Genome-wide Hierarchical Mixed Model Association Analysis

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

Di Liu (liudi@haas.cn)
College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

Zhiyu Hao
Institute of Animal husbandry, Heilongjiang Academy of Agricultural Sciences, Harbin 150086, China

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

Di Liu
College of Animal Scientific Technology, Northeast Agricultural University, Harbin 150030, China

Method Article

Keywords: Genome-wide association analysis, Hierarchical mixed model, Genomic breeding value, Joint association analysis, Statistical power

Posted Date: March 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-315869/v1

Abstract

We partitioned the genomic mixed model into two hierarchies to firstly estimate genomic breeding values (GBVs) using the genomic best linear unbiased prediction and then statistically infer the association of GBVs with each SNP using the generalized least square. The genome-wide hierarchical mixed model association study (named Hi-LMM) can correct effectively confounders with polygenic effects as residuals in association tests, preventing potential false negative errors produced with GRAMMAR or EMMAX. The Hi-LMM performs the same statistical power as the exact FaST-LMM with the same computing efficiency as EMMAX. When the GBVs have been estimated precisely, Hi-LMM outperforms existing methods in statistical power, especially through joint association analysis.

Introduction

In genome-wide association study (GWAS), it is important to dissect the confounding biases caused by population structures and cryptic relatedness. Linear mixed models (LMMs) [1,2] have the ability to separate true signals from a vast number of false signals caused by confounders, improving statistical power to detect quantitative trait nucleotides (QTNs). When applying an LMM to GWAS [3], the variance components or the polygenic effects in the LMM need to be estimated using a genome relationship matrix (GRM) [4], excluding the single nucleotide polymorphisms (SNPs) that are going to be tested, before the association tests are conducted. Even though using all markers to estimate variance components or polygenic effects, without repeatedly calculating the GRMs for each SNP, LMMs are much more computationally intensive at nonlinearly solving different variance components among high-throughput SNPs.

In initial genome-wide mixed model association studies, variance components were generally estimated using the maximum likelihood (ML) or restricted maximum likelihood (REML) methods [5], which have been implemented in various numerical optimization algorithms [3]. To reduce computationally expensive matrix operations at each iteration, EMMA [6], GEMMA [7], and FaST-LMM [8] use a single eigendecomposition of a GRM to rotate data. BOLT-LMM [9] introduces the Monte Carlo REML method [10,11] to estimate variance components, which only requires the solutions of linear mixed model equations. The H-E regression [12,13], which is another variance component method, is able to estimate polygenic and residual variances by linearly regressing the product of the phenotypes on the off-diagonal elements of the GRM in the most straightforward way [14]. Other than that, CMLM [15] and fastGWA [16] are more appropriate for accelerating the estimations of variance components in the stratified population or in the population with sparse GRMs. As approximations of the mixed model association analysis, two popular simplified algorithms such as EMMAX [17] or P3D [15] and GRAMMAR [18] attempt to replace different variance components and polygenic effects among candidate markers with the same variance components and genomic breeding values (GBVs), respectively, that were estimated under the null LMM, which greatly saves computing costs. In particular, GRAMMAR-Gamma [19] and BOLT-LMM [9] improve the statistical power that can be used to detect QTNs by calibrating GRAMMAR.

In the case that QTNs exist, over-estimation of polygenic variances and effects by genomic variances and GBVs may cause GRAMMAR and EMMAX to produce potential false negative errors. In this study, we divide genomic mixed model into two hierarchies: the LMM for the phenotypes of GBVs and the linear regression model of GBVs on the tested SNPs. Based on resulting the hierarchical mixed model, we firstly estimate GBVs using the genomic best linear unbiased prediction (GBLUP) method [4,20] and then statistically infer the genetic effects of each SNP using the generalized least square (GLS) method, regarding the GBVs with GRMs as "phenotypes." Especially in the linear regression model, the genetic effects for the tested SNPs were excluded from the polygenic effects or variance as the residuals to prevent the over-estimation of polygenic effects or variances by EMMAX and GRAMMAR. The utility of the genome-wide hierarchical mixed model association analysis is demonstrated by computer simulations and real data analysis.

Results

Statistical property of the Hi-LMM

The association results were obtained with the five competing methods and the Hi-LMM, a test at once, which are displayed selectively in Figure 1 for Q-Q profiles and Figure 2 for ROC profiles (Figure 1S and Figure 2S in detail). At the same time, genomic control values were recorded in Table 1S and Table 2S. Under genomic controls very close to 1.0, Hi-LMM performed almost the same statistical power to detect QTNs as FaST-LMM and EMMAX, regardless of how many QTNs and heritabilities were simulated. With the null model that had no QTNs, both FaST-LMM and EMMAX yielded slightly higher negative false rates than Hi-LMM and the negative false rates increased with the complexity of population structures, especially for EMMAX. Among the five competing methods, GRAMMAR had the lowest genomic controls and statistical power, and it more strongly deflated test statistics in more complex population structure. Although GRAMMAR-Gamma and BOLT-LMM genome-widely corrected the test statistics of GRAMMAR by their defined calibration factors [9,19], GRAMMAR-Gamma slightly deflated the test statistics for complex population structure, which generated a genomic control of below 1.0, and BOLT-LMM strongly increased the positive false rate due to overcorrecting test statistics.

Further, Hi-LMM jointly analyzed multiple QTN candidates chosen from one association test at a time, given a significance level of 0.05. For convenience to compare, we depicted the statistical powers obtained with joint analyses and those with a test at once together. Using a backward regression analysis, Hi-LMM evidently increased statistical power. In comparison, BOLT-LMM also significantly increased statistical power, but it did not control false positive errors in detecting QTNs, especially for complex population structure.

Sensitivities to estimate genomic heritability or GBVs

In the competing methods, EMMAX estimates genomic heritability to replace polygenic heritabilities, whereas GRAMMAR, GRAMMAR-Gamma, and BOLT-LMM estimate GBVs to replace polygenic effects.

Similarity, Hi-LMM also estimates genomic heritability and GBVs to associate with markers. What are the sensitivities of Hi-LMM, EMMAX, GRAMMAR, GRAMMAR-Gamma, and BOLT-LMM to estimate genomic heritability or GBVs?

Regarding the genomic heritability or GBVs simulated as polygenic estimates, we analyzed the simulated phenotypes with the five methods and the Hi-LMM, a test at once. As shown in Figure 3 and Figure 3S, Hi-LMM, one test at a time could achieve more highest statistical power under the more ideal genomic controls than joint association analysis, if genomic heritability or breeding values were completely accurately estimated. In contrast, EMMAX had somewhat decreases in both statistical power and genomic control. Additionally, GRAMMAR, GRAMMAR-Gamma, and BOLT-LMM did not find any QTNs from the residuals of GBLUP.

Instead of GBLUP, we adopted a Lasso technique implemented in R/glmnet [21] to rapidly estimate GBVs. Through association tests of the Hi-LMM, we also drawn corresponding ROC and Q-Q profiles in Figure 3 and Figure 3S, respectively. As could be seen, Hi-LMM achieved higher statistical power with the Lasso technique than GBLUP, and the tendency to improve the statistical power is consistent with that of the simulated GBVs. With selecting ridge estimation [22] in R/glmnet, we demonstrated that Hi-LMM also gained a statistical power as high as that GBLUP did. In conclusion, Hi-LMM could improve statistical powers by precisely estimating genomic heritability or breeding values, compared with EMMAX, GRAMMAR, GRAMMAR-Gamma, and BOLT-LMM.

Calculation of the GRM with the sampling markers

With GBLUP, estimation of genomic heritability and GBVs mainly depends on the density of markers used to calculate the GRMs in the structured population [4,23]. To improve computing efficiency, FaST-LMM, GRAMMAR-Gamma and BOLT-LMM sampled or screened a small proportion of the whole genomic SNPs to estimate the GBVs or genomic heritability as precisely as possible. Based on this, we also try to simplify the computation of Hi-LMM by sampling markers.

Figure 4 shows that the changes in genomic controls Hi-LMMs and four competing methods made by a test at once with numbers of sampling markers. Similar to FaST-LMM, EMMAX, and GRAMMAR-Gamma, Hi-LMM gradually controlled positive false errors as the number of sampling marker increased and yielded high statistical power using all genomic markers. GRAMMAR seemed to calibrate negative false rates by under-estimating GBVs with less markers, whereas BOLT-LMM produced serious positive false errors caused by inflating test statistics, regardless of how many the markers were drawn. To retain the ideal genomic control and statistical power to detect QTNs (See Figure 4S and 5S), Hi-LMM needed to draw no fewer than 40,000 markers to estimate GRMs in the simulation.

Real data analyses

Using previously published datasets on *Arabidopsis thaliana*, mice, and maize, we illustrated both genomic control and QTN mapping with Hi-LMM and compared our findings to those obtained using

FaST-LMM, GRAMMAR, GRAMMAR-Gamma, and BOLT-LMM. Using a visual test for normality, we selected 32 phenotypes with less than 120 records for GWAS in *Arabidopsis thaliana* and 109 phenotypes in mice. We did not record the computing times for these two datasets because either population size or the number of markers is enough to significantly differentiate these competing methods.

For all competing methods, we depicted the Q-Q and Manhattan profiles for the traits of detectable QTNs in Supplementary Figure 6S for *Arabidopsis thaliana*, Figure 7S for mice, and Figure 8S for maize. GRAMMAR, GRAMMAR-Gamma, and BOLT-LMM achieved almost the same statistical properties as they did in the simulations. GRAMMAR detected several QTNs with strong false negative errors, whereas with higher false positive errors, BOLT-LMM found more QTNs than the other competing methods but fewer QTNs than Hi-LMM did with joint analysis. Even though a test at once, Hi-LMM could identify more QTNs than GRAMMAR-Gamma. Using Hi-LMM with joint analyses, we found QTNs from 21 of 32 phenotypes in *Arabidopsis thaliana*, and 104 of 109 phenotypes in mice. With FaST-LMM, however, QTNs were not found for 1/21 and 51/104 of the traits, in *Arabidopsis thaliana* and mice, respectively. Under the ideal genomic control condition, Hi-LMM identified 19 and 94 more QTNs with joint analysis than exact FaST-LMM in *Arabidopsis thaliana* and mice, respectively. Moreover, for phenotypes in mice, Hi-LMM could cover 94% of the QTNs obtained using FaST-LMM, while approximately 72% for the traits analyzed in *Arabidopsis thaliana*, wherein FaST-LMM arose unstable genomic controls.

Finally, we applied the Hi-LMM to map the QTNs for flowering time and simultaneously executed the method using 50,000 SNPs randomly drawn from high-throughput markers. By a test at once, Hi-LMM detected 6 QTNs distributed on chromosomes 1, 2, 3, 8, and 10, and covered 3 of 4 QTNs on chromosomes 3, 8, and 10 detected using exact FaST-LMM. Further, Hi-LMM found the same QTNs using sampling markers besides the two QTNs located on chromosomes 5 and 2 detected using entire genomic markers. Joint analysis can separate all signals that correspond to the QTN candidates generated from a test at once, which improves the statistical power to detect QTNs and the comparability by sampling markers. Upon inputting the genotypes and phenotypes to obtain QTN mapping outputs, Hi-LMM ran for 3.200 and 1.900 mins in R software, respectively, for entire and sampling markers, whereas GRAMMAR and GRAMMAR-Gamma took 2.073 and 3.637 mins, respectively. Additionally, FaST-LMM ran for 32.147 mins in Single-Runking [24] and BOLT-LMM for 166.448 mins in BOLT [9]. 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.

Discussion

With the GBVs that included the genotypic effects of all the candidate markers, we stratified the genomic mixed model into the mixed model of random GBVs and the generalized linear regression model of the correlated GBVs to the tested markers. In contrast to GRAMMAR and EMMAX, which overestimated polygenic effects and variances with GBVs and genomic heritability, respectively, the Hi-LMM best and unbiasedly estimated polygenic effects by regressing the GBVs on candidate markers in association tests. As a result, it can avoid potential false negative errors, and achieve statistical power as high as the

exact mixed model association analysis. Theoretically, it has the same computational complexity as EMMAX, because EMMAX also uses GLS for association tests.

To improve statistical power to detect QTNs for economic traits in plant and animal, peoples always replaced the phenotypic values with the estimated in advance breeding values (EBVs) by pedigree or genomic markers. If the association of EBVs with candidate markers was statistically inferred by using a simple linear regression model rather than a generalized regression model, then it would produce higher false positive rates than that for phenotypes, especially in breeding populations with complex structures (simulations not shown). Exact and simplified mixed model association analyses for EBVs not only repeatedly estimate the heritability and breeding values, but also enhance computational complexity. Once the EBVs of traits have been provided before a GWAS, we recommend to efficiently use GLS method.

Under the assumption of minor polygenes, GBLUP is inappropriate to accurately estimate GBVs for quantitative traits controlled by less major genes. Meanwhile, it requires to estimate genomic heritability in advance, which increases computational complexity. For genomic selection, many methods, such as a series of Bayesian methods [25], can estimate GBVs by specifying various priors for markers effects, without estimate genomic heritability. Our simulations demonstrate that the statistical power to detect QTNs can be significantly improved as long as the GBVs have been accurately estimated. Therefore, highly efficient genomic selection methods play a critical role in achieving performance of the Hi-LMM.

With an increasing number of high-throughput SNP markers genotyped by deep resequencing, we can implement the GLS method at the second hierarchy of the Hi-LMM in a straightforward manner to finely map QTNs. This is because the GBVs obtained from previous GWAS in the same population are enough to ensure statistical power of the Hi-LMM. Once the GBVs for multiple correlated quantitative traits have been more accurately pre-estimated, the Hi-LMM can be easily extended to efficiently map pleiotropic QTNs within the framework of multivariate regression. For GWAS on dynamic quantitative traits, genomic random regression models can be divided into the three hierarchies: random regression model with individuals' dynamic trajectories, multivariate animal model for the parameters in dynamic trajectories, and generalized multivariate regression model for the GBVs of the parameters in dynamic trajectories, which would greatly improve computing efficiency. If the GBVs can be estimated once with a generalized linear mixed model, then the Hi-LMM is highly suited for binary disease traits because of the resulting normal distributions of the GBVs.

Declarations

Code availability

Hi-LMM software is 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 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, doi:10.1038/ng1702 (2006).

2 Henderson, C. R. *Applications of linear models in animal breeding*. (University of Guelph, 1984).

3 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, doi:10.1038/ng.2876 (2014).

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

5 Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika***58**, 545-554 (1971).

6 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).

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

8 Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods***8**, 833-835, doi:10.1038/nmeth.1681 (2011).

9 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).

10 García-Cortés, L. A. *et al.* Variance component estimation by resampling. *Journal of Animal Breeding and Genetics***109**, 358-363 (1992).

11 Matilainen, K., Mantysaari, E. A., Lidauer, M. H., Stranden, I. & Thompson, R. Employing a Monte Carlo algorithm in Newton-type methods for restricted maximum likelihood estimation of genetic parameters. *PLoS One***8**, e80821, doi:10.1371/journal.pone.0080821 (2013).

12 Chen, G.-B. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Frontiers in Genetics***5**, 1-14, doi:10.3389/fgene.2014.00107 (2014).

13 Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics***2**, 3-19 (1972).

14 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).

15 Zhang, Z. W. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature genetics***42**, 355-360, doi:10.1038/ng.546 (2010).

16 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).

17 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).

18 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, doi:10.1534/genetics.107.075614 (2007).

19 Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nature genetics***44**, 1166-1170, doi:10.1038/ng.2410 (2012).

20 Habier, D., Fernando, R. L. & Dekkers, J. C. The impact of genetic relationship information on genome-assisted breeding values. *Genetics***177**, 2389-2397, doi:10.1534/genetics.107.081190 (2007).

21 Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software***33**, 1-22 (2010).

22 Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics***12**, 55-67 (1970).

23 Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature genetics***42**, 565-569, doi:10.1038/ng.608 (2010).

24 Gao, J., Zhou, X., Hao, Z., Jiang, L. & Yang, R. Genome-wide barebones regression scan for mixed-model association analysis. *Theor Appl Genet*, doi:10.1007/s00122-019-03439-5 (2019).

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

Figures

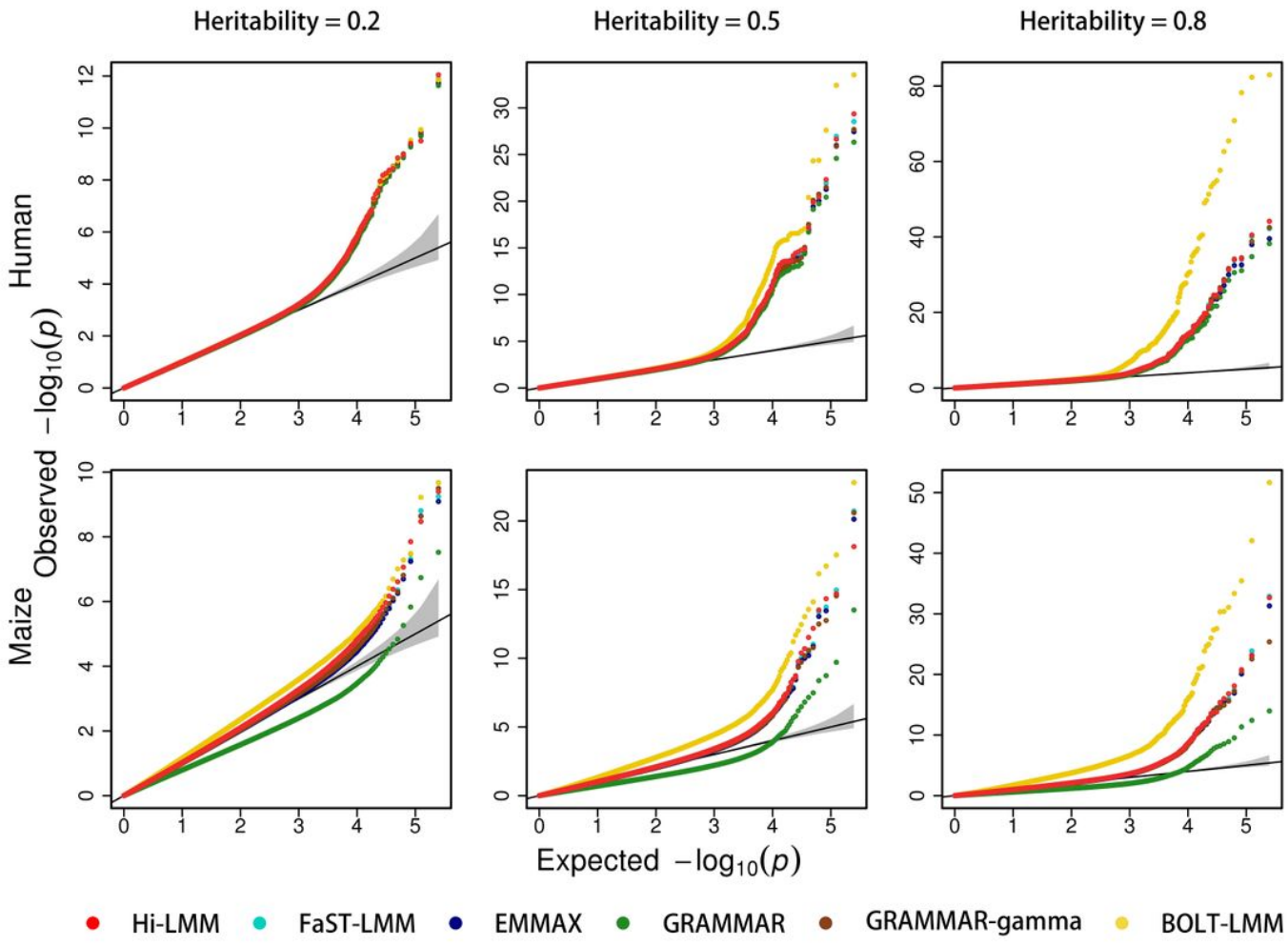

Figure 1

Comparison of Hi-LMM with the five 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.

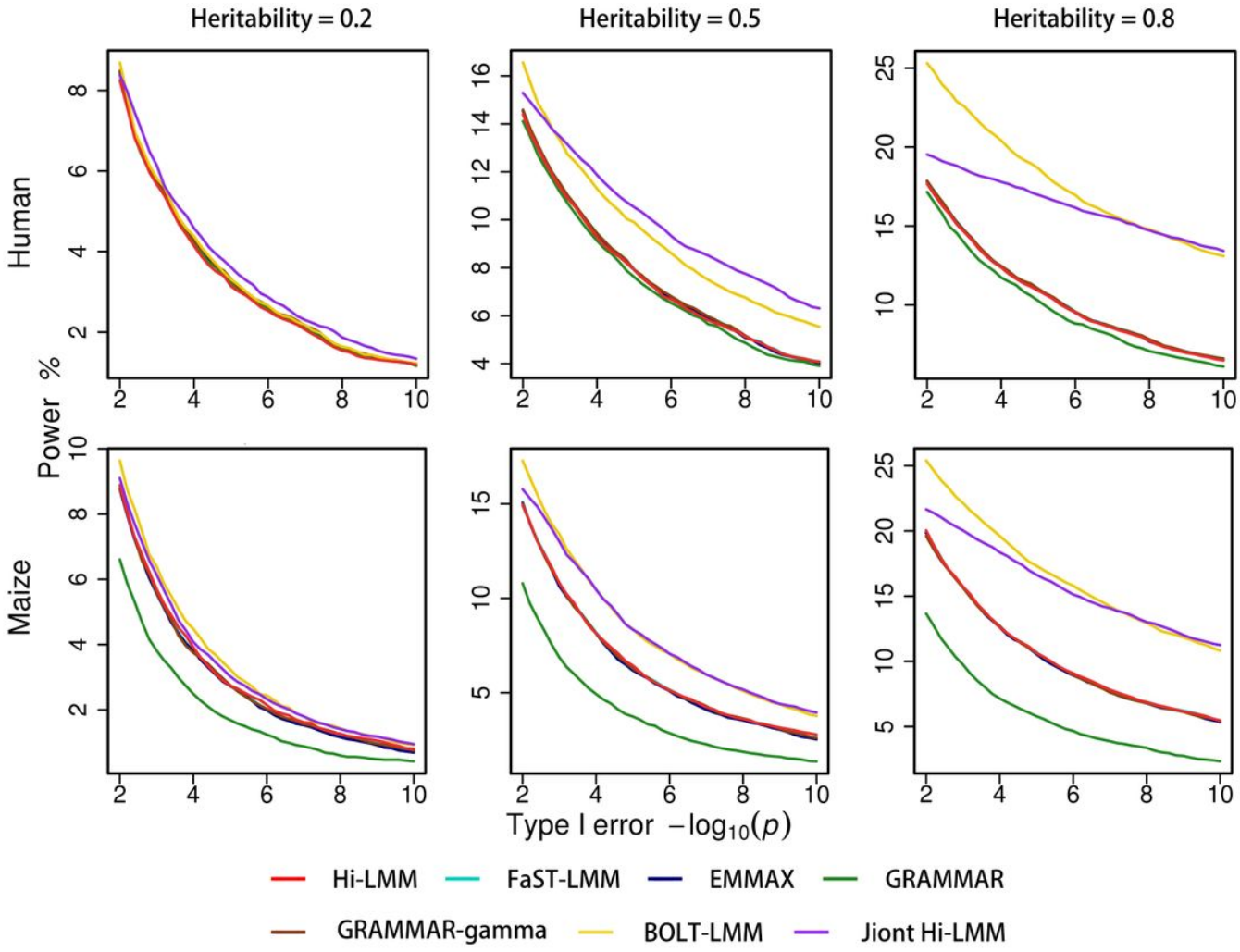

Figure 2

Comparison of Hi-LMM with the five 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.

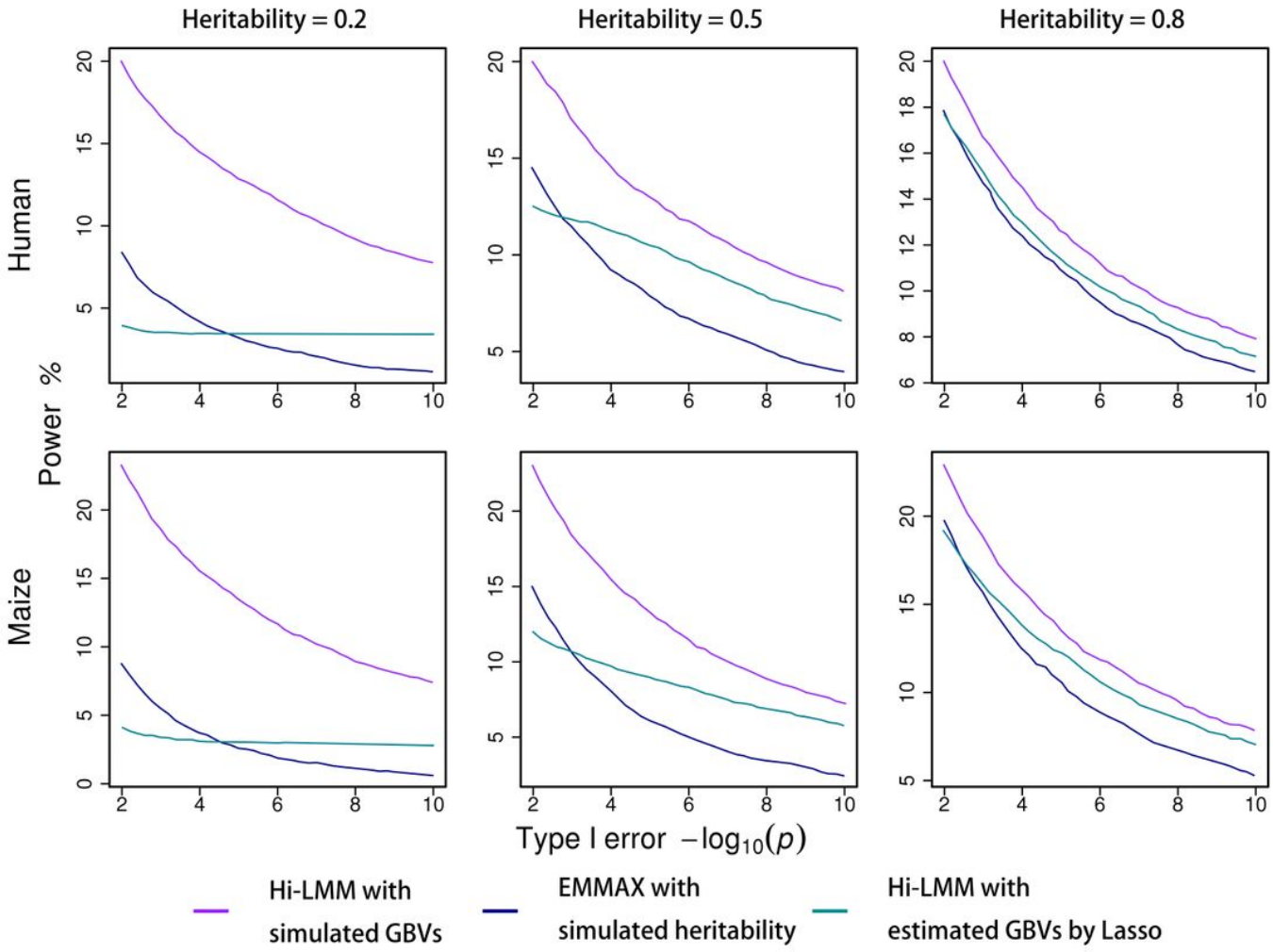

Figure 3

Sensitivity of statistical powers to estimate heritabilities or GBVs for Hi-LMM. Statistical powers are dynamically evaluated with the ROC profiles. Both GRAMMAR-Gamma and BOLT-LMM 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.

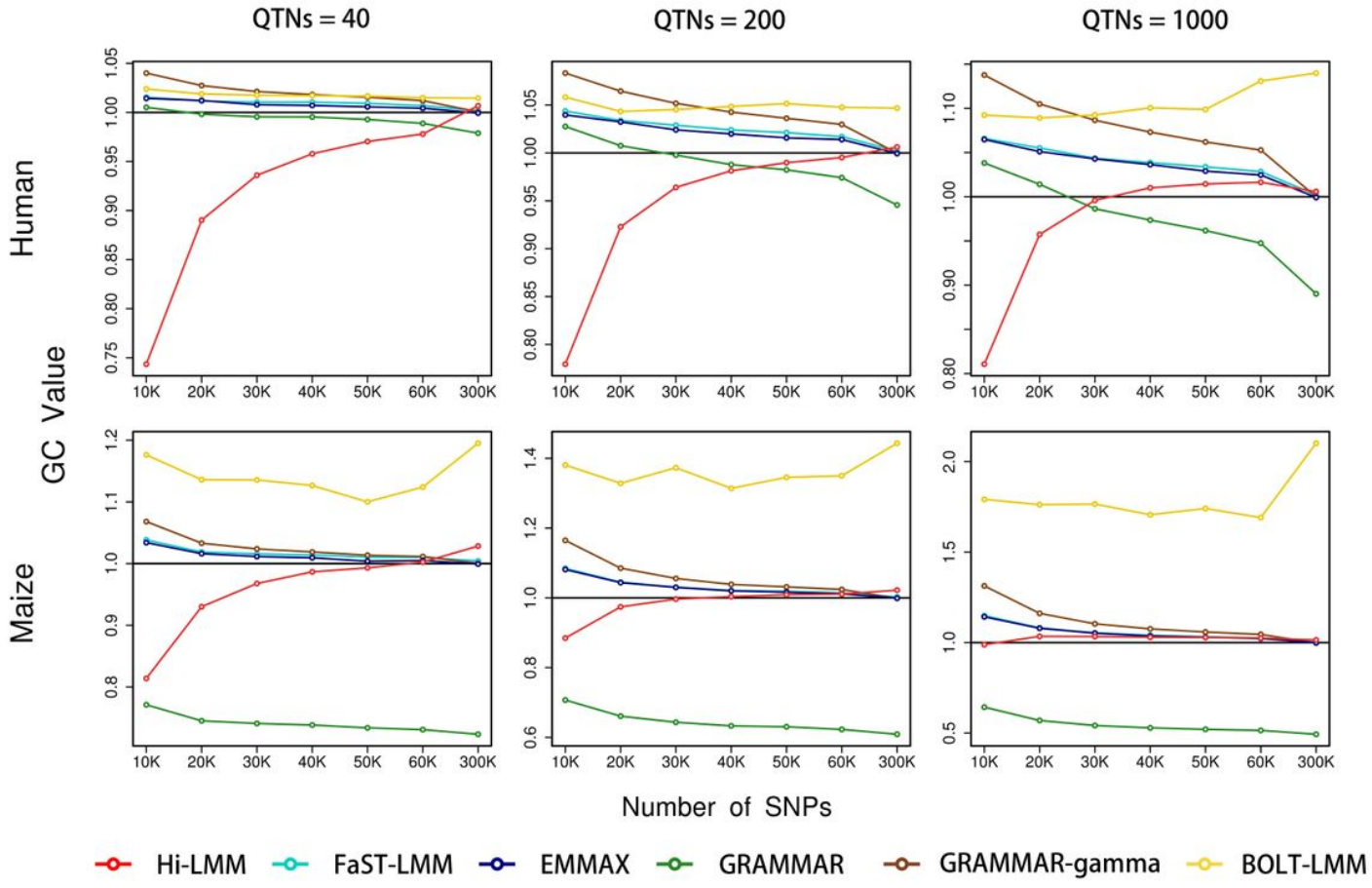

Figure 4

Changes in genomic controls with the number of sampling SNPs for Hi-LMM and the five competing methods. Genomic control is calculated by averaging genome-wide test statistics. The simulated phenotypes are controlled by 40, 200 and 1,000 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.

- SupplementaryfilesbinaryGRL.docx