nature genetics

Rapid variance components–based method for whole-genome association analysis

Gulnara R Svishcheva[1], Tatiana I Axenovich[1], Nadezhda M Belonogova[1], Cornelia M van Duijn[2] & Yurii S Aulchenko[1]

The variance component tests used in genome-wide association studies (GWAS) including large sample sizes become computationally exhaustive when the number of genetic markers is over a few hundred thousand. We present an extremely fast variance components–based two-step method, GRAMMAR-Gamma, developed as an analytical approximation within a framework of the score test approach. Using simulated and real human GWAS data sets, we show that this method provides unbiased estimates of the SNP effect and has a power close to that of the likelihood ratio test–based method. The computational complexity of our method is close to its theoretical minimum, that is, to the complexity of the analysis that ignores genetic structure. The running time of our method linearly depends on sample size, whereas this dependency is quadratic for other existing methods. Simulations suggest that GRAMMAR-Gamma may be used for association testing in whole-genome resequencing studies of large human cohorts.

In most GWAS, the participants are assumed to be unrelated and to come from a single population. However, even in carefully designed studies, some degrees of relatedness and population stratification are unavoidable. As the sample sizes of GWAS become increasingly large, it becomes virtually impossible to circumvent genetic substructure. Ignoring the sample structure increases the rate of false positive association results[1,2].

One of the most flexible and powerful methods of accounting for genetic substructure is the variance component approach that is based on mixed models. Although the method was originally proposed for pedigrees with known relationships[3–5], the variance component approach can be used for samples with unknown relationship when a large number of genetic markers are genotyped, thus allowing inference of genetic structure[6]. Likelihood ratio test (LRT)-based variance component analysis[6–8] is considered to be the gold standard of genetic association analysis using the variance component model. However, the method requires estimation of all model parameters for every tested genetic marker and is computationally demanding.

To solve this problem, a two-stage approach (further named as fast association score test–based analysis or FASTA) was proposed instead of the standard LRT[9]. This approach divides the model parameters into two categories, namely, segregation parameters related to trait heritability and parameters describing the effects of genetic marker(s) on this trait. The segregation parameters are estimated, and the variance-covariance matrix for the phenotypes of study participants is computed once for a given trait. In the second step, the effect of every SNP marker is evaluated, making corrections for the variance-covariance matrix. The two-step approach approximates the LRT well if many loci of small effects are involved in trait determination[9,10]. At the same time, the approach is much less computationally complex than the LRT-based method. A two-stage approach has been exploited in several fast methods, including efficient mixed models expedited (EMMAX)[10], trait analysis by association, evolution and linkage (TASSEL)/population parameters previously determined (P3D)[11] and fast linear mixed models (FaST-LMM)[8].

However, even these methods become rather slow when millions of SNP markers are analyzed in large samples. In cases when effective algorithm based on diagonalization of the relationship matrix is used, the computational complexity of the first step for the two-step methods is $O(n^3) + O(pn)$, where n is a sample size and p is the average number of iterations needed to find parameter estimates (**Table 1**). The term $O(n^3)$ corresponds to the time complexity of eigendecomposition of the relationship matrix, and the second corresponds to the parameter estimation process. As a variant, FaST-LMM[8] makes the performance of the first step of analysis with lower complexity of $O(ns_c^2)$ possible, where $s_c < n$ is the number of markers used to estimate relationship. Methods such as compressed mixed models[11] decrease n by 'compressing' the relationship matrix through identification of almost-equivalent rows, such as twins and siblings. For the second part of the analysis, the time complexity for all two-stage methods (FASTA, EMMAX and FaST-LMM) is linear with the number of markers s but quadratic with the number of individuals n in the study (time complexity of $O(sn^2)$). FaST-LMM can also decrease the complexity to $O(snk)$ if only the top k eigenvectors of the relationship matrix are used.

[1]Institute of Cytology and Genetics, Siberian Division of the Russian Academy of Sciences, Novosibirsk, Russia. [2]Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands. Correspondence should be addressed to Y.S.A. (yurii@bionet.nsc.ru).

Received 16 November 2011; accepted 16 August 2012; published online 16 September 2012; doi:10.1038/ng.2410

Table 1 Complexity of different two-step algorithms

Method	Time complexity	Space complexity
	Step 1	
FASTA	$O(n^3 + pn)$	$O(n^2)$
EMMAX	$O(n^3 + pn)$	$O(n^2)$
FaST-LMM	$O(n^3 + pn)$	$O(n^2)$
FaST-LMM-restricted	$O(ns_c^2 + pn)$	$O(ns_c)$
GRAMMAR-Gamma	$O(n^3 + pn)$	$O(n^2)$
	Step 2	
FASTA	$O(sn^2)$	$O(n)$
EMMAX	$O(sn^2)$	$O(n)$
FaST-LMM	$O(sn^2)$	$O(n)$
FaST-LMM-restricted	$O(snk)$	$O(n)$
GRAMMAR-Gamma	$O(sn)$	$O(n)$

n, sample size; s, number of tested SNPs; s_c, number of SNPs used for singular value decomposition; p, average number of iterations needed to find parameter estimates; k, number of eigenvectors used.

To further increase the computation speed, we have previously proposed the GRAMMAR method using environmental residuals estimated from the segregation model as a transformed trait[12,13]. Further analysis assumes independence among these transformed phenotypes of the study participants. However, GRAMMAR, although being computationally very fast, produces a conservative test and biased effect estimates[12].

Here, we describe a new variance component–based two-step method, GRAMMAR-Gamma, which is a fast approximation of FASTA. We show analytically that the ratio of GRAMMAR and FASTA tests can be approximated by a constant named the GRAMMAR-Gamma factor (Online Methods). Using this factor allows correction of the test statistic and SNP effect estimates obtained by the fast GRAMMAR approach and compensates for its conservativeness. Thus, the new method operates as fast as the original GRAMMAR approach and achieves the power of FASTA, which, in turn, is almost equivalent to the gold standard LRT-based variance component analysis.

In the first step, the new method estimates the segregation parameters and GRAMMAR-Gamma factor, and the trait is transformed. This step is accelerated by analytical optimization of the matrix operations on the basis of the eigendecomposition of the relationship matrix and analytical parameter estimation. Being mostly determined by the eigendecomposition, the computational complexity of this step is similar to that of other two-step methods, such as EMMAX[10] and FaST-LMM[8] (**Table 1**).

In the second step, the score test for the association between the transformed trait and genotypes is performed without the explicit use of the relationship matrix, and the resulting SNP effect estimates and test statistic values are corrected by dividing them by the GRAMMAR-Gamma correction factor. Computational complexity for the second step is linearly dependent on the number of individuals and markers (time complexity of $O(ns)$) and is much lower compared to that of other methods using the variance component model (**Table 1**).

We evaluated our method using a human GWAS data set from the Erasmus Rucphen Family (ERF) study[14], which is embedded into a genetically isolated population. Real genotypes in combination with simulated and real phenotypes were used for association analysis. In addition, the method was applied to five traits of *Arabidopsis thaliana* (published data of Atwell *et al.*[15]). This sample is very different from the human sample in its structure and size, heritability of traits and size of SNP effects.

RESULTS

Statistical properties of GRAMMAR-Gamma

We compared statistical properties of the new GRAMMAR-Gamma method with those of the FASTA and LRT-based variance component methods (see Online Methods for details).

We contrasted the type 1 error, power and SNP effect estimates using the approach described in previous publications[12,13]. The results are shown in **Figure 1** for selected scenarios and in **Supplementary Tables 1–4** for the extended set of methods and all scenarios. The estimates of SNP effect produced by GRAMMAR-Gamma were very close to those from FASTA and the LRT-based methods (two-sided Student's t-test P values > 0.22 and 0.11 for all scenarios, respectively) and to the simulated values of the SNP effects (all $P > 0.19$). The values of the GRAMMAR-Gamma test statistic were also very close to those from FASTA (all $P > 0.8$) and were only slightly but not significantly lower than the values from the LRT-based method (all $P > 0.11$). All ratios of the non-centrality parameter between GRAMMAR-Gamma and FASTA and LRT-based methods were greater than 0.997 and 0.97, respectively.

Thus, GRAMMAR-Gamma addresses the limitations of the original GRAMMAR approach. It gives unbiased estimates of the SNP effect and, therefore, the correct distribution of the test statistic. GRAMMAR-Gamma has a power that is essentially identical to that of FASTA and close to that of the LRT-based method.

Distribution of GRAMMAR-Gamma correction factor

The only assumption that should hold for good approximation of FASTA by our new method is that the GRAMMAR-Gamma correction factor (the ratio of the FASTA and GRAMMAR test statistics) is approximately the same for all markers. To examine this assumption empirically, we analyzed 6 traits in a large human pedigree (ERF study[14], sample size from 2,568 to 2,592) and 5 traits of *A. thaliana* (data from a previous study[15], sample size from 84 to 164) (see details in Online Methods). The marker-specific GRAMMAR-Gamma factors (expression (5) in Online Methods) were estimated using FASTA implemented in the 'mmscore' function of the GenABEL package[16]. The variance of

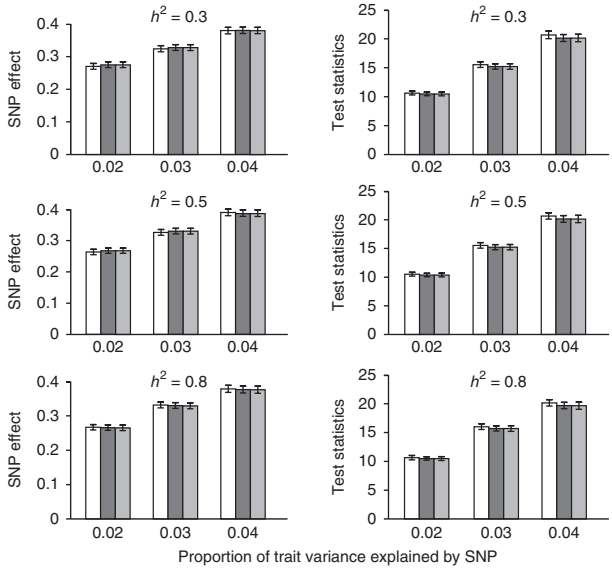

Figure 1 Comparison of mixed model–based methods. SNP effect estimates (left) and test statistics (right) obtained using LRT-based FMM (white), FASTA (dark gray) and GRAMMAR-Gamma (light gray) methods under nine selected simulation scenarios (modeled covariate effect = 0.1). Error bars, s.e.m.

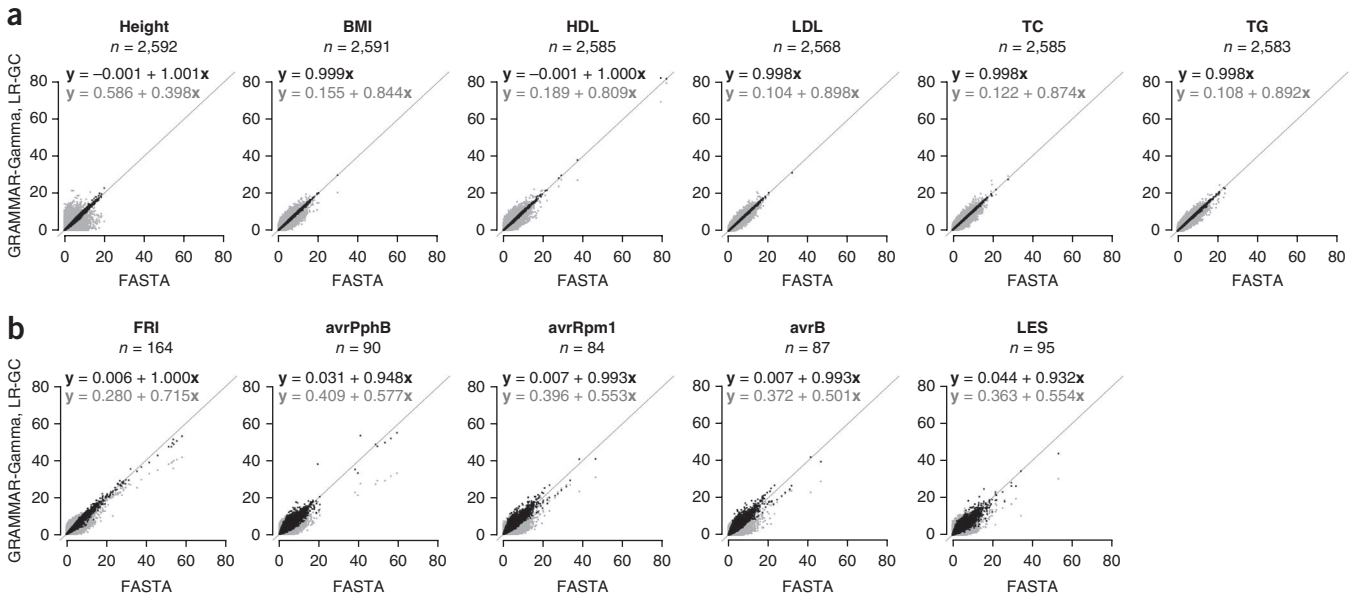

Figure 2 Correspondence between GRAMMAR-Gamma and FASTA test statistics. (**a**) Human data. (**b**) *A. thaliana* data. For comparison, the correspondence between linear regression with genomic control (LR-GC) and FASTA is shown in gray. Equations show the relationship between statistics ($\mathbf{y} = \mu + \beta\mathbf{x}$, where **y** is a vector of genome-wide test statistics obtained with GRAMMAR-Gamma or LR-GC and **x** is a vector of test statistics obtained with FASTA). Standard errors are <0.005 for all regression coefficients. Lines indicate one-to-one correspondence.

the GRAMMAR-Gamma factor was small for all human traits but was substantially higher for *A. thaliana* (**Supplementary Fig. 1**).

Association analysis of real traits

We performed GWAS on the selected human and *A. thaliana* traits using GRAMMAR-Gamma, FASTA, LRT, EMMAX and a method that does not incorporate knowledge about genetic (sub)structure, namely, linear regression with genomic control (LR-GC). In LR-GC, the test statistic from simple linear regression was corrected by dividing it by the genomic control inflation factor[17].

For all human traits studied, GRAMMAR-Gamma yielded genomic control inflation factors, *P* values and SNP effect estimates that were almost equivalent to those obtained using FASTA and EMMAX (correlation of test statistic values ranged from 0.997 to 1.0 for different traits) and LRT (correlation from 0.996 to 0.999) (**Fig. 2a** and **Supplementary Tables 5** and **6**).

The samples in the *A. thaliana* data set consist of pure lines, the phenotypes of which are averages across homozygous individuals within the lines. The data set is characterized by high heritabilities (0.84–1.00) and small sample sizes (84–164). However, despite these features, the statistics and SNP effect estimates of GRAMMAR-Gamma were close to those obtained using FASTA and EMMAX (correlation of test statistic values ranged from 0.932 to 0.985 for different traits) (**Fig. 2b** and **Supplementary Table 5**). Correlations of test statistic values obtained by LRT and GRAMMAR-Gamma were significantly smaller (from 0.629 to 0.953). However, similar correlations (from 0.690 to 0.969) were observed between test statistic values obtained by LRT and FASTA (**Supplementary Table 5**). This indicates that a two-step score-based approach may not be correct for approximating LRT when a trait with very high heritability is analyzed using small data sets with highly heterogeneous relationships.

Using GRAMMAR-Gamma for association testing, we replicated associations with all loci reported in a previous *A. thaliana* study[15] (**Supplementary Table 6**). We found good agreement between top *P* values obtained by FASTA, GRAMMAR-Gamma and LRT, whereas *P* values from EMMAX were several orders of magnitude lower. We suggest that different statistical tests, which are asymptotically equivalent (as we see with the large human data set), could give different results because of the small size of *A. thaliana* data set. For example, for association between the marker snp-1-4143161 and the avrPphB phenotype we obtained *P* values ranging from 2×10^{-14} to 4×10^{-13} when using FASTA, GRAMMAR-Gamma and LRT, but we obtained a *P* value of 6×10^{-22} when using EMMAX (**Supplementary Table 6**) and an even lower *P* value of 8×10^{-38} when using the Wald test–based analog of FASTA in the genome-wide feasible generalized least squares (GWFGLS) function of MixABEL.

To show the performance of GRAMMAR-Gamma approximation of FASTA, we also compared it with the other fast approximation, LR-GC. In this study, 95% of all absolute differences between the FASTA and GRAMMAR-Gamma test statistic values were <0.053 for low-density lipoprotein (LDL) concentration and <0.13 for human height. The absolute differences between the FASTA and LR-GC test statistics, however, were at least an order of magnitude larger (the 95% percentile was 1.40 for LDL and 3.32 for height). For the *A. thaliana* lesioning (LES) phenotype, where GRAMMAR-Gamma factors were quite variable, the 95% percentile of absolute differences between the FASTA and GRAMMAR-Gamma test statistic values was relatively large (0.67). However, this value was much smaller than that from LR-GC (2.71).

Running time

Finally, we addressed a question of the running time of practical implementations of different two-stage variance component methods. EMMAX[10], FaST-LMM[8] and FASTA implemented in the GenABEL package (mmscore) and the GRAMMAR-Gamma test in the current study were compared. The analyses were run on data from 500 to 3,000 individuals at 456,516 SNP markers (Online Methods) using a single-core Intel Xeon X5550 at 2.67 GHz with 36 GB of random-access memory (RAM).

The measured elapsed real times are presented in **Figure 3**. The proposed GRAMMAR-Gamma implementation provided the fastest

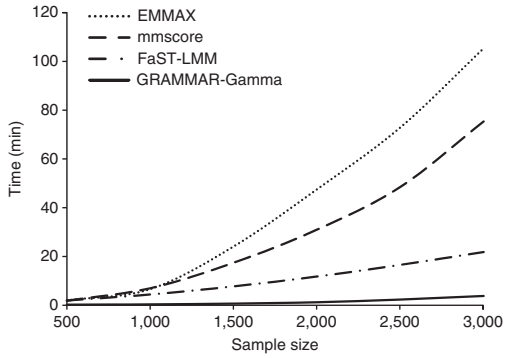

Figure 3 Run time on a single processor for different two-step methods, namely, EMMAX, mmscore realization of FASTA, FaST-LMM and GRAMMAR-Gamma. We tested 456,516 SNP markers for association in all runs.

means to run GWAS using the approximate variable component method. Compared with EMMAX and FaST-LMM (the two-stage option where the parameters are estimated only once for the null model), GRAMMAR-Gamma achieved increases in speed of up to 38 times and 10 times, respectively, on this data set. According to the nature of the algorithm, the more individuals and genetic markers that are analyzed, the larger the expected increase in speed will be compared to other methods.

Although the scenario above assumes the use of a SNP array, one of the current challenges in statistical genomics is the analysis of whole-genome resequencing data. Therefore, a scenario in which 36.5 million SNPs were available in 3,000 individuals was investigated (Online Methods). Using the new method, the analysis of this data set was completed in 38 min.

DISCUSSION

In the current study, we report a new fast method for GWAS, which is based on the variance component mixed-model approach and corrects for sample structure. The new method is the fast approximation of the two-step score test–based method proposed by Chen and Abecasis[9].

The two-stage approach is widely used to speed up GWAS when hundreds of thousands of SNPs are analyzed in large samples[8,10,11]. The recent works concentrated on increasing the speed of the first step of analysis where segregation parameters are estimated. The fastest approaches for the first step use additional algorithms, such as clustering individuals into groups (compressed mixed model)[11] or reducing the relationship matrix (singular value decomposition)[8]. At the same time, computational complexity of the second step in different methods increases quadratically with the number of individuals in the study (**Table 1**) because the $n \times n$ variance-covariance matrix is involved in the calculations. When the number of genetic markers reaches millions, the time complexity of the second step becomes restrictive. The main advantage of our new method is the reduction of computational complexity of this second step to a point close to theoretical minimum (defined by the complexity of the analysis of samples without genetic structure).

Two different approaches with computational complexity close to the minimum, particularly genomic control[17,18] and GRAMMAR[12], have been proposed earlier. The former ignores the genetic structure of the sample and considers correlated phenotypes and genotypes as independent values but adjusts the test statistic values by the whole-genome inflation factor, which is empirically calculated[17,18]. However, a study by Kang *et al.*[10] and the current study (**Fig. 2**) show that the genomic control approach leads to a substantial decrease in power. The latter approach, GRAMMAR, takes into account correlation between phenotypes, but it ignores correlation between genotypes[12]. Although the test statistic can be adjusted through genomic control, SNP effects are underestimated[13].

The current study analytically shows that the test statistics and SNP effect estimates from GRAMMAR and FASTA differ from one another by the GRAMMAR-Gamma factor. The effectiveness of the new method is shown using both simulated and real human data. The statistical properties of the new method are practically identical to the statistical properties of FASTA and are very close to that of the gold standard LRT-based variance component method. At the same time, the new method is extremely fast. The running time depends linearly on sample size, whereas dependency is quadratic for the majority of existing two-step methods.

The only assumption that needs to hold for GRAMMAR-Gamma approximation of FASTA is that marker-specific GRAMMAR-Gamma factors (equation (5) in Online Methods) do not vary much between the markers. The smaller the variability of the GRAMMAR-Gamma factors, the closer the GRAMMAR-Gamma statistic to those derived using the FASTA test. If the variability is high, the GRAMMAR-Gamma estimates and statistics (although asymptotically unbiased) may deviate from FASTA estimates for particular markers.

The assumption about low GRAMMAR-Gamma factor variability seems to hold well for large samples from human populations with hidden relationships. These samples require very fast methods for GWAS. However, caution needs to be exercised when GRAMMAR-Gamma is used for small data sets with highly heterogeneous relationships, such as the example *A. thaliana* data set. For this sample, the variance of the GRAMMAR-Gamma correction factor was substantially higher compared to the variance observed for human data.

The simplest way to evaluate GRAMMAR-Gamma applicability before analysis is to calculate GRAMMAR-Gamma factors for a subset of markers. This distribution can be used to estimate, for example, the 95% confidence intervals of the distribution of differences between the FASTA and GRAMMAR-Gamma test statistic values. The choice of method depends on the compromise between desired speed and accuracy.

The algorithm in the current study can easily be integrated into the second step of any two-step variance component–based method. Thus, the increase in speed can be achieved by combining the fastest method for the second step with the fastest approaches for the first step that uses additional algorithms. Use of the method proposed in the current study is expected to increase in the future because of the rapid generation of whole-genome resequencing data. The GRAMMAR-Gamma method is available through the 'grammar' procedure of the GenABEL package v 1.7-1 or later.

URLs. The GRAMMAR-Gamma method is available as the grammar procedure within the GenABEL package v 1.7-1 or later (GenABEL project, http://www.genabel.org/). The FASTA method is available in the GenABEL package (mmscore procedure). The FMM implementation of variance component LRT (developed by W. Astle) is available as the fmm procedure of the MixABEL package (part of the GenABEL project). EMMAX software is available at http://genetics.cs.ucla.edu/emmax/. FaST-LMM software is available at http://fastlmm.codeplex.com/. The *A. thaliana* data set used here is accessible at AtPolyDB (https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb).

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

We thank A. Kirichenko, D. Fabregat Traver and P. Bientinesi for technical support and advice and M. Axenovich, D. Balding, P. Borodin and W. Astle for discussion. This work was supported by grants from the Russian Foundation for Basic Research (RFBR) Programs of the Russian Academy of Sciences and the RFBR-Helmholtz Joint Research Groups program (research project 12-04-91322-СИГ_а).

AUTHOR CONTRIBUTIONS

G.R.S. developed the GRAMMAR-Gamma statistical test, ran the simulations and analyzed the simulated data. N.M.B. analyzed human and *A. thaliana* data and designed figures and tables. C.M.v.D. provided the human data and supervised its analyses. T.I.A. and Y.S.A. jointly designed and supervised the project and wrote the paper. All authors contributed to critical review of the manuscript during its preparation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2410.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J. & Stefánsson, K. An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* **37**, 90–95 (2005).
2. Astle, W. & Balding, D.J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
3. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
4. Henderson, C.R. Estimation of variance and covariance components. *Biometrics* **9**, 226–252 (1953).
5. Boerwinkle, E., Chakraborty, R. & Sing, C.F. The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann. Hum. Genet.* **50**, 181–194 (1986).
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. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
8. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
9. Chen, W.M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
10. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
11. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
12. 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).
13. Amin, N., van Duijn, C.M. & Aulchenko, Y.S. A genomic background based method for association analysis in related individuals. *PLoS ONE* **2**, e1274 (2007).
14. Pardo, L.M. *et al.* The effect of genetic drift in a young genetically isolated population. *Ann. Hum. Genet.* **69**, 288–295 (2005).
15. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
16. Aulchenko, Y.S. *et al.* GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
17. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
18. Bacanu, S.A., Devlin, B. & Roeder, K. Association studies for quantitative traits in structured populations. *Genet. Epidemiol.* **22**, 78–93 (2002).

ONLINE METHODS

Model. A standard mixed linear model for association analysis of quantitative traits can be written as $\mathbf{y} = \mu + X\boldsymbol{\beta}_X + \mathbf{g}\beta_g + \mathbf{u} + \mathbf{e}$, where $\mathbf{y}$ $(n \times 1)$ and $\mathbf{g}$ $(n \times 1)$ are vectors of phenotypes and genotypes of an analyzed SNP marker, respectively, observed for n individuals; X $(n \times k)$ is a matrix of k observed covariates; μ is an intercept; β_g is fixed effect for an analyzed SNP; $\boldsymbol{\beta}_X$ $(k \times 1)$ is a vector of covariate effects; and $\mathbf{u}$ $(n \times 1)$ and $\mathbf{e}$ $(n \times 1)$ are vectors of unobserved polygenic and random residual effects, respectively. For the analyzed marker having two alleles, A and a, a genotype score, g, is defined as 1, 0.5 and 0 for the AA, Aa and aa genotypes, respectively.

This model assumes that the phenotypes follow a multivariate normal distribution and a logarithm of the likelihood (log likelihood) is given by

$$\log l = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Omega| - \frac{1}{2}\left(\mathbf{y} - E(\mathbf{y})\right)^T \Omega^{-1}\left(\mathbf{y} - \mathrm{E}(\mathbf{y})\right)$$

where $\mathrm{E}(\mathbf{y})$ is a mean vector

$$E(\mathbf{y}) = X\boldsymbol{\beta}_X + \mathbf{g}\beta g$$

and Ω is a covariance matrix

$$\Omega = \sigma^2(h^2R + (1-h^2)I) \quad (1)$$

where σ^2 is a total trait variance; h^2 is a trait heritability; and R and I are a relationship and an identity matrix, respectively. Relationship coefficients between relatives are defined by the pedigree structure or may be estimated on the basis of the large number of genotyped markers (genomic relationship).

Score-based association test. Use of the score approach allows estimation of all model parameters (except for β_g) only once for a given trait. Then, a score statistic for a given marker is defined via one parameter, β_g, by

$$T^2_{\mathrm{score}} = \frac{\beta_g^2}{\mathrm{var}(\beta_g)}$$

where $\mathrm{var}(\beta_g)$ is the variance of β_g. This statistic is approximately distributed as χ^2 with 1 degree of freedom.

Estimates

$$\beta_g = \frac{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{y}}}{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{g}}}$$

and

$$\mathrm{var}(\beta_g) = \frac{1}{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{g}}}$$

obtained as a result of maximization of the log likelihood under the alternative hypothesis lead to the score statistic proposed by Chen and Abecasis[9]

$$T^2_{\mathrm{score}} = \frac{\left(\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{y}}\right)^2}{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{g}}} \quad (2)$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \mathrm{E}(\mathbf{y})$ and $\tilde{\mathbf{g}} = \mathbf{g} - \mathrm{E}(\mathbf{g})$.

The time complexity of the score test for each genetic marker is a quadratic function of the sample size. A lot of time could be required for GWAS, when millions of genetic markers are analyzed in large samples. To speed up computations, we have recently proposed a fast approximation of the score test–based method named the GRAMMAR approach[12]. Its improved speed is the result of transforming the vector of dependent phenotype residuals $\tilde{\mathbf{y}}$ into the vector of the independent ones, $\tilde{\mathbf{y}}^* = \sigma_e^2\Omega^{-1}\tilde{\mathbf{y}}$. Then $\tilde{\mathbf{y}}^*$ can be analyzed by standard linear regression methods using the following score statistic:

$$T^2_{\mathrm{GRAMMAR}} = \frac{n\left(\tilde{\mathbf{g}}^T\tilde{\mathbf{y}}^*\right)^2}{\left(\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}\right)\left(\tilde{\mathbf{y}}^{*T}\tilde{\mathbf{y}}^*\right)} \quad (3)$$

This equation holds because the GRAMMAR test estimates the effect of a quantitative trait locus (QTL) and its variance as

$$\beta_{g_{\mathrm{GRAMMAR}}} = \frac{\tilde{\mathbf{g}}^T\tilde{\mathbf{y}}^*}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}$$

and

$$\mathrm{var}\left(\beta_{g_{\mathrm{GRAMMAR}}}\right) = \frac{\tilde{\mathbf{y}}^{*T}\tilde{\mathbf{y}}^*}{n\left(\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}\right)}$$

respectively. However, it was demonstrated that the GRAMMAR test is conservative and gives biased estimates of marker effect[13].

The GRAMMAR-Gamma test. Here, we describe a new test, which is based on the GRAMMAR method but, unlike GRAMMAR, results in unbiased test statistic and marker effect estimates.

The score test (2) can be presented as follows:

$$T^2_{\mathrm{score}} = \frac{\left(\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{y}}\right)^2}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}} \Bigg/ \left(\frac{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{g}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}\right) \quad (4)$$

The numerator of equation (4) is a new statistic

$$T^2_{\mathrm{new}} = \frac{\left(\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{y}}\right)^2}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}$$

which is similar to equation (3) and does not take into consideration the correlations within genotype data $\tilde{\mathbf{g}}$. In this case, transforming the phenotype data by use of $\Omega^{-1}\tilde{\mathbf{y}}$ makes them uncorrelated with the genotypic data. Then, standard linear regression methods can be used to calculate this statistic. The denominator of equation (4)

$$\frac{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{g}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}$$

is indeed a correction factor, which allows us to obtain the score statistic for the analyzed marker from this new statistic.

The denominator of expression (4) for genetic marker m can be written as

$$\gamma_m = \frac{\sum_{i,j=1}^{n}\left(g_{mi} - \mathrm{E}(g_m)\right)\omega_{ij}^{-1}\left(g_{mj} - \mathrm{E}(g_m)\right)}{\sum_{l=1}^{n}\left(g_{ml} - \mathrm{E}(g_m)\right)^2} \quad (5)$$

where subscripts i and j define a pair of relatives, ω^{-1} is an element of the Ω^{-1} matrix, $\mathrm{E}(g_m)$ is a genotype mean and $\frac{1}{n-1}\sum_{l=1}^{n}\left(g_{ml} - \mathrm{E}(g_m)\right)^2$ is a genotype variance, $\mathrm{var}(g_m)$.

We assume here that, if many loci with small effects control the trait, then the correction factors (5) for different markers $(m = \overline{1, M})$ are similar to each other. This assumption follows from observing the linear dependence between statistics (2) and (3), which has been demonstrated empirically[12]. In this case, we can introduce the GRAMMAR-Gamma factor γ, which is equal to the arithmetic mean

$$\gamma = \frac{1}{M}\sum_{m=1}^{M}\gamma_m$$

leading to the following expression for the factor:

$$\gamma = \frac{1}{M}\sum_{m=1}^{M}\frac{\sum_{i,j=1}^{n}\left(g_{mi} - \mathrm{E}(g_m)\right)\omega_{ij}^{-1}\left(g_{mj} - \mathrm{E}(g_m)\right)}{\sum_{l=1}^{n}\left(g_{ml} - \mathrm{E}(g_m)\right)^2} \quad (6)$$

Because permuting the operations of summation on markers and individuals does not change the value of γ, expression (6) can be rewritten as follows:

$$\gamma = \frac{1}{n-1}\sum_{i,j=1}^{n}\omega_{ij}^{-1}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\left(g_{mi}-\mathrm{E}(g_m)\right)\left(g_{mj}-\mathrm{E}(g_m)\right)}{\mathrm{var}(g_m)}\right]$$

The expression in square brackets defines the r_{ij} genomic relationship between i and j relatives[2,13], resulting in the following definition of the GRAMMAR-gamma factor:

$$\gamma = \frac{1}{n-1}\sum_{i,j=1}^{n}\omega_{ij}^{-1}r_{ij}$$

As follows from expression (1),

$$R = \frac{1}{\sigma^2 h^2}\left(\Omega - \sigma^2\left(1-h^2\right)I\right)$$

and

$$\gamma = \frac{1}{\sigma^2 h^2}\left(1 - \frac{1-h^2}{n-1}\mathrm{trace}\left(V^{-1}\right)\right) \qquad (7)$$

were V^{-1} is an inverse correlation matrix, $V^{-1} = \sigma^2\Omega^{-1}$. As can be seen from expression (7), the factor γ does not depend on marker information and can be analytically defined through h^2 and σ^2. Therefore, it can be estimated only once at the first step of analysis.

Thus, the score statistic (2) is approximated by the new statistic

$$T^2_{\mathrm{new}} = \frac{\left(\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{y}}\right)^2}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}$$

which results from linear regression analysis, divided on the correction factor γ (7).

$$T^2_{\mathrm{score}} \approx \frac{1}{\gamma}T^2_{\mathrm{new}}$$

In light of the new statistic, the QTL effect and its variance are estimated as

$$\beta_{g_{\mathrm{new}}} \approx \frac{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{y}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}$$

and

$$\mathrm{var}\left(\beta_{g_{\mathrm{new}}}\right) = \frac{\tilde{\mathbf{y}}^T\Omega^{-1}\tilde{\mathbf{y}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}$$

respectively. These estimates are biased because they do not take into consideration the correlations within genotype data $\tilde{\mathbf{g}}$. It is easy to show that γ is also the correction factor for a new estimates of the QTL effect and its variance.

$$\beta_g = \left(\frac{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{y}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}\right)\Big/\left(\frac{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{g}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}\right) \approx \frac{1}{\gamma}\beta_{g_{\mathrm{new}}}$$

$$\mathrm{var}\left(\beta_g\right) = \left(\frac{\tilde{\mathbf{y}}^T\Omega^{-1}\tilde{\mathbf{y}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}\right)\Big/\left(\frac{\tilde{\mathbf{g}}^T\Omega^{-1}\tilde{\mathbf{g}}}{\tilde{\mathbf{g}}^T\tilde{\mathbf{g}}}\right) \approx \frac{1}{\gamma}\mathrm{var}\left(\beta_{g_{\mathrm{new}}}\right)$$

Fast calculations under the null hypothesis. To accelerate the first step of the method in the current study, where segregation parameters are estimated using a maximum-likelihood approach, two algorithms are proposed. The first is the use of the theory of eigenanalysis in likelihood calculation to replace 'expensive' operations over matrices with 'inexpensive' operations over vectors. The second is the implementation of the analytical procedure for estimation of all model parameters, except for the heritability. The study demonstrates that a system of equations with $(k+3)$ variables, where k is the number of covariates, may be analytically reduced to an equation with a single variable. The description of these algorithms is presented in the **Supplementary Note**.

Comparison of methods. The majority of LRT-based methods implementing the variance component mixed model for GWAS are rather slow when large samples and great numbers of genotypes are analyzed. Two LRT-based methods, FMM[19] and FaST-LMM_full[8], run relatively fast and can be used for analysis of large data sets. The test statistics obtained for simulated QTLs in 15 experiments using FMM and FaST-LMM_full were compared. Both methods gave identical values of the test statistics (Pearson's correlation was 0.99999) (**Supplementary Fig. 2**). However, FMM was running about four times faster than FaST-LMM_full. Therefore, FMM is the method of choice for the comparison. The mmscore function of GenABEL software[16] was used to implement FASTA, which is described in detail in the score based–association test section. The methods EMMAX[10], which is a two-step approximation of the LRT-based method EMMA[7], and FaST-LMM[8], where parameters of the null hypothesis are estimated only once for every trait, were used to compare the running time and the implementation of additional testing of statistical properties.

Simulated data. For the simulation study, real data from the ERF study, which was performed among a young genetically isolated Dutch population[14], were used. All study protocols were approved by the Medical Ethics Committee of Erasmus University, and all participants gave written informed consent in accordance with the Declaration of Helsinki. The trait was simulated on the basis of real genotypes as a sum of four independent effects, namely, QTL, polygenic, fixed covariate and residual random effects. Several scenarios concerning three parameters, namely, total trait heritability (0.3, 0.5 and 0.8), proportion of variance explained by a QTL (0.02, 0.03 and 0.04) and covariates (0.01, 0.1 and 0.5), were considered for the tests.

To estimate the type 1 error and power, 1,000 replicas of phenotypes using 50,000 SNPs randomly distributed over the autosomes of 500 individuals from the ERF study were generated.

To estimate the running time of the different methods, the size of the data set was made to vary from 500 to 3,000 individuals with a step size of 500. The genotypes of 456,516 SNP markers with minor allele frequency (MAF) of >1% were included in the analysis.

To estimate the running time of the analysis of the whole-genome resequencing data, 36.5 million genotypes from 3,000 people were used. The genotype data were generated as 80 repeats of real genotypes of 456,516 SNP markers.

Real data. The data used for the analysis of human traits were a part of the ERF study described in the previous section. The sample included data for 2,596 individuals with a call rate of ≥0.95 genotyped at 239,843 SNP markers with MAF of ≥0.05 and call rate of ≥0.99. Six traits were analyzed, namely, height, body mass index (BMI) and serum levels of high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol and triglycerides. All traits were adjusted for age and sex.

The *A. thaliana* data analyzed in this study are described in ref. 15 and are freely available for access at AtPolyDB (see URLs). We tested 206,603 SNP markers for association with the 5 traits that showed the strongest association signals[15], namely, *FRI* gene expression (FRI), three phenotypes describing the hypersensitive response for bacterial inoculation (avrPphB, avrRpm1 and avrB) and presence or absence of lesioning (LES).

19. Astle, W. *Population Structure and Cryptic Relatedness in Genetic Association Studies*. PhD Thesis, University of London (2009).

doi:10.1038/ng.2410