nature genetics

Genome-wide efficient mixed-model analysis for association studies

Xiang Zhou[1] & Matthew Stephens[1,2]

Linear mixed models have attracted considerable attention recently as a powerful and effective tool for accounting for population stratification and relatedness in genetic association tests. However, existing methods for exact computation of standard test statistics are computationally impractical for even moderate-sized genome-wide association studies. To address this issue, several approximate methods have been proposed. Here, we present an efficient exact method, which we refer to as genome-wide efficient mixed-model association (GEMMA), that makes approximations unnecessary in many contexts. This method is approximately *n* times faster than the widely used exact method known as efficient mixed-model association (EMMA), where *n* is the sample size, making exact genome-wide association analysis computationally practical for large numbers of individuals.

There is an increasing interest in using linear mixed models (LMMs, also known as mixed linear models (MLMs)) to test for association in genome-wide association studies (GWAS) because of their demonstrated effectiveness in accounting for relatedness among samples and in controlling for population stratification and other confounding factors[1–7]. However, these models present substantial computational challenges. For example, at the time that this work was submitted for publication, the most efficient algorithm for effectively computing exact association test statistics (either the Wald test or the likelihood-ratio test) implemented in the EMMA software[3] had a per-SNP computation time that increased with the cube of the number of individuals (n). As a result, an average-sized GWAS including a few thousand individuals and half a million SNPs would take years of central processing unit (CPU) time to analyze[1,7]. While this paper was in review, Lippert *et al.*[8] also published an efficient algorithm for this model, implemented in the FaST-LMM software; the relationship between this algorithm and ours is discussed.

Several approximation methods have been proposed to make genome-wide analysis using linear mixed models possible. Probably the simplest and fastest of these approximations, genome-wide rapid association using mixed model and regression (GRAMMAR) implemented in the GenABEL software[9] first estimates the residuals from the LMM under the null model (no SNP effect) and then treats these residuals as phenotypes for further genome-wide analysis by a standard linear model[10]. This substantially reduces per-SNP computation time, making it linear with respect to the number of individuals included. More recently, two more sophisticated approximate approaches have been suggested. Zhang *et al.*[7] use population parameters previously determined (P3D), which avoids repeatedly estimating variance components when performing each test by simply using the pre-estimated variance components from the null model; their method is implemented in the TASSEL software. Kang *et al.*[1] also avoid repeatedly estimating variance components by a slightly different strategy, which keeps the heritability estimated from the null model fixed when testing individual SNPs. Their approach is implemented in the EMMA eXpedited (EMMAX) software. (This approximation and related ideas were also considered by previous authors[10,11].) Both approximations have per-SNP computation time that increases quadratically with the number of individuals, which makes them practical on a single desktop computer for GWAS involving thousands of individuals.

Although in some settings the approximate methods provide results almost identical to those from the exact method[1,7], this is not guaranteed in general, and in practice it is hard to know how accurate the approximations will be without running an exact calculation. One possible consequence of inaccuracy in the approximation could be a reduction in power relative to exact methods. For these reasons, the ability to perform exact calculations remains of interest. Here, we present a new, more efficient method for exact calculations that provides numerically identical results to EMMA (exact Wald or likelihood-ratio test statistics) but is roughly n times faster (computation time per SNP, when using the usual genome-wide relatedness matrix, increases quadratically with the number of individuals, with a run time similar to that of EMMAX). This makes exact calculations feasible for large GWAS, thereby obviating the need for approximate methods in most common settings.

RESULTS

The method and its computational complexity are described in detail and derived in the Online Methods. Briefly, the method requires complete or imputed genotype data[12,13] for all SNPs and involves only one eigen decomposition of the relatedness matrix at the beginning (computational complexity of $O(n^3)$, where O is the big O notation[14]). For each SNP tested, it effectively replaces the expensive additional eigen-decomposition step in EMMA with one matrix-vector multiplication (computational complexity

[1]Department of Human Genetics, University of Chicago, Chicago, Illinois, USA. [2]Department of Statistics, University of Chicago, Chicago, Illinois, USA. Correspondence should be addressed to X.Z. (xz7@uchicago.edu) or M.S. (mstephens@uchicago.edu).

Received 11 July 2011; accepted 4 May 2012; published online 17 June 2012; doi:10.1038/ng.2310

Table 1 Performance of different methods for GWAS with the linear mixed model

Methods		Time complexity[a]	Computing time	
			HDL-C[b]	Crohn's disease[c]
Exact methods	GEMMA	$O(mn^2 + cn^2 + pn^2 + pt_2c^2n)$	33 min	3.3 h
	EMMA	$O(mn^2 + pmn^2 + pt_2n)$	~9 d	~27 years
	FaST-LMM[d]	$O(mn^2 + cn^2 + pn^2 + pt_1c^2n)$	6.8 h	6.2 h
Approximate methods	EMMAX	$O(mn^2 + t_2n + pn^2)$	44 min	6.4 h
	GRAMMAR	$O(mn^2 + t_2n + pn)$	1.6 min	12 min

All computing was performed on a single core of an Intel Xeon L5420 2.50 GHz CPU. The time for the EMMA method is projected from a selection of 10,000 and 100 genetic markers in the HMDP and WTCCC data sets, respectively. Note that EMMA was implemented in R, whereas others were implemented in C. A C implementation of EMMA could be a few times faster. p, the number of genetic markers; n, the number of individuals; m, the number of strains (equal to n for human studies); c, the number of covariates (fixed effects) in addition to the genotypes. t_1 and t_2 are the number of optimization iterations required for Brent's method (super-linear rate of convergence) and the Newton-Raphson method (quadratic rate of convergence), respectively. Note that t_2 is expected to be smaller than t_1. [a]Complexities are given assuming the usual genome-wide relatedness matrix, which has rank n. In the current implementation of various methods except EMMA, the first terms are actually n^3, but it would in principle be straightforward to convert them to mn^2. [b]$m = 99$, $n = 681$, and $p = 1{,}885{,}197$. [c]$m = n = 4{,}686$, and $p = 442{,}001$. [d]These results are for the algorithm in FaST-LMM that uses the standard full-rank relatedness matrix, which produces P values that are identical to those generated in GEMMA and EMMA.

of $O(n^2)$). After this, as in EMMA, each iteration of the following optimization step requires inexpensive operations (complexity of $O(n)$) to evaluate both first and second derivatives of the target functions. We refer to our method as genome-wide efficient mixed-model association (GEMMA) because it builds on EMMA and facilitates its genome-wide application.

We implemented our method and compared the analysis results with those obtained using the exact method, EMMA, and the approximation methods, EMMAX and GRAMMAR, using two examples: a mouse GWAS for high-density lipoprotein–cholesterol (HDL-C) levels from the Hybrid Mouse Diversity Panel (HMDP)[15] and a human GWAS for Crohn's disease from the Wellcome Trust Case Control Consortium (WTCCC)[16]. The size of this second study makes it computationally impractical to analyze it with EMMA[3]. The computational complexity for the four methods and the CPU time for the analysis of the two data sets on a single desktop CPU are summarized in **Table 1**. We also include the results obtained with the recently published FaST-LMM[8], which can produce identical P values to those generated by EMMA and GEMMA in the same time complexity as GEMMA. As expected, GEMMA was comparable in speed to EMMAX, completing the larger (WTCCC) example in less than 4 h.

To verify the correctness of our algorithm and implementation, we first validated it by comparing the P values calculated by GEMMA with those from EMMA on a subset of SNPs from both example data sets. For all SNPs examined, the P values from the two methods matched exactly (Wald test results shown in **Fig. 1a**,**b**; likelihood-ratio test results not shown).

Because GEMMA provides exact computations in essentially the same time as EMMAX, the accuracy of the approximations in EMMAX and other methods may seem irrelevant. However, in some settings and in particular for mixed models with more than one random effect (variance component), the computational trick used by GEMMA does not apply, and approximations along the lines of EMMAX may remain necessary. For this reason, the accuracy of different approximation methods remains of some interest, and we therefore present a comparison between the Wald test P values from GEMMA, EMMAX and GRAMMAR across the genome for both the HMDP and WTCCC data sets.

The HMDP GWAS represents a situation where approximation methods such as EMMAX or GRAMMAR may yield inaccurate test statistics. In particular, because individuals in the data set are closely related and because the strongly associated SNPs contribute to a significant proportion of phenotypic variation in HDL-C[13], using estimates of variance components or fitted residuals from the null model for testing might be expected to yield conservative P values, potentially leading to a loss of power. Our empirical comparison (**Fig. 1c**) confirms this: approximation by EMMAX led to systematic and appreciable underestimation of the most significant P values (by almost two orders of magnitude), whereas approximation by GRAMMAR led to marked underestimation of all P values. Indeed, in contrast to the exact P values, no P values generated by EMMAX were significant at the conventional $P = 0.05$ level after Bonferroni correction, and no P values generated by GRAMMAR were significant even before Bonferroni correction. The fact that the exact P values for the most significant results are substantially more significant than the approximate P values from EMMAX suggests that, in this type of setting, the exact P values may produce a more powerful test, and simulation results confirm this (**Supplementary Fig. 1**).

In contrast, the WTCCC example represents a very different situation, where the approximation methods might be expected to yield accurate test statistics. This is because there is relatively little population stratification in these data (the individuals are all from the UK, with the relatedness matrix approximately diagonal), and the effect sizes of the most strongly associated SNPs for Crohn's disease are small compared with the effect sizes in the HMDP data[15]. Both conditions favor the approximation assumptions in EMMAX and GRAMMAR. Empirical comparisons (**Fig. 1d**) showed that, for this particular data set, the P values from EMMAX differed only negligibly from the exact values. However, the P values from GRAMMAR still departed noticeably from the exact values.

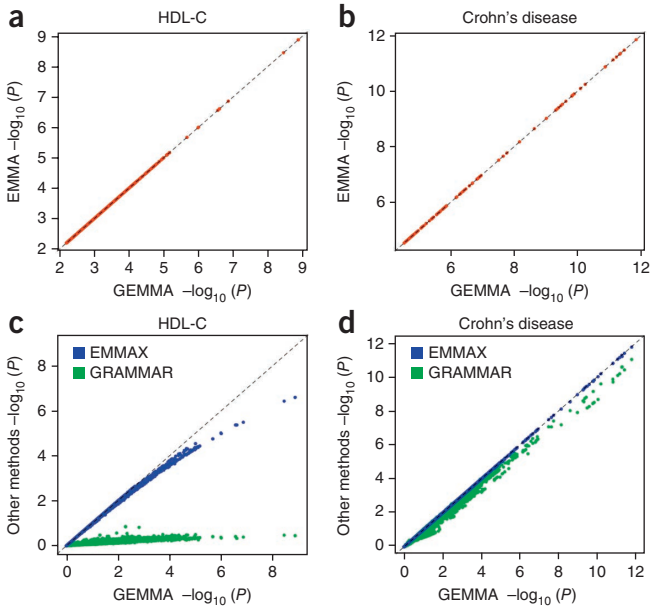

Figure 1 Comparison of GEMMA with EMMA, EMMAX and GRAMMAR on HMDP HDL-C data and WTCCC Crohn's disease data. (**a**,**b**) Comparison of $-\log_{10}$ P values obtained from GEMMA with those from EMMA. P values are shown for the top 10,000 markers (**a**) and the top 100 markers (**b**). (**c**,**d**) Comparison of $-\log_{10}$ P values obtained from GEMMA with those from EMMAX and GRAMMAR. P values are shown for all markers: 1.9 million (**c**) and 442,000 (**d**).

Taken together, these results confirm that approximation by EMMAX is appreciably more accurate than with GRAMMAR, even in cases such as the WTCCC data where the sample structure is subtle. The comparisons also show that the accuracy of EMMAX approximation can vary from case to case. Consequently, the potential gain in power from performing exact instead of approximate tests will also vary among data sets. For the HMDP data set, the potential gain in power from the exact calculations seems considerable, and this is confirmed by simulations (**Supplementary Fig. 1**). For the WTCCC Crohn's disease data set, the power gain is negligible, and, as noted in ref. 1, only a small gain in power is generally expected at SNPs with small effect sizes. Of course, one advantage of being able to perform the exact tests is that it obviates the need to consider which approximations work best under which circumstances or to consider ways in which the approximations could be improved. We also note that the computational methods employed here can be applied in other contexts, including, for example, the combined variable selection plus random effects model that has been widely studied for phenotype and breeding value prediction[17].

DISCUSSION

In summary, we have presented an efficient method for computing exact values of standard test statistics in linear mixed models. This method is comparable in speed to approximation methods such as EMMAX but yields exact test statistics. By analyzing two example data sets, we demonstrate the use of our method and show that the approximation methods can yield inaccurate *P* values when the sample structure is strong and/or when the marker effect size is large. We also find that approximation by EMMAX is more accurate than approximation by GRAMMAR across the genome (a comparison made possible only by the availability of an efficient exact method).

Lippert *et al.*[8] also recently published an efficient method for computing likelihoods for LMMs that, similar to our method, requires only one singular value decomposition of the relatedness matrix. They use this method in combination with Brent's optimization algorithm to produce an algorithm for computing exact test statistics with effectively the same computational complexity as GEMMA: $O(mn^2 + cn^2 + pn^2 + ptc^2n)$, where n is the number of individuals, m is the number of strains (equal to n for human studies), p is the number of genetic markers, c is the number of covariates in addition to the genotypes and t is the number of optimization iterations required for convergence (**Table 1**). (Lippert *et al.*[8] also suggest a further innovation in which a low-rank relatedness matrix is used in place of the usual relatedness matrix computed from all SNPs across the genome, which produces an algorithm that is linear with respect to n and therefore is feasible for very large GWAS samples containing more than 100,000 individuals; however, changing the relatedness matrix in this way changes the resulting *P* values appreciably, and in this sense this linear complexity algorithm is not directly comparable with either GEMMA or EMMA.) The main additional contribution of our work here, beyond that described by Lippert *et al.*, is that we provide and demonstrate the use of efficient methods for the evaluation of not only the likelihood but also both its first and second derivatives. This allows use of the Newton-Raphson optimization method, which has better theoretical convergence properties than Brent's algorithm (quadratic versus super-linear, respectively), potentially reducing per-SNP computation time by reducing the number of iterations (t) required for convergence. The practical effect of this is expected to depend on the sample size n. Examining the theoretical computational complexity, if p is large (and assuming the simplest case with no additional covariates, such that $c = 1$), the per-SNP complexity of the algorithms is $O(n^2 + tn)$. Thus, if n is large, the n^2 term will dominate, and the number of iterations will have only a small effect on computation time; if n is moderate, then the number of iterations may have a more substantial contribution. Consistent with this idea, we found that GEMMA was 12 times faster than the algorithm developed by Lippert *et al.* when implemented in FaST-LMM for the smaller HMDP data set (33 min versus 6.8 h, respectively) but was only 2 times faster for the WTCCC data set (3.3 h versus 6.2 h, respectively). It is possible that implementation issues, which are important but conceptually less fundamental, also contribute to differences in speed. In addition to allowing slightly faster computational speed, which might be considered a minor issue, by providing efficient methods to compute derivatives, our work here lays the foundation for similar efficient analyses for LMMs with multivariate phenotypes[18], where multidimensional optimization is required, and evaluating the target functions alone is unlikely to suffice.

Here, we have focused on computations using the usual relatedness matrix computed from all SNPs across the genome whose rank r is typically equal to the number of individuals n. However, as noted by Lippert *et al.*[8], using a lower-rank relatedness matrix reduces computing time (computational complexity of the singular value decomposition can scale with nr^2) and, in some cases, memory requirements (for example, Lippert *et al.*[8] suggest using a relatedness matrix based on only a few thousand SNPs; this is advantageous in that the required singular value decompositions can be completed without computing the $n \times n$ relatedness matrix itself). Using the usual full-rank relatedness matrix, our current implementation of GEMMA can handle approximately 23,000 individuals on a machine with 64 GB of memory (in double precision); using a lower-rank relatedness matrix, much larger problems can be addressed. However, we note that changing the relatedness matrix can produce much larger changes in *P* values than, for example, using EMMAX versus exact calculations (**Supplementary Fig. 2**), and, for both the HMDP and WTCCC data sets, using a lower-rank relatedness matrix seems to compromise the ability of the LMM to control for sample structure (**Supplementary Table 1**). Thus, choice of relatedness matrix could affect statistical efficiency (both power and correct control of type I error due to stratification or relatedness), as well as computational efficiency. Notably, statistical and computational considerations may not necessarily conflict: for example, Zhang *et al.*[7] suggest that the use of compressed MLM, which yields a lower-rank relatedness matrix by clustering individuals, can both reduce computation and increase power compared with the full-rank matrix. The general question of which low-rank relatedness matrices produce the best combination of computational and statistical performance seems to be an interesting avenue for further study.

URLs. Freely available implementation of the GEMMA software, http://stephenslab.uchicago.edu/software.html; WTCCC, http://www.wtccc.org.uk/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This research is supported in part by grants from the US National Institutes of Health (NIH) (HL092206 to Y. Gilad and HG02585 to M.S.). We thank A.J. Lusis for making the mouse genotype and phenotype data available. This study also makes use of data generated by the WTCCC[15]. A full list of the investigators who contributed to the generation of the data is available from the WTCCC website. Funding for the WTCCC project was provided by the Wellcome Trust (award 085475).

AUTHOR CONTRIBUTIONS

X.Z. and M.S. designed the study, developed methods and wrote the manuscript. X.Z. implemented software and analyzed data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2310.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
2. Kang, H.M., Ye, C. & Eskin, E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* **180**, 1909–1925 (2008).
3. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
4. Listgarten, J., Kadie, C., Schadt, E.E. & Heckerman, D. Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. USA* **107**, 16465–16470 (2010).
5. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
6. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
7. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
8. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
9. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
10. 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).
11. Abney, M., Ober, C. & McPeek, M.S. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am. J. Hum. Genet.* **70**, 920–934 (2002).
12. Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genet.* **4**, e1000279 (2008).
13. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
14. Knuth, D.E. Big Omicron and big Omega and big Theta. *ACM SIGACT News.* **8**, 18–24 (1976).
15. Bennett, B.J. *et al.* A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res.* **20**, 281–290 (2010).
16. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
17. Lee, S.H., van der Werf, J.H., Hayes, B.J., Goddard, M.E. & Visscher, P.M. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* **4**, e1000231 (2008).
18. Meyer, K. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genet. Sel. Evol.* **23**, 67–83 (1991).

ONLINE METHODS

Linear mixed-model and target-optimization functions. We consider the following standard linear mixed model

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\mathbf{u} \sim \mathrm{MVN}_{m}(0, \lambda\tau^{-1}\mathbf{K})$$

$$\boldsymbol{\varepsilon} \sim \mathrm{MVN}_{n}(0, \tau^{-1}\mathbf{I}_{n})$$

where n is the number of individuals, m is the number of groups, strains or clusters, $\mathbf{y}$ is an $n \times 1$ vector of quantitative traits, $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2 \ldots \mathbf{w}_c)$ is an $n \times c$ matrix of covariates (fixed effects) including a column vector of 1, $\boldsymbol{\alpha}$ is a $c \times 1$ vector of corresponding coefficients including the intercept, $\mathbf{x}$ is an $n \times 1$ vector of marker genotypes, β is the effect size of the marker, $\mathbf{Z}$ is an $n \times m$ loading matrix, $\mathbf{u}$ is an $m \times 1$ vector of random effects, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors, τ^{-1} is the variance of the residual errors, λ is the ratio between the two variance components, $\mathbf{K}$ is a known $m \times m$ relatedness matrix, $\mathbf{I}_n$ is an $n \times n$ identity matrix and MVN denotes multivariate normal distribution.

In the case of the HMDP data set, m is the number of strains, n is the number of animals and matrix $\mathbf{Z}$ indicates which strain each animal arises from ($\mathbf{z}_{ij} = 1$ if individual i comes from strain j and = 0 otherwise). In the case of the WTCCC data set, $m = n$ and $\mathbf{Z}$ is an identity matrix. Multiple covariates, such as cluster memberships or eigenvectors[5–7], can be incorporated into $\mathbf{W}$.

We are interested in obtaining both the maximum-likelihood estimates (MLEs) and the restricted/residual maximum-likelihood (REML) estimates and further exact test statistics. We use the term 'exact' for brevity, although the more precise term is 'effectively exact'. This is because computing the statistics involves an optimization problem that is not guaranteed to be convex, and, therefore, in general one cannot be guaranteed of finding the global optimum. However, existing optimization methods seem to be highly effective in practice. The following description and derivation of the GEMMA algorithm uses a few properties that have been described previously[19].

The log-likelihood and log-restricted likelihood functions for the standard linear mixed model are

$$l(\lambda, \tau, \boldsymbol{\alpha}, \beta) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\tau(\mathbf{y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{x}\beta)^{T}\mathbf{H}^{-1}(\mathbf{y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{x}\beta) \quad (1)$$

and

$$l_r(\lambda, \tau) = \frac{n-c-1}{2}\log(\tau) - \frac{n-c-1}{2}\log(2\pi) + \frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^{T}(\mathbf{W},\mathbf{x})\right| - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^{T}\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})\right| - \frac{1}{2}\tau\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y} \quad (2)$$

where $\mathbf{G} = \mathbf{Z}\mathbf{K}\mathbf{Z}^{T}$, $\mathbf{H} = \lambda\mathbf{G} + \mathbf{I}_n$ and $\mathbf{P}_X = \mathbf{H}^{-1} - \mathbf{H}^{-1}(\mathbf{W},\mathbf{x})((\mathbf{W},\mathbf{x})^{T}\mathbf{H}^{-1}(\mathbf{W},\mathbf{x}))^{-1}(\mathbf{W},\mathbf{x})^{T}\mathbf{H}^{-1}$.

If λ is known, we can easily obtain $\hat{\boldsymbol{\alpha}}$, $\hat{\beta}$ and $\hat{\tau}$ for both log-likelihood and log-restricted likelihood functions (**Supplementary Note**). Therefore, finding MLEs and REML estimates is equivalent to optimizing the following target functions with respect to λ:

$$l(\lambda) = \frac{n}{2}\log\left(\frac{n}{2\pi}\right) - \frac{n}{2} - \frac{1}{2}\log|\mathbf{H}| - \frac{n}{2}\log\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}\right) \quad (3)$$

$$l_r(\lambda) = \frac{n-c-1}{2}\log\left(\frac{n-c-1}{2\pi}\right) - \frac{n-c-1}{2} + \frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^{T}(\mathbf{W},\mathbf{x})\right| - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^{T}\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})\right| - \frac{n-c-1}{2}\log\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}\right) \quad (4)$$

Optimization method overview. A direct, naive evaluation of the likelihood function or the restricted-likelihood function has a computational time that increases with the cube of the number of individuals because it involves calculating a matrix determinant and a matrix inversion. A similarly expensive computation involving a matrix inversion and a few matrix-vector multiplications is used for each update step in the standard Henderson's iterative optimization procedure[20]. Therefore, Henderson's optimization algorithm is relatively slow. The algorithm in EMMA[3] solves this problem by eigen decompositions of matrix $\mathbf{G}$ and matrix $\mathbf{P}_x$ before optimization. After that, each target function involves only a summation of n scalar functions, thus making the generation of the derivatives straightforward and their evaluation efficient. As a result, EMMA performs a single expensive calculation for each marker (decomposition of $\mathbf{P}_x$) followed by an iterative maximization scheme that involves only inexpensive operations (linear complexity in the number of individuals for each iteration).

We take a different approach and obtain the first and second derivatives in vector and matrix forms before eigen decomposition of the relatedness matrix $\mathbf{G}$. Using three key recursions, we further show that both target functions and derivatives in vector/matrix forms for each marker, despite their complicated appearance, are easy and efficient to evaluate during each optimization step. Therefore, we effectively replace the expensive eigen decomposition of matrix $\mathbf{P}_x$ for each SNP with an inexpensive matrix-vector multiplication followed by a few recursions involving only scalar multiplications. As in EMMA, each iteration of iterative maximization involves only inexpensive operations (linear complexity in the number of individuals n, quadratic complexity in the number of covariates c).

For numeric optimization, we start with Brent's method on the first derivative for stability and follow with the Newton-Raphson method, using the second derivative for efficiency. Details are given in the **Supplementary Note**.

Note that eigen decomposition can be completed more quickly when $m < n$ with a modification of the Gram-Schmidt process[3]; however, this trick is not expected to substantially improve analysis for a genome-wide analysis and has not been implemented in the current version of the software.

Derivatives of target functions. We obtain the first and second derivatives for the log-likelihood function with

$$\frac{\partial l(\lambda)}{\partial\lambda} = -\frac{1}{2}\mathrm{trace}\left(\mathbf{H}^{-1}\mathbf{G}\right) + \frac{n}{2}\frac{\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{y}}{\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}} \quad (5)$$

$$\frac{\partial^2 l(\lambda)}{\partial\lambda^2} = \frac{1}{2}\mathrm{trace}\left(\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}\mathbf{G}\right) - \frac{n}{2}\frac{2\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{y}\right)\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}\right) - \left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{y}\right)^{2}}{\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}\right)^{2}} \quad (6)$$

and the first and second derivatives for the log-restricted likelihood function with the following equations:

$$\frac{\partial l_r(\lambda)}{\partial\lambda} = -\frac{1}{2}\mathrm{trace}\left(\mathbf{P}_{x}\mathbf{G}\right) + \frac{n-c-1}{2}\frac{\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{y}}{\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}} \quad (7)$$

$$\frac{\partial^2 l_r(\lambda)}{\partial\lambda^2} = \frac{1}{2}\mathrm{trace}\left(\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{G}\right) - \frac{n-c-1}{2}\frac{2\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{y}\right)\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}\right) - \left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{G}\mathbf{P}_{x}\mathbf{y}\right)^{2}}{\left(\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}\right)^{2}} \quad (8)$$

The above equations are obtained using a few matrix calculus properties, which are listed in detail in the **Supplementary Note**.

Several quantities require efficient evaluation. There are a few quantities that need to be efficiently evaluated for each genetic marker in each optimization step. For the log-likelihood and log-restricted likelihood functions (3) and (4), we need to evaluate $|\mathbf{H}|$, $\left|(\mathbf{W},\mathbf{x})^{T}\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})\right|$ and $\mathbf{y}^{T}\mathbf{P}_{x}\mathbf{y}$. For the derivatives of the log-likelihood and log-restricted likelihood functions (5)–(8), we need to evaluate two types of quantities: trace terms ($\mathrm{trace}(\mathbf{H}^{-1}\mathbf{G})$, $\mathrm{trace}(\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}\mathbf{G})$, $\mathrm{trace}(\mathbf{P}_x\mathbf{G})$ and $\mathrm{trace}(\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G})$), and vector-matrix-vector

product terms ($\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}$ and $\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}$). We note that the trace terms can be derived from trace($\mathbf{H}^{-1}$), trace($\mathbf{H}^{-1}\mathbf{H}^{-1}$), trace($\mathbf{P}_x$) and trace($\mathbf{P}_x\mathbf{P}_x$), and the vector-matrix-vector product terms can be derived from $\mathbf{y}^T\mathbf{P}_x\mathbf{y}$, $\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{y}$ and $\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{P}_x\mathbf{y}$ (**Supplementary Note**). Therefore, we need to efficiently evaluate three types of quantities for each SNP for any given λ

1. Determinant terms $|\mathbf{H}|$ and $\left|(\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})\right|$
2. Trace terms trace($\mathbf{H}^{-1}$), trace($\mathbf{H}^{-1}\mathbf{H}^{-1}$), trace($\mathbf{P}_x$) and trace($\mathbf{P}_x\mathbf{P}_x$)
3. Vector-matrix-vector product terms $\mathbf{y}^T\mathbf{P}_x\mathbf{y}$, $\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{y}$ and $\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{P}_x\mathbf{y}$

We separate the above terms into basic quantities (which involve matrix $\mathbf{H}$) and induced quantities (which involve matrix $\mathbf{P}_x$). The next two sections describe how these quantities were evaluated.

Calculation of the basic quantities. Here, we describe the efficient calculations of three basic quantities: the determinant term $|\mathbf{H}|$, the trace terms trace($\mathbf{H}^{-1}$) and trace($\mathbf{H}^{-1}\mathbf{H}^{-1}$), and the vector-matrix-vector product terms in the forms of $\mathbf{a}^T\mathbf{H}^{-1}\mathbf{b}$, $\mathbf{a}^T\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{b}$ and $\mathbf{a}^T\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{b}$, with $\mathbf{a}$ and $\mathbf{b}$ being equal to one of $\mathbf{w}_i$, $\mathbf{x}$ and $\mathbf{y}$.

Before the genome-wide analysis, we first obtain an eigen decomposition $\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ with time complexity $O(mn^2)$, where $\mathbf{D} = \mathrm{diag}(\delta_1, \delta_2 \ldots, \delta_n)$ and δ_i values are the eigen values. As $\mathbf{I}_n = \mathbf{U}\mathbf{U}^T$, we have $\mathbf{H} = \mathbf{U}\mathrm{diag}(\lambda\delta_1 + 1, \ldots, \lambda\delta_n + 1)\mathbf{U}^T$. Therefore, during each optimization step, the determinant term can be calculated with time complexity $O(n)$: $|\mathbf{H}| = \prod_{i=1}^{n}(\lambda\delta_i + 1)$. Similarly, the trace terms can be evaluated with time $O(n)$: $\mathrm{trace}\left(\mathbf{H}^{-1}\right) = \sum_{i=1}^{n}(\lambda\delta_i + 1)^{-1}$ and $\mathrm{trace}\left(\mathbf{H}^{-1}\mathbf{H}^{-1}\right) = \sum_{i=1}^{n}(\lambda\delta_i + 1)^{-2}$.

Next, we define and calculate $(\mathbf{v}_{w1}, \mathbf{v}_{w2} \ldots \mathbf{v}_{wc}) = \mathbf{U}^T\mathbf{W}$, $\mathbf{v}_y = \mathbf{U}^T\mathbf{y}$ and $\mathbf{v}_x = \mathbf{U}^T\mathbf{x}$, each with time complexity $O(n^2)$, and only $\mathbf{v}_x$ needs to be calculated for each SNP. Then, for any $\mathbf{a}$ and $\mathbf{b}$ that are equal to one of $\mathbf{w}_i$, $\mathbf{x}$ and $\mathbf{y}$ during each optimization step with time complexity $O(n)$, we obtain

$$\mathbf{a}^T\mathbf{H}^{-1}\mathbf{b} = \sum_{i=1}^{n} v_{ai}v_{bi}(\lambda\delta_i + 1)^{-1}$$

$$\mathbf{a}^T\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{b} = \sum_{i=1}^{n} v_{ai}v_{bi}(\lambda\delta_i + 1)^{-2}$$

$$\mathbf{a}^T\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{b} = \sum_{i=1}^{n} v_{ai}v_{bi}(\lambda\delta_i + 1)^{-3}$$

where v_{ai} and v_{bi} are the corresponding ith elements in the vectors $\mathbf{U}^T\mathbf{a}$ and $\mathbf{U}^T\mathbf{b}$, respectively.

Recursions for the induced quantities. Here, we describe three recursions to efficiently evaluate the induced quantities from the basic quantities. The induced quantities are the determinant term $\left|(\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})\right|$, the trace terms trace($\mathbf{P}_x$) and trace($\mathbf{P}_x\mathbf{P}_x$), and the vector-matrix-vector product terms $\mathbf{y}^T\mathbf{P}_x\mathbf{y}$, $\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{y}$ and $\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{P}_x\mathbf{y}$.

First, we define $\mathbf{P}_0 = \mathbf{H}^{-1}$, $\mathbf{P}_{c+1} = \mathbf{P}_x$, $\mathbf{w}_{c+1} = \mathbf{x}$, $\mathbf{W}_i = (\mathbf{w}_1, \mathbf{w}_2 \ldots \mathbf{w}_i)$ and $\mathbf{P}_i = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{W}_i\left(\mathbf{W}_i^T\mathbf{H}^{-1}\mathbf{W}_i\right)^{-1}\mathbf{W}_i^T\mathbf{H}^{-1}$ for i $\in$ {1, 2… c+1}. With the Leibniz formula, we obtain a recursion for the determinant term $\left|\mathbf{W}_i^T\mathbf{H}^{-1}\mathbf{W}_i\right| = \left|\mathbf{W}_{i-1}^T\mathbf{H}^{-1}\mathbf{W}_{i-1}\right|\left(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i\right)$.

Next, with blockwise matrix inversion, we obtain $\mathbf{P}_i = \mathbf{P}_{i-1} - \mathbf{P}_{i-1}\mathbf{w}_i\left(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i\right)^{-1}\mathbf{w}_i^T\mathbf{P}_{i-1}$. This leads to a recursion for the trace terms trace($\mathbf{P}_i$) and trace($\mathbf{P}_i\mathbf{P}_i$) and another recursion for the vector-matrix-vector product terms $\mathbf{a}^T\mathbf{P}_i\mathbf{b}$, $\mathbf{a}^T\mathbf{P}_i\mathbf{P}_i\mathbf{b}$ and $\mathbf{a}^T\mathbf{P}_i\mathbf{P}_i\mathbf{P}_i\mathbf{b}$ for any vectors $\mathbf{a}$ and $\mathbf{b}$ of the right size (**Supplementary Note**).

All the above recursions only involve scalar multiplications, and calculations do not depend on the number of individuals. Therefore, the overall time complexity for GEMMA is $O(mn^2)$ (eigen decomposition of $\mathbf{G}$) + $O(cn^2)$ (evaluations of $\mathbf{v}_{wi}$ and $\mathbf{v}_y$) + $O(pn^2)$ (evaluation of $\mathbf{v}_x$ for each SNP) + $O(ptc^2n)$ (evaluations of the basic quantities for each SNP during each optimization iteration) = $O(mn^2 + cn^2 + pn^2 + ptc^2n)$.

19. Searle, S.R., Casella, G. & McCulloch, C.E. *Variance Components*. (Wiley, New York, 2006).
20. Henderson, C.R. *Applications of Linear Models in Animal Breeding* (University of Guelph, Guelph, Canada, 1984).

doi:10.1038/ng.2310