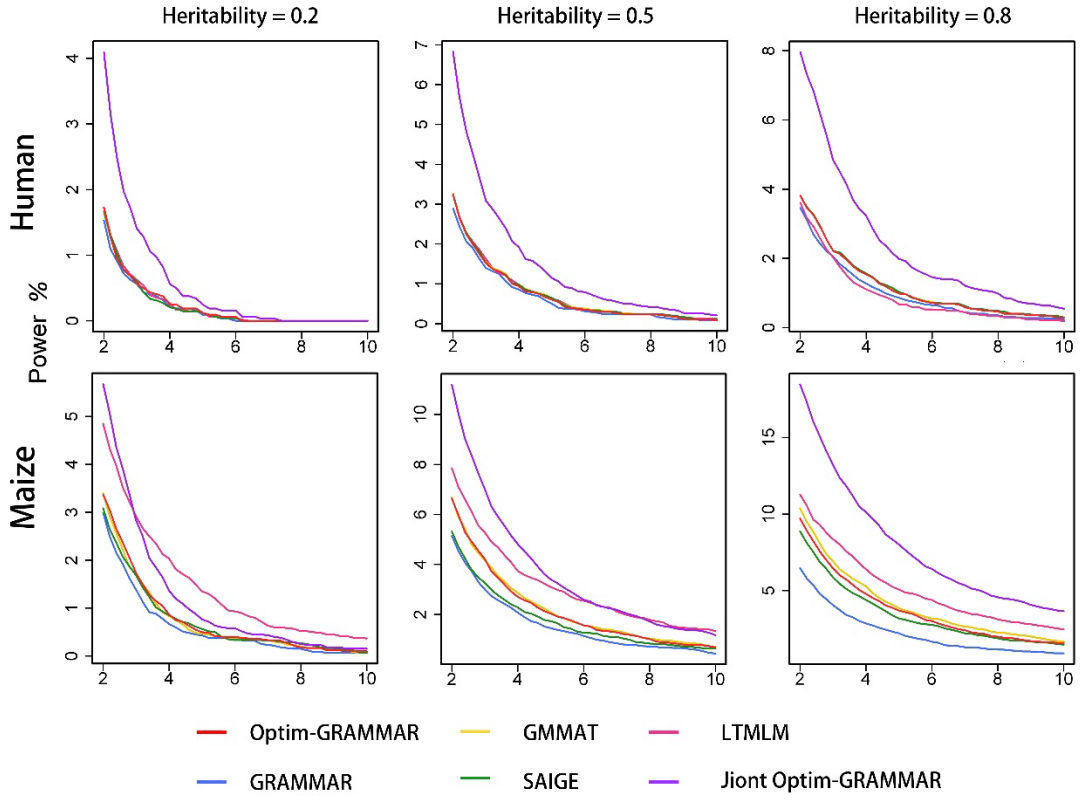

Figure 2: Comparison in the ROC profiles between Optim-GRAMMAR and the four competing methods. The ROC profiles are plotted using the statistical powers to detect QTNs relative to the given series of Type I errors. Here, the simulated phenotypes are controlled by 200 QTNs with the low, moderate and high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

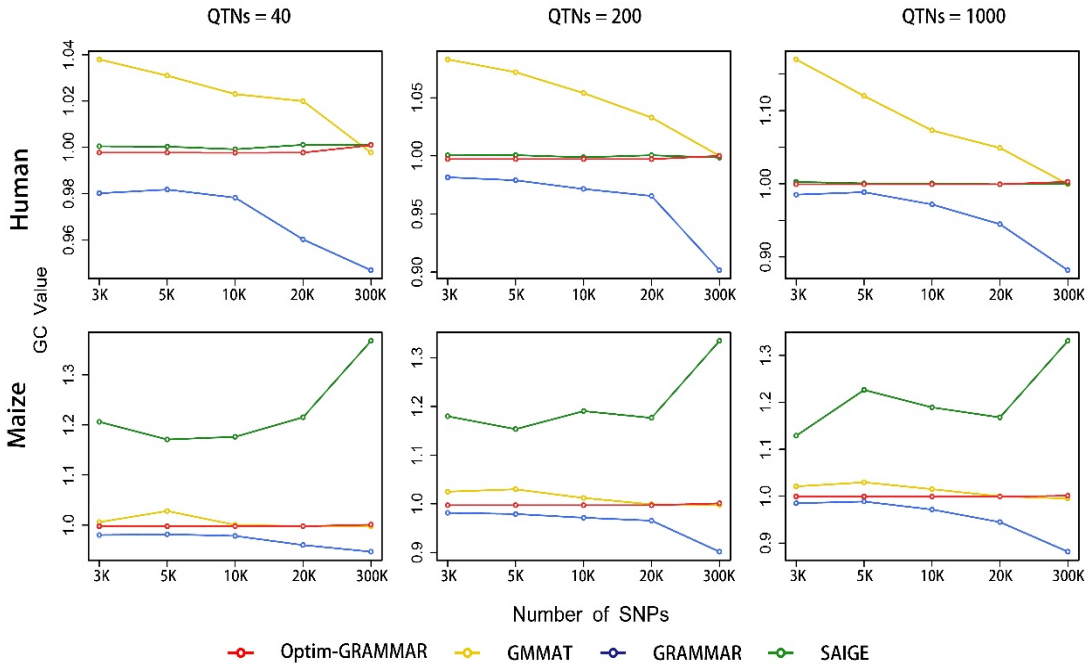

Figure 3: Changes in genomic controls with the number of sampling SNPs for Optim-GRAMMAR and the three competing methods. Genomic control is calculated by averaging genome-wide test statistics. The simulated phenotypes are controlled by 40, 200 and 1000 QTNs with the moderate heritability in human and maize.

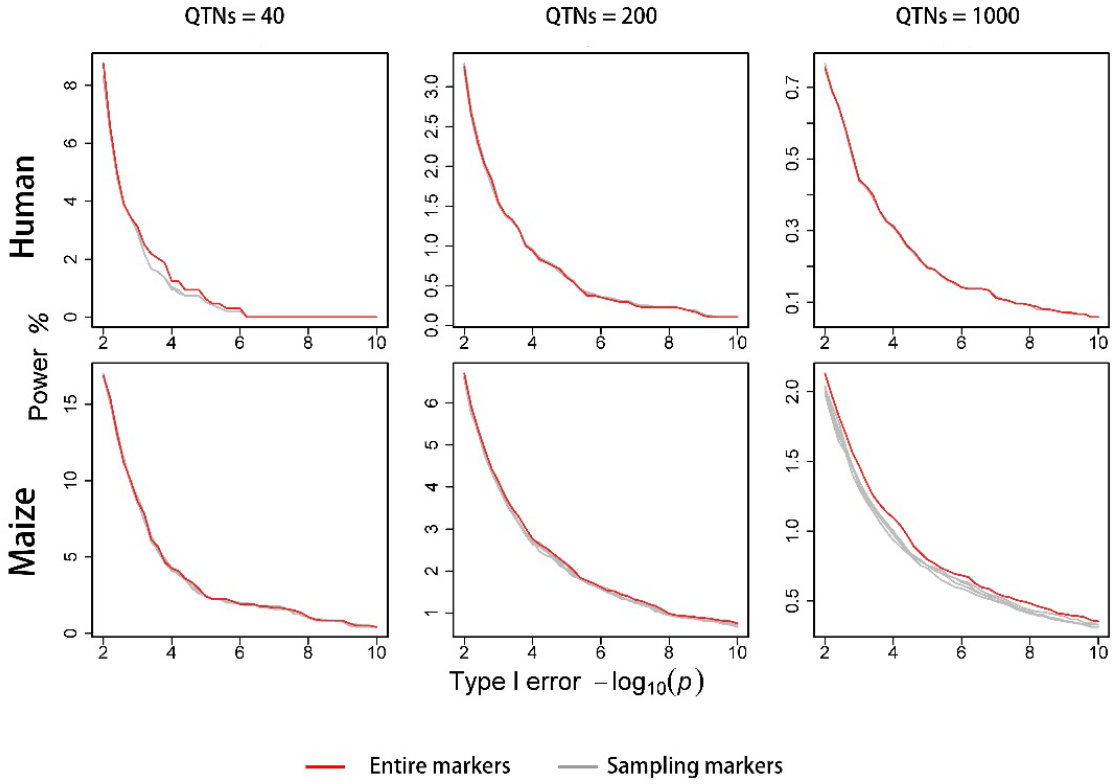

Figure 4: Changes in ROC profiles with the number of sampling SNPs for Optim-GRAMMAR. The simulated phenotypes are controlled by 40, 200, 1,000 QTNs with the moderate heritability in human and maize.

Online Methods

Genomic logit regression

Complex disease traits, as binary ones, usually follow binomial or Poisson distributions, so the generalized linear model (GLM) [1,2] is used to map QTLs controlling the traits. Assume that *n* individuals are recorded for phenotypic values and genotyped for *m* genetic markers. Distinguishing these markers from major and common alleles in a magnitude of effects of the markers on quantitative traits, we describe the relationship between all markers (predictors) and the mean of the exponential distribution family in the following logit regression:

$$\ln\left(\frac{\mu}{1-\mu}\right) = \mathbf{Xb} + \mathbf{Z}_1\mathbf{a}_1 + \mathbf{Z}_2\mathbf{a}_2 \quad (1)$$

where μ denotes the expectations of phenotypic distribution, **b** is the systematic environment effect; the population structure (stratification) which results in phenotypic differences among subpopulations is always considered here, except for sex, age, and some initial experimental conditions. $\mathbf{a}_1$ is the large genetic effect of *q* markers on phenotype, $\mathbf{a}_2$ is the minor or zero effect of the *m*-*q* markers on phenotype, and **X**, $\mathbf{Z}_1$, and $\mathbf{Z}_2$ are the corresponding design matrices of **b**, $\mathbf{a}_1$, and $\mathbf{a}_2$, respectively.

GRAMMAR for binary disease traits

We define the GBVs as

$$\mathbf{g} = \mathbf{Z}_1\mathbf{a}_1 + \mathbf{Z}_2\mathbf{a}_2 \quad (2)$$

Then, model (1) becomes

$$\ln\left(\frac{\mu}{1-\mu}\right) = \mathbf{Xb} + \mathbf{g} \quad (3)$$

which is a the GLMM [3]. Under the assumption that $(\mathbf{a}_1, \mathbf{a}_2)$: $N_m(\mathbf{0}, \mathbf{I}\sigma_a^2)$ with minor σ_a^2 for each marker, the GBVs are turned into random effects and $\mathbf{g}$: $N_n(\mathbf{0}, \mathbf{K}\sigma_g^2)$ with genomic variance of traits $\sigma_g^2 = m\sigma_a^2$ and the genomic relationship matrix (GRM) **K** [4]. Based on the model (3), the GBVs can be estimated with the following GBLUP equations:

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{W} \\ \mathbf{W}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z}+\frac{1-h^2}{h^2}\mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}^{*} \\ \mathbf{W}\mathbf{y}^{*} \end{bmatrix} \quad (4)$$

with

$$\mathbf{W} = \boldsymbol{\mu}(\mathbf{1}-\boldsymbol{\mu}) \text{ and } \mathbf{y}^{*} = \ln\left(\frac{\boldsymbol{\mu}}{\mathbf{1}-\boldsymbol{\mu}}\right) + \frac{\mathbf{y}-\boldsymbol{\mu}}{\boldsymbol{\mu}(\mathbf{1}-\boldsymbol{\mu})}.$$

where $\mathbf{y}$ is a binary phenotype, and $h^2 = \frac{\sigma_g^2}{\sigma_g^2+1}$ is the unknown genomic heritability of liability with the residual variance of 1 to be assumed in GLMM, which can be estimated in advance using the REML for GLMM [3,5].

Unlike the normally distributed quantitative traits, the residuals for binary traits cannot be directly obtained due to the difference in scale between the predictors and response variables. Within the framework of GRAMMAR, thus, we eliminated polygenic effects on the binary phenotype by regarding the estimated GBVs $\hat{\mathbf{g}}$ as a known predictor in the following GLM:

$$\ln\left(\frac{\boldsymbol{\mu}}{\mathbf{1}-\boldsymbol{\mu}}\right) = \mathbf{z}_{\mathrm{SNP}} a_{\mathrm{SNP}} + \hat{\mathbf{g}} \quad (5)$$

with a regression item $\mathbf{z}_{\mathrm{SNP}} a_{\mathrm{SNP}}$ of the SNP tested.

With the iteratively re-weighted least square method [1], we obtained the maximum likelihood estimate for the SNP effect as:

$$\hat{a}_{\mathrm{SNP}} = (\mathbf{z}_{\mathrm{SNP}}^{\mathrm{T}}\mathbf{W}\mathbf{z}_{\mathrm{SNP}})^{-1}\mathbf{z}_{\mathrm{SNP}}^{\mathrm{T}}\mathbf{W}(\mathbf{y}^{*}-\hat{\mathbf{g}}) \quad (6)$$

The test statistic to infer the association of the SNP with binary traits is generally formulated by

$$\chi^2 = \frac{\hat{a}_{\mathrm{SNP}}^2}{\mathbf{z}_{\mathrm{SNP}}^{\mathrm{T}}\mathbf{W}\mathbf{z}_{\mathrm{SNP}}} \quad (7)$$

which is subject to the chi-squared distribution with the 1 degree of freedom.

Optimal Genomic control

In GRAMMAR for binary traits, replacement of polygenic effects excluding QTNs with GBVs deflates the test association statistics, which yields a high false negative rate. By regulating the downward genomic heritability, we can more accurately estimate the polygenic effects with the GBLUP Equation (4). The polygenic heritability less than genomic heritability is determined by optimizing genomic control for association tests.

Such an optimization for GRAMMAR can be summarized in the following steps:

1) Set the searching open interval of h^2 to (0, 1)
2) Estimate the GBVs $\hat{\mathbf{g}}$ using Equation (4);
3) Statistically infer the genetic effect for each SNP by the chi-squared statistic (7);
4) Calculate the genome-wide chi-squared mean or statistical probability for each SNP;
5) Plot the quantile-quantile (Q-Q) profile for genome-wide statistical probabilities;
6) Update h^2 with Brent's method [6];
7) Repeat step (2)-(6) until the genome-wide chi-squared mean reaches 1.0 plus or yields a satisfactory Q-Q plot.

Joint association analysis

After optimizing genomic control for GRAMMAR, we jointly analyzed multiple QTN candidates to improve the statistical power to detect QTNs for binary disease traits. Multiple QTN candidates were chosen within the interval of significance level 0.05 to the Bonferroni corrected criterion [7], so that the number of QTN candidates was limited to be no greater than the population size. Backward regression was adopted to optimize the multiple GLM with known optimized polygenic effects in a stepwise manner:

$$\ln\left(\frac{\boldsymbol{\mu}}{\mathbf{1}-\boldsymbol{\mu}}\right) = \mathbf{Xb}+\mathbf{Z}_1\mathbf{a}_1 + \hat{\mathbf{g}} \quad (8)$$

Given the Bonferroni corrected significance level, the significant QTN effects remained in the model (8) according to the corrected statistic (7).

Simulations

Two genomic datasets of human [8]and maize [3] samples were used to simulate the adaptability of GRAMMAR for binary traits to population structure. The maize population has a more complex structure than the human population. Then, 300,000 SNPs for both 12000 people and 2640 maize were extracted through higher quality control. In whole simulations, control and case samples were constrained to 1:1 for the maize population, and 3000 cases were selected from the human population with low incidence rate of 5% simulated in advance. QTNs were distributed randomly over the entire SNPs, whose additive effects were sampled from a gamma distribution with

shape = 1.66 and scale = 0.4. Given the genomic heritability of liability, phenotypes of control (0) and case (1) can be generated from the genomic logit model (1).

In addition to population structure, the number of QTNs, genome heritability, and sampling number of SNPs were considered as experimental factors in the simulations. Under the optimized genomic control infinitely close to 1.0, the ROC profiles can be plotted by statistical powers to detect the 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, and in each simulation, the positions and effects of QTNs simulated were varied and the average results were recorded.

Method References

1. Wedderburn, R.W.M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439-447 (1974).
2. McCullagh, P. & Nelder, J.A. *Generalized linear models, 2nd ed.*, (Chapman and Hall, New York, 1989).
3. Breslow, N.E. & Clayton, D.G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9-25 (1993).
4. Vanraden, P.M. Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414-4423 (2008).
5. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* **98**, 653-66 (2016).
6. Brent, R.P. *Algorithms for minimization without derivatives*, (Prentice-Hall, New Jersey, 1973).
7. Hochberg, Y. & Tamhane, A.C. *Multiple Comparison Procedures*, (John Wiley & Sons, Inc., New York, 1987).
8. Romay, M.C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. *Genome biology* **14**, R55 (2013).