A Shortcut Approach for Large-scale Mixed Model Associations with Binary Traits

Runqing Yang (runqingyang@cafs.ac.cn)
Research Centre for Aquatic Biotechnology, Chinese Academy of Fishery Sciences, Beijing 100141, People's Republic of China & College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

Jun Bao (jbao@neau.edu.cn)
College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

Runqing Yang
Research Centre for Aquatic Biotechnology, Chinese Academy of Fishery Sciences, Beijing 100141, People's Republic of China & College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

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

Zhiyu Hao
College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

Jun Bao
College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

Method Article

Keywords: Generalized linear mixed model, Genomic control, Computational efficiency, Binary trait, Large-scale population

Posted Date: March 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-312421/v2

Abstract

Generalized linear mixed models exhibit computationally intensive and biasness in mapping quantitative trait nucleotides for binary diseases. In genomic logit regression, we consider genomic breeding values estimated in advance as a known predictor, and then correct the deflated association test statistics by using genomic control, thereby successfully extending GRAMMAR-Lambda to analyze binary diseases in a complex structured population. Because there is no need to estimate genomic heritability and genomic breeding values can be estimated by a small number of sampling markers, the generalized mixed-model association analysis has been extremely simplified to handle large-scale data. With almost perfect genomic control, joint analysis for the candidate quantitative trait nucleotides chosen by multiple testing offered a significant improvement in statistical power.

Introduction

Complex diseases are generally thought to be the quantitative traits controlled by a number of loci each having a small effect [1, 2]. Conventionally, logistic regression in generalized linear models (GLMs) [3, 4] instead of linear regression models was used to map quantitative trait nucleotides (QTNs) for binary phenotypes. Although logistic regression can correct for fixed-effect covariates [5-7], it still leads to inflation of associated test statistics. Therefore, the generalized linear mixed-model (GLMM)[8] which considers random polygenic effects to increase the power to map QTNs for disease traits was introduced. However, the computation for genome-wide GLMM association is much more intensive than that for mixed-model association with either restricted maximum likelihood estimation (REML)[9] or Markov chain Monte Carlo (MCMC) iterations[10]. If using the maximum likelihood in estimation and approximations to avoid numerical integration, the GLMM produces a problem of serious bias induced by the approximations [11], especially solutions tending toward positive/negative infinity.

In ascertained case-control studies in which the proportions of cases and controls are not a random sample from the population, GLMM provides biased estimates of the genomic heritability for disease traits [12] and therefore suffers a loss in the power to detect QTNs. Based on the calibrated genomic hereditability for case-control ascertainment, a Chi-squared score statistic for association tests was computed from the posterior mean liabilities under the liability-threshold model [13]. For simplification of GLMM-based association analysis, GMMAT [14] and SAIGE [15] separately extend the EMMAX [16] and BOLT-LMM [17] for normally distributed traits to binary diseases. Dividing the deflated test statistics from the GRAMMAR by genomic control, GRAMMAR-lambda was developed for extremely efficient genome-wide mixed-model association analysis (Not published yet). In this study, we extend the GRAMMAR-Lambda for quantitative traits normally distributed to handle binary diseases by regarding genomic breeding values (GBVs) estimated in advance as a known predictor in genomic logit regression and then calibrating the test statistics by genomic control. Because genomic heritability does not need be estimated and GBVs can be obtained by a small number of sampling markers, GLMM-based association

analysis is extremely simplified. Moreover, joint analysis for the QTN candidates chosen by multiple testing significantly improves the statistical power to detect QTNs for binary diseases.

Results

Statistical properties of GRAMMAR-lambda for binary diseases

For the two genomic datasets, phenotypes were simulated to be controlled by 40, 200, and 1000 QTNs at the low (0.2), moderate (0.5), and high (0.8) genomic heritability, respectively. The GRAMMAR-Lambda, a test at once is comparable with the four competing methods—GRAMMAR, GMMAT, LTMLM, and SAIGE. Association results are displayed selectively in Figure 1 for the Q-Q profiles and Figure 2 for ROC profiles (Figure 1S and Figure 2S for details); the genomic controls are estimated in Table 1S. Under genomic control of exact 1, GRAMMAR-Lambda performs stably and with high statistical power, irrespective of how much the QTNs control the quantitative traits and how complex the population structures are. In contrast, GMMAT exhibits the same statistical power as GRAMMAR-Lambda, but with slightly low, even instable genomic control. For GRAMMAR, genomic controls are the lowest among GRAMMAR-Lambda and the four competing methods, the population structure is more complex, and the false negative rate is higher. Although LTMLM detected most QTNs for all simulated phenotypes in maize and SAIGE provided higher statistical power for those controlled by 1,000 QTNs at the genomic heritability of 0.2, they produce strong false positive errors. For humans, there were no distinct differences in statistical properties of GRAMMAR-Lambda and the four competing methods, although GRAMMAR produced a slight false negative error.

Furthermore, GRAMMAR-Lambda could jointly analyze multiple QTN candidates chosen from a test at once at a significance level of 0.05. For convenience of comparison, statistical powers of detecting QTNs are depicted together in Figure 1 selectively and in Supplementary Figure 1S in detail. Through backward multiple regression analysis, GRAMMAR-Lambda offered significant improved statistical power, but also kept almost perfect genomic control that was infinitely close to 1. Even at the highest false positive rates, LTMLM was inferior to GRAMMAR-Lambda with joint analysis in terms of statistical power.

Sensitivity to estimate genomic heritability

For normally distributed traits, GRAMMAR-Lambda does not need to estimate genomic heritability. To test whether this finding fits the binary diseases or not, we pre-specify the genomic heritability at 0.5 to analyze the simulated phenotypes controlled by 200 QTNs at different levels of heritabilities (see Figure 3S). Figure 3 compares the statistical behaviors of GRAMMAR-Lambda and the other methods by specifying and estimating genomic heritability. As shown in each plot, both QQ and ROC curves obtained by pre-specifying the genomic heritability almost overlapped with those obtained by estimating genomic heritability. Extensive simulations in optimizing GRAMMAR showed that the lower bound of the given heritability may be lower for binary diseases than that for normally distributed ones (data not shown). This supports the fact that while implementing GRAMMAR-Lambda for binary diseases, genomic heritability is set to 0.5 by default or to the empirical heritability of traits, if available.

Calculation of GRMs with the sampling markers

When the genomic heritability is pre-specified to 0.5, we randomly took 3 K, 5 K, 10 K, 20 K, and 25 K SNPs from the entire genomic markers to analyze the simulated phenotypes controlled by the varied numbers of QTNs at the heritability 0.5 using all methods, except LTMLM which cannot sample SNPs to estimate heritability. Changes in genomic control at the varied sampling levels of SNPs are depicted in Figure 4 for GRAMMAR-Lambda, GRAMMAR, GMMAT, and SAIGE. No competing method can stably control the positive/negative false errors using less than 20 K sampling SNPs, and SAIGE for human phenotypes simulated. Specifically, GMMAT gradually controls the positive false errors when the sampling markers increased; GRAMMAR seemed to reduce the negative false rate by sampling less markers, while SAIGE produced a serious false negative error in the complex maize population. In comparison, GRAMMAR-Lambda still retained high statistical power to detect QTNs through perfect genomic control, even by using less than 3000 sampling markers (see Figure 4S and Figure 5S).

Application of GRAMMAR-Lambda to WTCCC study

We were authorized to collect the data of 11,985 cases of six common diseases and 3,004 shared controls, genotyped at a total of 490,032 SNPs from the Wellcome Trust Case Control Consortium (WTCCC) study 1 [18]. 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 was 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 GRAMMAR-Lambda in two ways: to estimate the genomic heritability and GBVs together using all genomic markers and to estimate only GBVs by randomly sampling 5,000 SNPs with a given heritability of 0.5. The Q-Q and Manhattan profiles for the six common diseases are shown in Figure 6S and Figure 7S, which depict GRAMMAR-Lambda and the four competing methods used in simulations. We concluded that (1) under perfect genomic control, GRAMMAR-Lambda found the QTNs for each disease on each chromosome, and the numbers of the detected QTNs were not less than all the competing methods; and (2) in GRAMMAR-Lambda, joint association analyses detected more QTNs than a test at once. Compared with GRAMMAR-Lambda, GRAMMAR detected less QTNs with the lowest genomic control among all methods, while GMMAT yielded more SNPs whose -log(p) exceeded the Bonferroni corrected thresholds for CAD, T1D, T2D, and HT, but it behaved the largest

genomic control. Additionally, LTMLM estimated the abnormal genomic heritabilities for CAD, BD, T2D, and HT, producing an unstable genomic control.

For better estimating the genomic heritability, furthermore, we conducted strict QC for each dataset, as performed previously in [12]. Despite this, the missing heritabilities could not be normally evaluated 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 the genomic heritability for each disease after strict QC. GRAMMAR-lambda could extremely 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 phenotype to the output of mapping QTNs for all the methods. Table 2S shows that GRAMMAR-Lambda several times to dozens of times reduced the computing time by almost dozens of times with the lowest memory footprint.

Discussion

Although interpretable and predictable for discrete traits, the GLMM cannot be efficiently applied into GWAS for complex diseases because of intractable solutions and computation. Several GLMM-based association methods such as LTMLM, GMMAT, and SAIGE have simplified the genome-wide mixed-model association analysis for binary diseases to certain extent, but they are more likely to appropriately handle less complex populations like humans [14, 15]. In this study, we successfully extended GRAMMAR-Lambda from normally distributed traits to binary diseases, extremely simplifying the generalized mixed-model association analysis. In complex structured populations, the extended GRAMMAR-Lambda can more efficiently and robustly map QTNs for binary diseases than the existing GLMM-based association methods. With almost perfect genomic control, joint analysis for the candidate QTNs chosen by multiple testing significantly improved the statistical power, under the framework of GRAMMAR-Lambda.

Prior to applying GRAMMAR-Lambda for binary diseases, GRAMMAR needs to be constructed to rapidly associate binary phenotypes with candidate markers. Within the framework of GLM, however, no binary residuals could be produced because of the difference in scale between the binary phenotype and predictors. Therefore, we took the GBVs estimated in advance as a known predictor in genomic logit regression and then executed association tests for candidate markers. Moreover, the heritability for binary diseases could not be robustly and precisely estimated using genomic markers [9, 11, 12], which also limits the efficient application of the existing GLMM-based association methods. Inheriting the advantages for the normally distributed traits, GRAMMAR-Lambda extremely efficiently performed genome-wide GLMM association analysis in three ways: 1) the genomic heritability was not required to be estimated for binary diseases; 2) it used fewer sampling markers to calculate the GRM; and 3) the computing complexity of association tests was the lowest for binary phenotypes, as that of the PLINK [19].

Generally, computing costs are attributed to building the GRM, estimating variance components or polygenic heritability, and computing association statistics in genome-wide mixed-model association

analysis [20]. To date, no algorithm has been developed that can comprehensively reduce these three computational charges. For a genomic dataset containing m SNPs genotyped on n individuals, GRAMMAR-Lambda for binary diseases took only the computing complexity of $O(mn^2)$ to build the relationship matrix and $O(mn)$ for association tests. When analyzing a large-scale population, we solved the effects of m_0 of the sampled markers using ridge regression [21] with given heritability and then estimated GBVs as $\mathbf{Z_0 a_0}$, which reduced the computing time to build the information matrix to , as in FaST-LMM-Select [22]. For the simulated 8 million SNPs on 400,000 individuals, GRAMMAR-Lambda required only 4.7hr to analyze single binary phenotype by sampling 5,000 SNPs to calculate GRM, while SAIGE did about 534 hr [15]. A user friendly GRL-Binary software was developed, which is freely available at https://github.com/RunKingProgram/ BinaryGRAMMAR-Lambda.

Declarations

Acknowledgements

The research is financially supported by the National Key R&D Program of China

(2018YFD0900201) and the National Natural Science Foundations of China (32072726).

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 Genetic, Edn. 2ed. (Longman, London; 1981).
3. McCullagh, J. Generalized linear models. *New York* (1989).
4. Wedderburn Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447 (1974).
5. Zaitlen, N. et al. Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet* **8**, e1003032 (2012).
6. Zaitlen, N. et al. Analysis of case-control association studies with known risk variants. *Bioinformatics* **28**, 1729-1737 (2012).
7. Joel, M. & Witte, J.S. The Covariate's Dilemma. *Plos Genetics* **8**, e1003096 (2012).
8. Breslow, N.E. & Clayton, D.G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9-25 (1993).
9. Schall, R. Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727 (1991).

10. Sorenrsen, D. & Gianola, D. Likelihood, Bayesian, and MCMC methods in quantitative genetics. (2002).
11. 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).
12. 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 (2011).
13. Hayeck, T. et al. Mixed Model with Correction for Case-Control Ascertainment Increases Association Power. *American Journal of Human Genetics* **96**, 720 (2015).
14. 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 (2016).
15. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics* **50**, 1335-1341 (2018).
16. Kang, H.M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-354 (2010).
17. Loh, P.R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284-290 (2015).
18. 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 (2007).
19. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575 (2007).
20. 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).
21. Hoerl, A.E. & Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55-67 (1970).
22. Jennifer, L. et al. Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525-526 (2012).

Figures

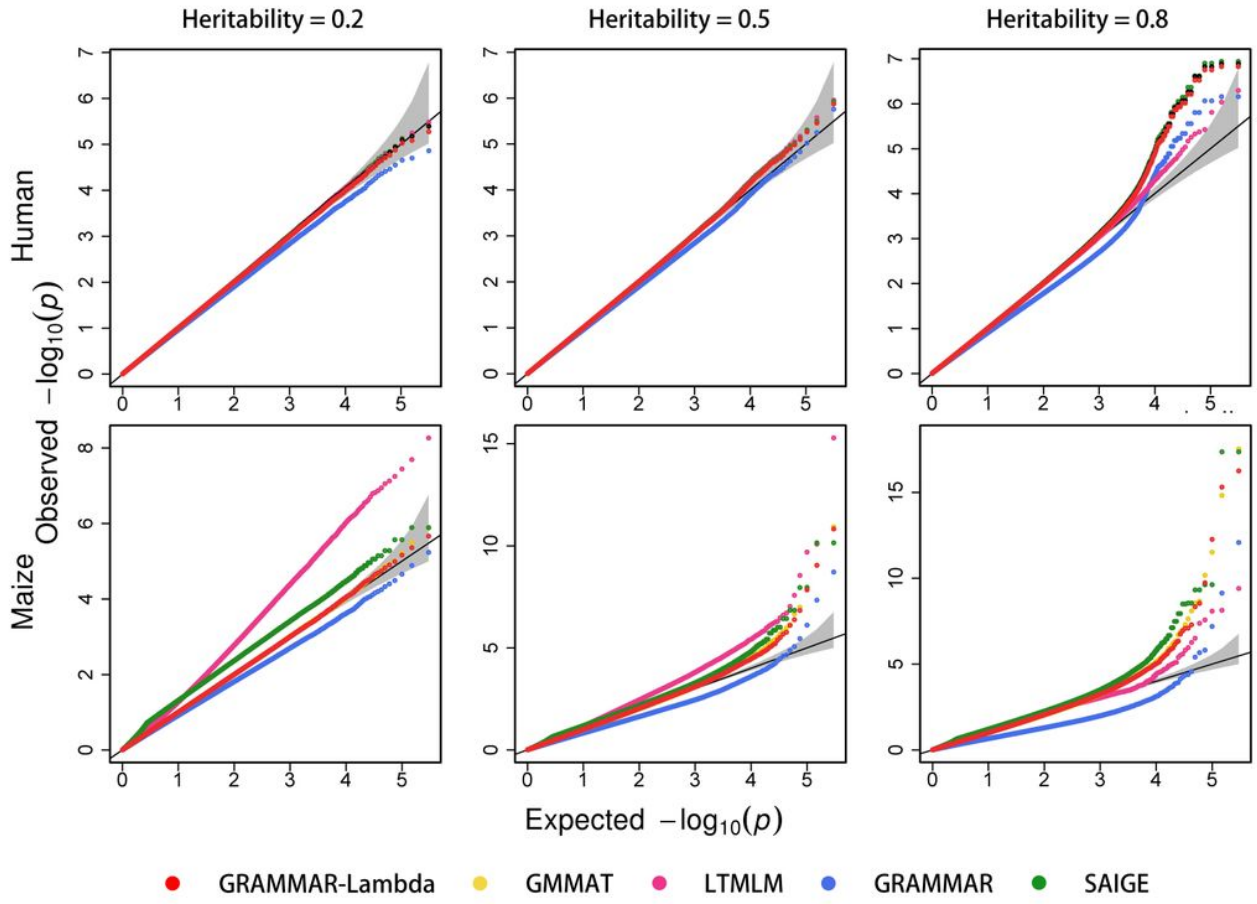

Figure 1

Comparison in the Q-Q profiles between GRAMMAR-Lambda 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.

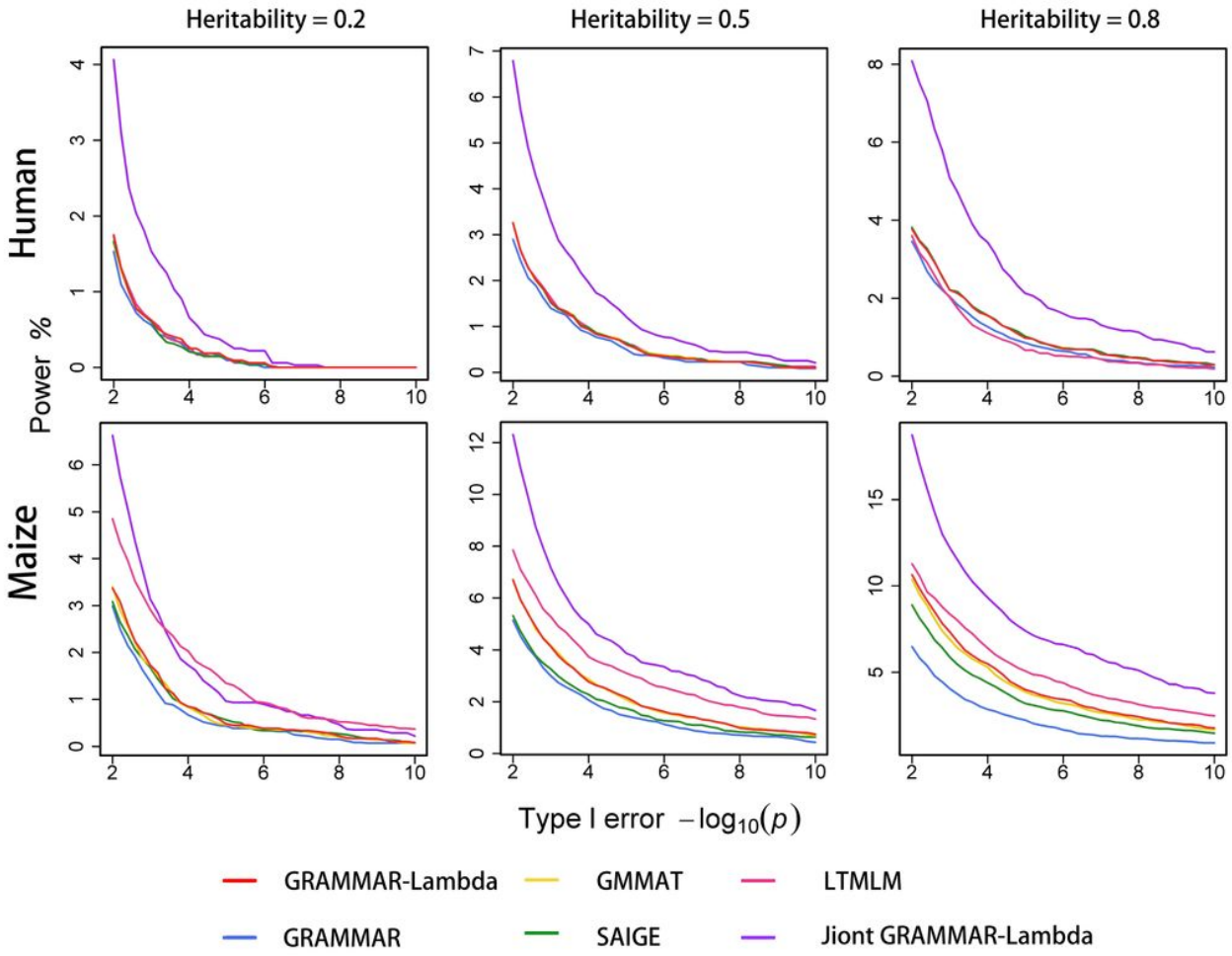

Figure 2

Comparison in the ROC profiles between GRAMMAR-Lambda 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.

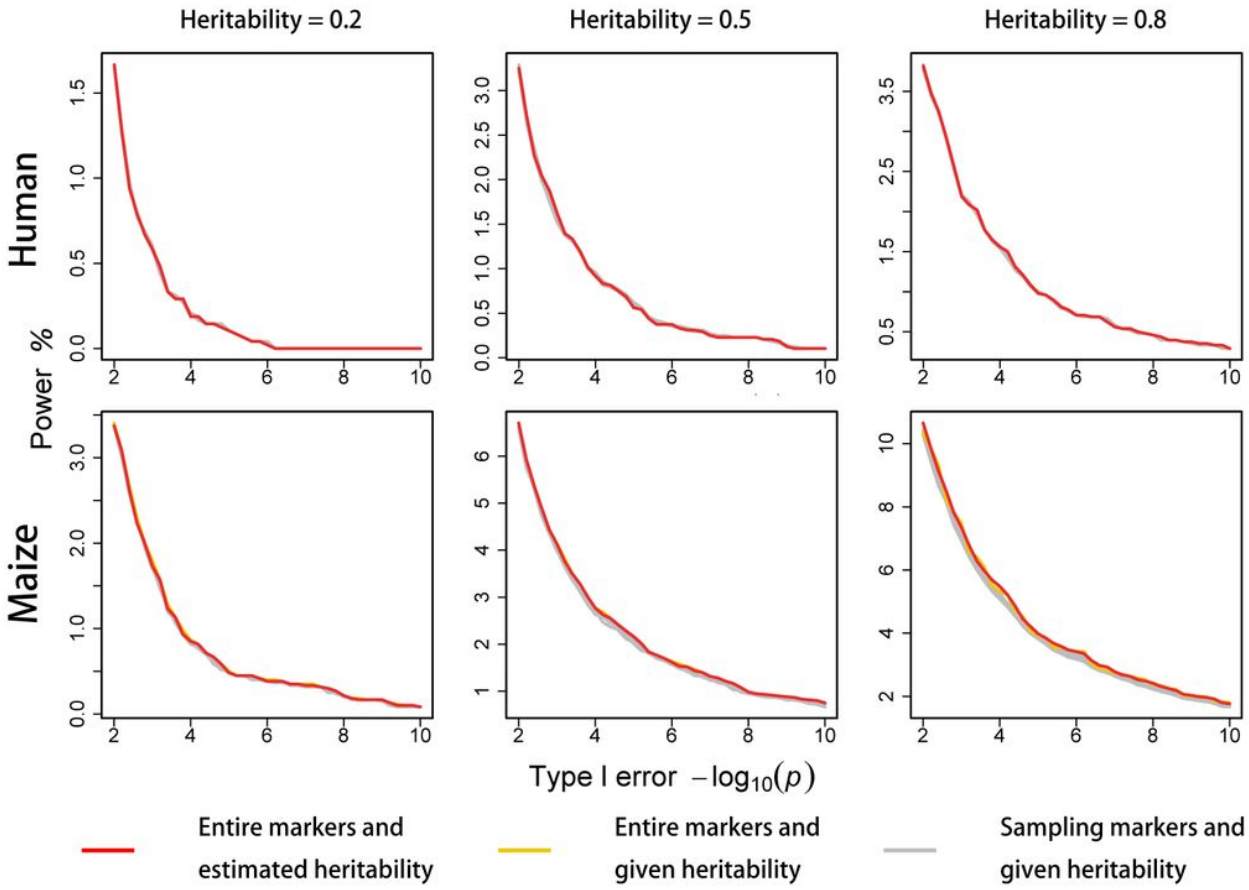

Figure 3

Sensitivity of statistical powers to the specified heritabilities for GRAMMAR-Lambda. Statistical powers are dynamically evaluated with the ROC profiles. The simulated phenotypes are controlled by 200 QTNs with the low, moderate and high heritabilities in human and maize.

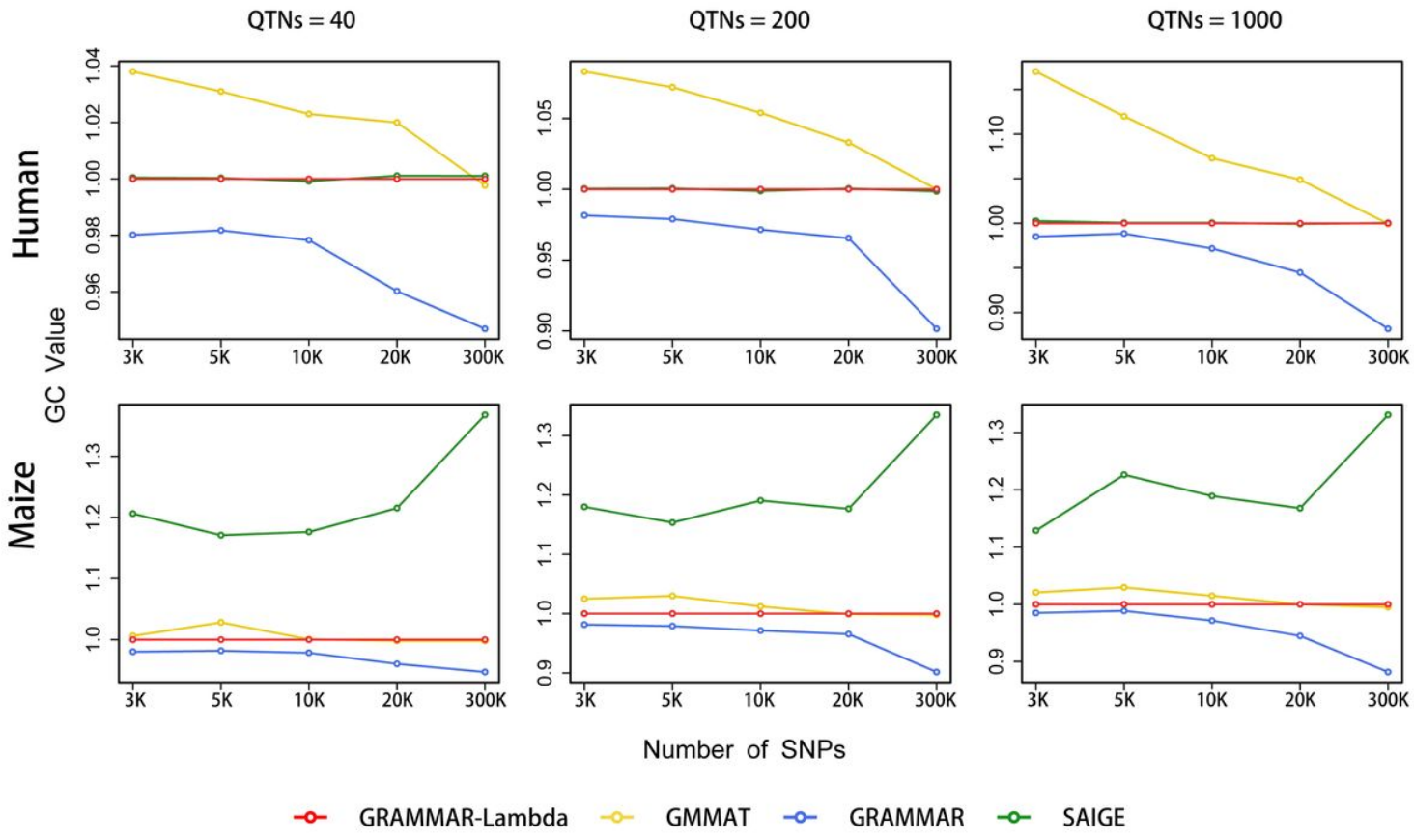

Figure 4

Changes in genomic controls with the number of sampling SNPs for GRAMMAR-Lambda and the four competing methods. Genomic control (GC) 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineMethods.docx