GRAMMAR-Lambda: An Extreme Simplification for Genome-wide Mixed Model Association Analysis

Runqing Yang[1, 2], Jin Gao[1, 2], Yuxin Song[1, 2], Zhiyu Hao[3] and Pao Xu[1]

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

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

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

Corresponding Authors: Runqing Yang (runqingyang@cafs.ac.cn)

Pao Xu (xup@ffrc.cn)

Abstract

A highly efficient genome-wide association method, GRAMMAR-Lambda is proposed to make simple genomic control for the test statistics deflated by GRAMMAR, producing statistical power as high as exact mixed model association method. Using the simulated and real phenotypes, we show that at a moderate or above genomic heritability, polygenic effects can be estimated using a small number of randomly selected markers, which extremely simplify genome-wide association analysis with an approximate computational complexity to naïve method in large-scale complex population. Upon a test at once, joint association analysis offers significant increase in statistical power over existing methods.

Introduction

Population stratification, family structure, and cryptic relativeness, as widespread confounding factors, inflate test statistics in genome-wide association studies (GWAS), which increases the false-positive rate of association tests. The extent of this inflation is measured using genomic control [1]. The most straightforward approach to correct these confounding effects is to divide genome-wide test statistics by genomic control calculated as the average or median genome-wide test statistics [2-4]. Principal component analysis and family-based association tests are applied to correct population stratification [5-10] and family structure [11-14], respectively. In comparison, the

linear mixed-model (LMM) [15,16] considers not only population stratification and family structure, but also cryptic relatedness through known kinship, offering a practical and comprehensive genomic control approach.

Even though the genomic relationship matrix (GRM) [17], calculated with whole genomic markers, is used to estimate corresponding polygenic variances to candidate markers, the LMMs are much more computation-intensive for high-throughput single nucleotide polymorphisms (SNPs). Several simplified algorithms for mixed model association have been developed to improve computational efficiency. Assuming that a quantitative trait nucleotide (QTN) has a minor variance, both EMMAX [18] and P3D [19] replaced different polygenic variances with the same genomic variance estimated by null LMM. Subsequently, FaST-LMM [20] and GEMMA [21] use a single eigen decomposition of the GRM [22] to reduce the estimation of polygenic or genomic variances. After moving the polygenic effects estimated by the pedigree-based best linear unbiased prediction (PBLUP) [15] from phenotypes, initial GRAMMAR [23] associates each SNP with phenotype residuals. Instead of the PBLUP, however, GRM-based best linear unbiased prediction (GBLUP) [24,25] increases GRAMMAR's false-negative error rate because of accurately estimating genomic breeding values (GBVs) [17,24,25]. Therefore, GRAMMAR-Gamma [26] and BOLT-LMM [27] have been proposed to improve the statistical power of QTN detection by calibrating the deflated test statistics by the GRAMMAR.

Before association tests, all LMM-based methods are required to calculate GRM and estimate variance components. Using these estimated components, GRAMMAR determines the least computing time for association tests, $O(nm)$, where n is the population size and m is the number of markers. Based on the GRAMMAR, BOLT-LMM estimates variance components with Monte Carlo REML [28,29] only in the computing complexity of $O(inm)$ [27], where i is iterative times. For large-scale data of humans, a resource-efficient tool for mixed-model association analysis has been developed to enhance computing efficiency of GRAMMAR-Gamma significantly, via calculating sparse GRM [30].

In this study, we propose a GRAMMAR-Lambda method to adjust GRAMMAR using genomic control, extremely simplifying genome-wide mixed model analysis. We show, using extensive simulations, that for a complex population structure, a high false-negative error of GRAMMAR can be efficiently corrected by dividing genome-wide test statistics by genomic control. Statistical power is kept constant,

almost the same as that of exact FaST-LMM, even with randomly selected markers across the whole genome. In estimating GBVs with GBLUP, variance components or genomic heritability do not require to be estimated in advance. Within the framework of the GRAMMAR-Lambda, the joint analysis for QTN candidates significantly improves the statistical power, as compared to a test at once.

Results

Statistical property of GRAMMAR-Lambda

GRAMMAR-Lambda, a test at once was comparable to the four competing methods, such as FaST-LMM, GRAMMAR, GRAMMAR-Gamma and BOLT-LMM. Association results from the simulation (1) are displayed selectively in Figure 1 for Q-Q profiles and Figure 2 for ROC profiles, and genomic control values are calculated in Table 1S. In general, GRAMMAR-Lambda showed stable and good statistical properties that did not depend on how many the simulated QTNs were responsible for quantitative traits and how complex the population structures were. Among GRAMMAR-Lambda and four competing methods, GRAMMAR had the lowest genomic controls and statistical powers. At the same time, it showed more complex population structures and more strong deflation of genome-wide test statistics. Under perfect genomic control of exact 1, GRAMMAR-Lambda could detect QTNs with the slightly high statistical powers, compared to FaST-LMM and GRAMMAR-Gamma. Although GRAMMAR-Gamma and BOLT-LMM achieved genomic control for GRAMMAR via different calibration factors 26,27, GRAMMAR-Gamma a little deflates test statistics for complex maize population, making the genomic control below 1.0. On the contrary, BOLT-LMM inflates test statistics, so that increases the false-positive rate as the complexity of population structure increased.

After genomic control for GRAMMAR, GRAMMAR-Lambda jointly analysed multiple QTN candidates chosen with a test at once, at a significance level of 0.05. For convenience to compare, the statistical powers with joint association analyses were depicted along with those using a test at once. Though backward multiple regression analysis, GRAMMAR-Lambda gained significant increase in statistical powers. Although BOLT-LMM also showed an increase in the statistical power to detect QTNs, it did not control the false-positive error, especially for the complex maize population.

Sensitivity to estimate genomic heritability

All LMM-based GWASs, except for GRAMMAR, required the exact estimation of polygenic variances or effects. Since GRAMMAR-Lambda efficiently controlled the deflation of test statistics caused by overestimation of polygenic effects, we hypothesised that with GBLUP, the appropriate estimation or overestimation of polygenic effects could be produced by pre-setting a genomic heritability value greater than the actual genomic heritability or no less than 50% of it. Note that the given lower bound of genomic heritability was obtained via our extensive simulations for optimising GRAMMAR (not shown).

With the specified genomic heritabilities of 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.8, 0.9 and 0.95, we analysed the phenotypes from the simulation (2) using only GRAMMAR-Lambda to examine the sensitivity to estimate genomic heritability. Figure 3 compared statistical powers generated from the estimated genomic heritabilities with those from the specified genomic heritabilities. As shown in each plot for human population, most ROC curves of the estimated genomic heritabilities almost overlapped with those of pre-specified genomic heritabilities. The maximum thickness of each overlapped curve was less than one QTN converted by the difference in the statistical power. For the simulated phenotypes in maize, however, the decrease in statistical powers was observed at very low or high genomic heritabilities simulated. Despite this, GRAMMAR-Lambda could retain high statistical powers with perfect genomic controls (see Figure 3S), when genomic heritability was chosen from any value within a broader interval to estimate GBVs. Thus, there was no need to estimate variance components or genomic heritability. When implementing GRAMMAR-Lambda, genomic heritability was set to 0.5 by default or empirical heritability of traits if available.

Calculation of GRMs with the sampling markers

Estimation of genomic heritability and GBVs using GBLUP mainly depends on the density of markers used to calculate the GRMs in the structured population 17,31. To improve computing efficiency, several simplified algorithms such as 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. To underestimate the GBVs, GRAMMAR-Lambda would require fewer sampling markers than the competing methods, as demonstrated by simulation (3),

wherein GRAMMAR-Lambda pre-specified a genomic heritability of 0.5.

We randomly sampled 3,000, 5,000, 10,000, 15,000, 20,000, and 25,000 SNPs from entire genomic markers to calculate the GRMs. Figure 3 shows the changes in the genomic controls of GRAMMAR-Lambda and four competing methods with numbers of sampling markers. No competing method could control false-positive/negative errors with less than 50,000 sampling SNPs. Specifically, both FaST-LMM and GRAMMAR-Gamma gradually controlled false-positive errors with the increasing number of sampling markers. GRAMMAR calibrated the false-negative rate by under-estimating GBVs with fewer markers, while BOLT-LMM yielded a high rate of false-positive errors due to inflating test statistics, independent of where the markers were drawn from. In comparison, GRAMMAR-Lambda retained reasonable powers to detect QTNs through prefect genomic control, even with fewer than 3,000 sampling markers (see Figures 4S and 5S).

Application to real data

Previously published datasets of *Arabidopsis thaliana*, mice, and maize were used to illustrate GRAMMAR-Lambda and compare it with FaST-LMM, GRAMMAR, GRAMMAR-Gamma and BOLT-LMM in terms of QTN mapping and genomic control. Thirty-two phenotypes with greater than 120 records were selected for GWAS for *Arabidopsis thaliana* and 109 phenotypes for mice using the normality test. For the two datasets, we did not record computing times, because either a small population or a few markers were enough to make a distinction between these competing methods.

Using all compared methods, we depicted Q-Q and Manhattan profiles for the traits of detectable QTNs (Figures 6S–8S for *Arabidopsis thaliana*, mice, and maize, respectively). GRAMMAR, GRAMMAR-Gamma, and BOLT-LMM performed almost the same statistical properties as in simulations. GRAMMAR detected a handful QTNs with substantial false-negative errors, while BOLT-LMM detected more QTNs with higher false-positive errors than other competing methods but still fewer QTNs than GRAMMAR-Lambda with joint association analysis. Under reasonable genomic control, overall, GRAMMAR-Gamma identified fewer QTNs than GRAMMAR-Lambda with an association test at a time. In what follows, we focused on comparing the mapping results obtained with GRAMMAR-Lambda and

exact FaST-LMM.

As shown in the Q-Q profiles, GRAMMAR-Lambda with either one test at a time or joint analysis performed almost perfect genomic control in the exact or infinitely close to the averaged Chi-squared statistics of 1.0. Using GRAMMAR-Lambda with joint association analyses, we found QTNs in 21 of 32 phenotypes of *Arabidopsis thaliana* and 56 of 109 phenotypes of mice. However, FaST-LMM did not find any QTN in 1 and 5 of these 21 and 56 traits, respectively. Meanwhile, GRAMMAR-Lambda identified 32 and 242, respectively, more QTNs with joint association analyses than exact FaST-LMM. Even with one association test at a time, GRAMMAR-Lambda covered 90% of the QTNs detected with FaST-LMM for mice phenotypes, while about 60% for *Arabidopsis thaliana* phenotypes. Notably, FaST-LMM produced unstable genomic controls for the analysed traits in *Arabidopsis thaliana*.

To analyse the flowering time in maize, we implemented GRAMMAR-Lambda in estimating genomic heritability and GBVs together by using entire genomic markers or estimating only GBVs by randomly sampling 5,000 SNPs with an empirical heritability of 0.8. GRAMMAR-Lambda, one association test at a time detected the same QTNs as GRAMMAR-Gamma and completely covered 4 QTNs distributed on chromosomes 3, 6, 8, 10 identified with exact FaST-LMM. The GRAMMAR-Lambda revealed fewer QTNs with sampling markers than those with entire genomic markers by a two-QTN difference, but joint association analysis could separate more signals corresponding to QTNs detected with either entire or sampling markers than a test at once. From the input of genotypes and phenotypes to the output of QTN mapping, the GWAS for flowering time consumed computing time of 2.073, 3.637, 0.101, and 0.058 min with GRAMMAR, GRAMMAR-Gamma, and the GRAMMAR-Lambdas by using entire and sampling markers. Additionally, FaST-LMM ran for 32.147 min in Single-RunKing [32] and BOLT-LMM for 166.448 min in BOLT [27].

Discussion

The modified GRAMMARs, GRAMMAR-Gamma, and BOLT-LMM have the same computing complexity for association tests as GRAMMAR-Lambda, but both methods require estimated variance components or polygenic heritability. To prevent false-positive errors, GRAMMAR-Gamma is not used to calculate the GRM with fewer selected markers. Although BOLT-LMM estimates the calibration factor with

only 30 pseudorandom SNPs in human datasets [27], it produces a high false-positive error rate in a complex structured population. Extensive simulations demonstrated the following: 1) In estimating GBVs, variance components or genomic heritability could be empirically specified over a broad range, without estimation in advance; 2) even using fewer randomly drawn markers from the whole genome to calculate the GRM, GRAMMAR-Lambda had almost the same statistical power as the exact mixed-model association analysis.

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 [33]. To date, no simplified algorithm has been published that can comprehensively reduce these three computational charges, while simultaneously improving the statistical power to detect QTNs. For a genomic dataset containing m SNPs genotyped on n individuals, GRAMMAR-Lambda took only the computing time of $O(mn^2)$ to build the GRM and $O(mn)$ for association tests. Instead of the GBLUP, we adopt ridge regression [34] to estimate the GBVs, reducing the computing time to build the GRM to $O(m_0^2 n)$, as in FaST-LMM-Select [35]. When analysing a large-scale population with complex structure, we can simplify GBVs' estimation in following way: with a specified heritability, to firstly solve the effects of m_0 sampling markers $\mathbf{a}_0 = \left[\mathbf{Z}_0\mathbf{Z}_0^{\mathbf{T}} + diag(\frac{1-h^2}{h^2})m_0\right]^{-1} \cdot \mathbf{Z}_0\mathbf{y}$ by using ridge regression [34], and then to calculate GBVs as $\mathbf{Z_0a_0}$. As a result, such a negligible computational complexity to estimate the GBVs allows GRAMMAR-Lambda to implement large mixed model association analyses in naïve association tests of PLINK [36] or GCTA [37]. A user-friendly GRAMMAR-Lambda software with these simplifications was developed, which is freely available at https://github.com/RunKingProgram/GRAMMAR-Lambda.

With the consideration of possible linkage disequilibrium among candidate markers, joint association analysis for multiple QTN candidates obtained with multiple testing significantly improves the statistical power to detect QTNs. Moreover, it is not required to re-estimate the residuals and genomic control in stepwise regression, generating a higher estimating robustness and computing efficiency than the Gaussian mixture-model association analysis [38-40] of BOLT-LMM. Since there is no need to estimate variance components (heritability or genetic relationship), GRAMMAR-Lambda can easily be extended to handle multiple correlated

phenotypes and binary disease traits, which is expected to simplify genome-wide multivariate and generalized mixed-model association studies significantly.

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).

References

1. Price, A.L., Zaitlen, N.A., David, R. & Nick, P. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459-463 (2010).
2. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
3. Pritchard, J.K. & Rosenberg, N.A. Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies. *American Journal of Human Genetics* **65**, 220-228 (1999).
4. Reich, D.E. & Goldstein, D.B. Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology* **20**, 4-16 (2001).
5. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).
6. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170-181 (2000).
7. Rosenberg, N.A. *et al*. Genetic structure of human populations. *Science* **298**, 2381-2385 (2002).
8. Zhu, X.F., Zhang, S.L., Zhao, H.Y. & Cooper, R.S. Association mapping, using a mixture model for complex traits. *Genetic epidemiology* **23**, 181-196 (2002).
9. Price, A.L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909 (2006).
10. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655-1664 (2009).
11. Spielman, R.S., McGinnis, R.E. & Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Joural of Human Genetics* **52**, 506-516 (1993).
12. Abecasis, G.R., Cardon, L.R. & Cookson, W.O. A general test of association for quantitative traits in nuclear families. *American Joural of Human Genetics* **66**, 279-292 (2000).
13. Lange, C., Demeo, D.L. & Laird, N.M. Power and Design Considerations for a General Class of Family-Based Association Tests: Quantitative Traits. *American Journal of Human Genetics* **71**, 1330-1341 (2002).
14. Laird, N.M. & Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* **7**, 385-394 (2006).
15. Henderson, C.R. *Applications of linear models in animal breeding*, xxiii, 462 pages (University of Guelph, Guelph, Ont., 1984).
16. 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).
17. Vanraden, P.M. Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414-4423 (2008).

18. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-54 (2010).
19. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355-360 (2010).
20. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833-5 (2011).
21. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821-824 (2012).
22. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23 (2008).
23. 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).
24. Vanraden, P.M. *et al.* Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16-24 (2009).
25. Habier, D., Fernando, R.L. & Dekkers, J.C. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-97 (2007).
26. 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 (2012).
27. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284-290 (2015).
28. García-Cortés, L.A. *et al.* Variance component estimation by resampling. *Journal of Animal Breeding and Genetics* **109**, 358-363 (1992).
29. 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 (2013).
30. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51**, 1749-1755 (2019).
31. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565-569 (2010).
32. Gao, J., Zhou, X., Hao, Z., Jiang, L. & Yang, R. Genome-wide barebones regression scan for mixed-model association analysis. *Theor Appl Genet* (2019).
33. 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).
34. Hoerl, A.E. & Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55-67 (1970).
35. Jennifer, L. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525-526 (2012).
36. 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).
37. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76-82 (2011).
38. Logsdon, B.A., Hoffman, G.E. & Mezey, J.G. A variational Bayes algorithm for fast and accurate

multiple locus genome-wide association analysis. *BMC Bioinformatics* **11**, 58 (2010).

39. Logsdon, B.A., Carty, C.L., Reiner, A.P., Dai, J.Y. & Kooperberg, C. A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. *Bioinformatics* **28**, 1738-44 (2012).
40. Stephens, M. & Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**, 681-90 (2009).

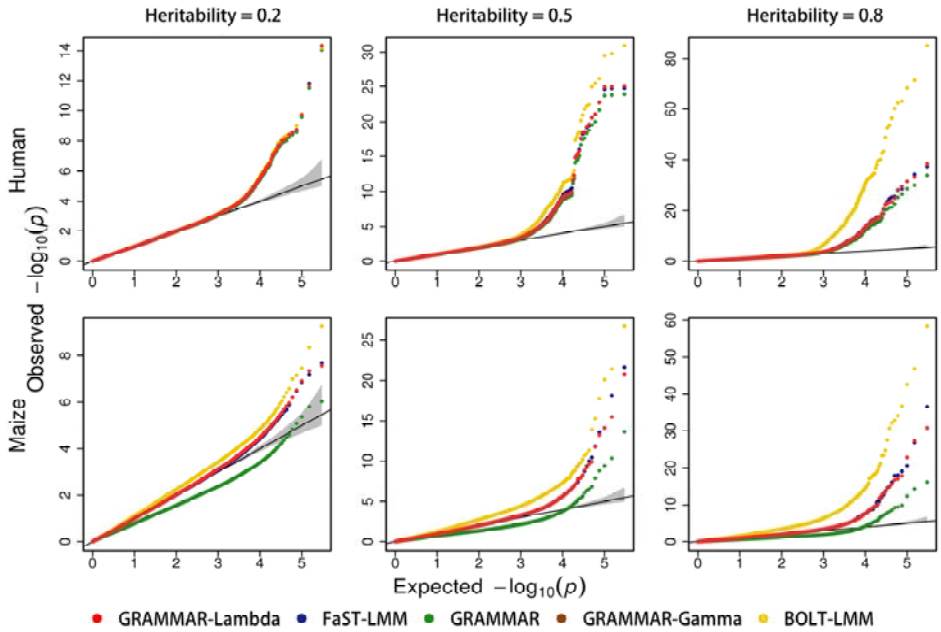

Figure 1: Comparison in the Q-Q plots between GRAMMAR-Lambda and the four competing methods. The simulated phenotypes are controlled by 200 QTNs with the low, moderate and high heritabilities in human and maize. The Q-Q plots for all simulated phenotypes are reported in Supplementary Figure 1S.

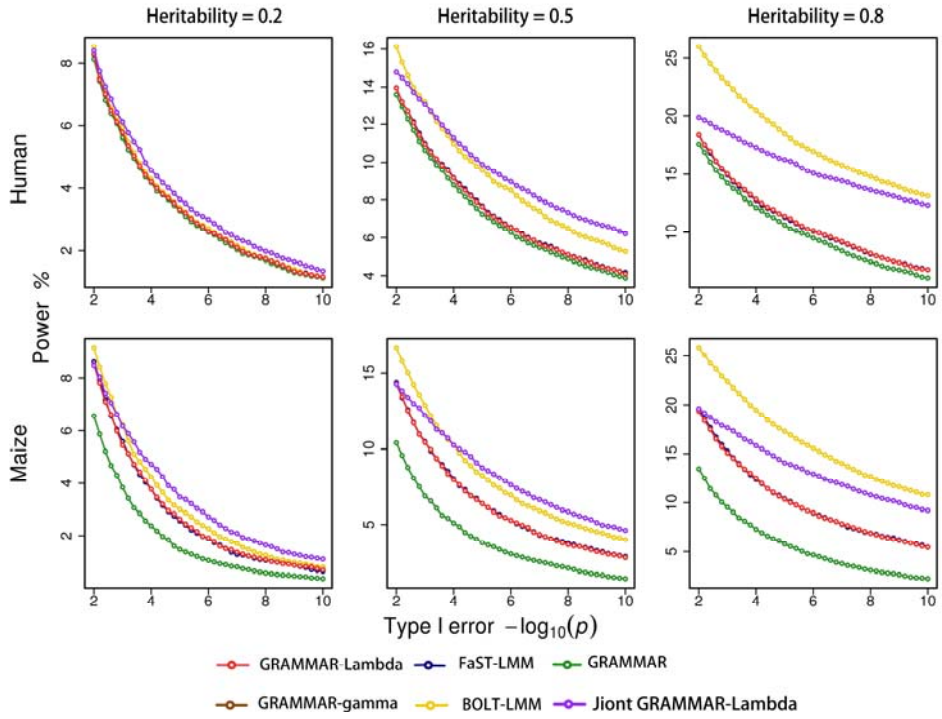

Figure 2: Comparison in the ROC profiles between GRAMMAR-Lambda and the four competing methods. The ROC profiles are plotted using the statistical powers to detect QTNs relative to the given series of Type I errors. Here, the simulated phenotypes are controlled by 200 QTNs with the low, moderate and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

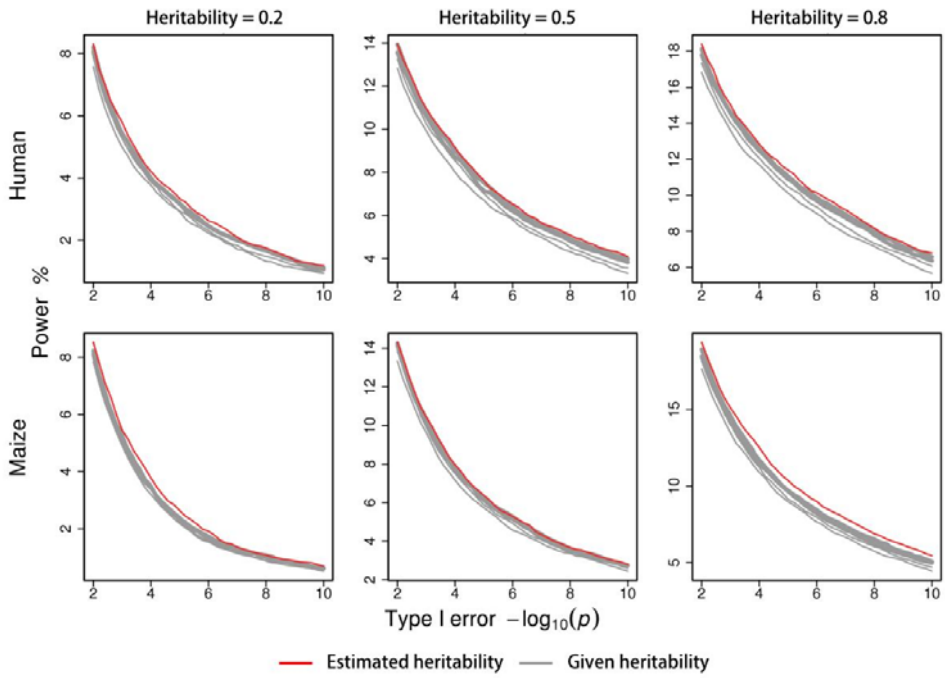

Figure 3: Sensitivity of statistical powers to the specified heritabilities for GRAMMAR-Lambda. Statistical powers are dynamically evaluated with the ROC profiles. The simulated phenotypes are controlled by 200 QTNs with the low, moderate and high heritabilities in human and maize.

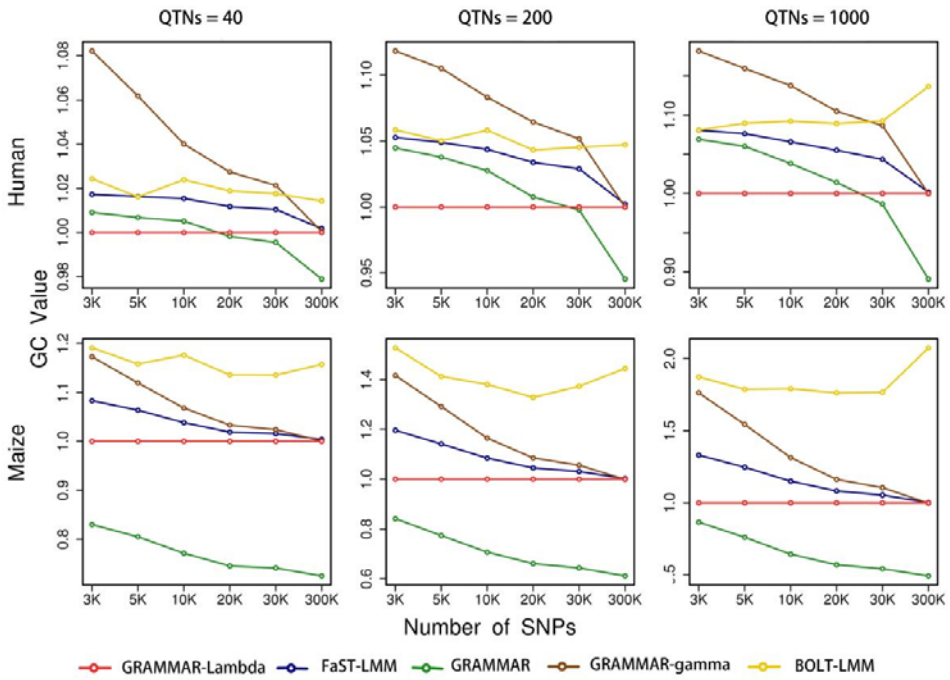

Figure 4: Changes in genomic controls with the number of sampling SNPs for GRAMMAR-Lambda and the four competing methods. Genomic control is calculated by averaging genome-wide test statistics. The simulated phenotypes are controlled by 40, 200 and 1,000 QTNs with the moderate heritability in human and maize.

Online Methods

Genomic mixed-model

Before genome-wide association analysis, the phenotype is generally justified for some fixed factors, such as ethnicity, sex and age. For comprehensive genomic control of population stratification, family structure and cryptic relatedness, we described the relationship between phenotype and the tested SNPs using the following LMM:

$$\mathbf{y} = \mathbf{z}a + \mathbf{g} + \mathbf{e} \tag{1}$$

where **y** is a vector of the justified phenotype; a is a scalar for genetic effect of the tested SNPs on phenotype; **z** is a vector of indicator variables for SNP genotypes, which is generally coded as 0, 1 or 2 for the three genotypes AA, AB and BB; **g** is a vector of polygenic effects excluding the tested SNP; $\mathbf{g}$: $N(\mathbf{0}, \mathbf{K}\sigma_g^2)$ with σ_g^2 being the polygenic variance and **K** the GRM among individuals; **e** is the residual error distributed as $\mathbf{e}$: $N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ with σ_e^2 being the residual variance and **I** the identity matrix.

GRAMMAR algorithm

In essence, GRAMMAR solves genomic mixed model using the two-stage regression method. The first stage is to estimate polygenic effects with the null model:

$$\mathbf{y} = \mathbf{g} + \mathbf{e} \tag{2}$$

To avoid computing the inverse of GRM, we construct GBLUP equations as

$$\mathbf{K}\left(\mathbf{K} + \tfrac{h^2}{1-h^2}\mathbf{I}\right)\hat{\mathbf{g}} = \mathbf{K}\mathbf{y} \tag{3}$$

where h^2 is the genomic heritability. Equation (3) can be solved using the gradient descent algorithm for a large population.

The second stage is to statistically infer association of resulting phenotypic residuals at the first stage with the tested SNP using simple regression model:

$$\mathbf{y} - \hat{\mathbf{g}} = \mathbf{z}a + \mathbf{e} \tag{4}$$

We formulate test statistic as

$$U = \frac{|\hat{a}|}{\sqrt{(\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\hat{\sigma}_e^2}} \tag{5}$$

with $\hat{a} = (\mathbf{z}^{\mathrm{T}}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{g}})$ and $\hat{\sigma}_e^2 = \frac{1}{df}(\mathbf{y} - \hat{\mathbf{g}} - \hat{a})^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{g}} - \hat{a})$, where df is residual degree of freedom.

The U statistic subjects to a standard normal distribution and in what of follows, we transform normal U to Chi-squared statistic $\chi^2 = U^2$ to evaluate genomic control.

Genomic control

At the first stage of GRAMMAR, we estimate the GBVs rather than polygenic effects. in addition of polygenic effects, as known, there are a part of QTN effects in the GBVs. Moving partial QTN effects from phenotypic residuals makes genome-wide test statistics to be deflated at the second stage, which yields serious false-negative error, especially when GEBV was precisely estimated for complex population structure. In this case, the test statistic U will subject to

$$U: \ N(0, \lambda) \tag{6}$$

rather than the standard normal distribution, where λ is the variance deflation factor (so-called genomic control). According to the frequentist outlier test [1], we could efficiently control the deflation of the test statistics χ^2 by

$$\chi_c^2 = \frac{\chi^2}{\lambda} \tag{7}$$

where λ is estimated as the mean or median of χ^2 obtained with the GRAMMAR. χ_c^2 follows Chi-squared distributions with the degree of freedom 1.

Joint association analysis

After genomic control for GRAMMAR, some SNPs with relatively low linkage disequilibrium were chosen as QTN candidates. In general, a lot of markers may be chosen as QTN candidates at the significance level lower than stringent Bonferroni corrected criterion [2], but number of QTN candidates should not be limited to less than population size to enhance efficiency for model selection. For the residual $\mathbf{y} - \hat{\mathbf{g}}$ obtained with the GBLUP, we jointly analysed multiple QTN candidates to improve the statistical power to detect QTNs. Backward regression was adopted to optimise the following multiple linear model:

$$\mathbf{y} - \hat{\mathbf{g}} = \mathbf{Z}_c \mathbf{a}_c + \mathbf{e} \tag{8}$$

where, $\mathbf{Z}_c \mathbf{a}_c$ is the regression terms of QTN candidates. Given the Bonferroni corrected level of significance, the genetic effects were selected stepwise, and corresponding QTNs identified according to the corrected statistics (7).

Simulations

To verify the adaptability of GRAMMAR-Lambda to population structure, we used human genomic data [3] with simple population stratification, and maize genomic data [4] with complex cryptic relativeness among individuals to simulate trait phenotypes. With higher quality control, 300,000 SNPs were extracted from both 3000 human and 2640 maize. QTNs were randomly distributed over the whole genomic SNPs. Additive effects of simulated QTNs were drawn from the gamma distribution with shape=1.66 and scale=0.4 so that there were few with large effects and more with minor effects in the simulated QTNs. The genetic effect for each individual was calculated as the sum of all QTN genotypic effects. In sampling residual errors using the normal distribution, the residual variance was regulated by the given genomic heritability of traits. The trait phenotypes were generated by adding up the genetic effects and residual errors.

In addition to population structure, the number of QTNs, genomic heritability and sampling number of SNPs were considered as experimental factors in the simulations. We carried out three simulations to investigate (1) statistical properties of GRAMMAR-Lambda by simulating phenotypes controlled by 40, 200 and 1000 QTNs with varied levels of heritabilities (low (0.2), moderate (0.5), and high (0.8)); (2) sensitivity to estimates of variance components or genomic heritability by using the phenotypes controlled by 200 QTNs in the first simulation; and (3) calculation of GRMs with sampling markers by using the phenotypes at the moderate genomic heritability in the first simulation. In the first and third simulations, GRAMMAR-Lambda was compared to FaST-LMM, GRAMMAR, GRAMMAR-Gamma and BOLT-LMM in terms of Q-Q profiles, or genomic controls, and ROC profiles. Especially in the first simulation, GRAMMAR-Gamma jointly analyses multiple QTN candidates with a test at once.

Under perfect genomic control (1.0), the ROC profiles can be plotted using the statistical powers to detect QTNs relative to the given series of Type I errors. Statistical powers were defined as the percentage of identified QTNs that had the maximum test statistics among their 20 closest neighbours over the total number of simulated QTNs. Simulations were repeated 50 times, in each the positions and effects of simulated QTNs varied, and the average results were recorded.

Real data

The datasets were collected from three species: 1) the *Arabidopsis thaliana* dataset included 199 samples, with 216,130 SNPs and 107 phenotypes [5]; 2) the mouse dataset had 1,940 samples with 12,226 SNPs and 123 phenotypes [6]; 3) the maize datasets involved 2,279 inbred lines, each with 681,258 SNPs genotyped and the flowering time phenotype measured as days to silk [4] (URL: http://www.panzea.org/!#genotypes/cctl). 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.

Method References

1. Barnett, V. & Lewis, T. *Outliers in statistical data analysis*, (Wiley, New York, 1994).
2. Hochberg, Y. & Tamhane, A.C. *Multiple Comparison Procedures*, (John Wiley & Sons, Inc., New York, 1987).
3. 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).
4. Romay, M.C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. *Genome biology* **14**, R55 (2013).
5. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627-31 (2010).
6. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* **38**, 879-87 (2006).