FaST linear mixed models for genome-wide association studies

Christoph Lippert[1–3], Jennifer Listgarten[1,3], Ying Liu[1], Carl M Kadie[1], Robert I Davidson[1] & David Heckerman[1,3]

We describe factored spectrally transformed linear mixed models (FaST-LMM), an algorithm for genome-wide association studies (GWAS) that scales linearly with cohort size in both run time and memory use. On Wellcome Trust data for 15,000 individuals, FaST-LMM ran an order of magnitude faster than current efficient algorithms. Our algorithm can analyze data for 120,000 individuals in just a few hours, whereas current algorithms fail on data for even 20,000 individuals (http://mscompbio.codeplex.com/).

The problem of confounding by population structure, family structure and cryptic relatedness in genome-wide association studies (GWAS) is widely appreciated[1–7]. Statistical methods for correcting these confounders include linear mixed models (LMMs)[2–10], genomic control, family-based association tests, structured association and Eigenstrat[7]. In contrast to other methods, LMMs can capture all of these confounders simultaneously, without knowledge of which are present and without the need to tease them apart[7]. Unfortunately, LMMs are computationally expensive relative to simpler models. In particular, the run time and memory footprint required by these models scale as the cube and square of the cohort size (the number of individuals represented in the dataset), respectively. This bottleneck means that LMMs run slowly or not at all on currently or soon to be available large datasets.

Roughly speaking, LMMs tackle confounders by using measures of genetic similarity to capture the probabilities that pairs of individuals have causative alleles in common. Such measures include those based on identity by descent[10,11] and the realized relationship matrix (RRM)[9,10,12], and have been estimated with a small sample of markers (200–2,000 markers)[2,4]. Here we take advantage of such sampling to make LMM analysis applicable to extremely large datasets, introducing a reformulation of LMMs called factored spectrally transformed LMM (FaST-LMM). We show that, provided (i) the number of single-nucleotide polymorphisms (SNPs) used to estimate genetic similarity is less than the cohort size (regardless of how many SNPs are to be tested) and (ii) the RRM is used to determine these similarities, then FaST-LMM produces exactly the same results as a standard LMM but with a run time and memory footprint that is only linear in the cohort size. FaST-LMM thus dramatically increases the size of datasets that can be analyzed with LMMs and additionally makes currently feasible analyses much faster.

Our FaST-LMM algorithm builds on the insight that the maximum likelihood (or the restricted maximum likelihood (REML)) of an LMM can be rewritten as a function of just a single parameter, δ, the ratio of the genetic variance to the residual variance[3,13]. Consequently, the identification of the maximum likelihood (or REML) parameters becomes an optimization problem over δ only. The algorithm 'efficient mixed model association' (EMMA)[3] speeds up the evaluation of the log likelihood for any value of δ, which is ordinarily cubic in the cohort size, by clever use of spectral decompositions. However, the approach requires a new spectral decomposition for each SNP tested (a cubic operation). The algorithms 'EMMA expedited' (called EMMAX) and 'population parameters previously determined' (called P3D)[4,5] provide additional computational savings by assuming that variance parameters for each tested SNP are the same, removing the expensive cubic computation per SNP.

In contrast to these methods, FaST-LMM requires only a single spectral decomposition to test all SNPs, even without assuming variance parameters to be the same across SNPs, and offers a decrease in memory footprint and additional speedups. A key insight behind our approach is that the spectral decomposition of the genetic similarity matrix makes it possible to transform (rotate) the phenotypes, SNPs to be tested and covariates in such a way that the rotated data become uncorrelated. These data are then amenable to analysis with a linear regression model, which has a run time and memory footprint linear in the cohort size.

In general, the number of entries in the required rotation matrix is quadratic in the cohort size, and computing this matrix by way of a spectral decomposition has a cubic run time in the cohort size. When the number of SNPs used to construct the genetic similarity matrix is less than the cohort size, however, the number of entries in the matrix required to perform the rotations is linear in the cohort size (and linear in the number of SNPs), and the time required to compute the matrix is linear in the cohort size (and quadratic in the number of SNPs). Intuitively, these savings can be achieved because the intrinsic dimensionality of the space spanned by the SNPs used to construct the similarity matrix can never be higher than the smaller of the number of such SNPs and the cohort size. Thus, we can always perform operations in the smaller space without any loss of information, and the computations remain exact. This basic idea has been exploited

[1]Microsoft Research, Los Angeles, California, USA. [2]Max Planck Institutes Tübingen, Tübingen, Germany. [3]These authors contributed equally to this work. Correspondence should be addressed to C.L. (christoph.lippert@tuebingen.mpg.de), J.L. (jennl@microsoft.com) or D.H. (heckerma@microsoft.com).

RECEIVED 5 APRIL; ACCEPTED 2 AUGUST; PUBLISHED ONLINE 4 SEPTEMBER 2011; DOI:10.1038/NMETH.1681

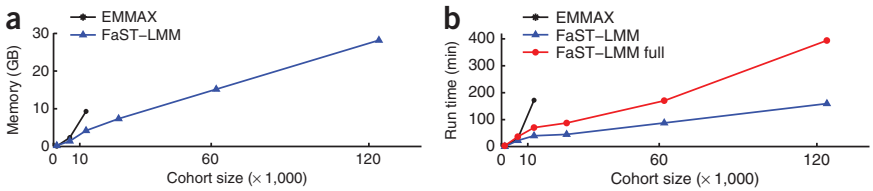

Figure 1 | Computational costs of FaST-LMM and EMMAX. (**a**,**b**) Memory footprint (**a**) and run time (**b**) of the algorithms running on a single processor as a function of the cohort size in synthetic datasets based on GAW14 data. In each run, we used 7,579 SNPs both to estimate genetic similarity (RRM for FaST-LMM and identity by state for EMMAX) and to test for association. In the 'FaST-LMM full' analysis, the variance parameters were re-estimated for each test, and in the FaST-LMM analysis these parameters were estimated only once for the null model, as in EMMAX. FaST-LMM and FaST-LMM full had the same memory footprint. EMMAX would not run on the datasets that contained 20 or more times the cohort size of the GAW14 data because the memory required to store the large matrices exceeded the 32 GB available.

previously[8,14] but would require expensive computations per SNP when applied to GWAS, making these approaches far less efficient than FaST-LMM.

To achieve our linear run time and memory footprint, the spectral decomposition of the genetic similarity matrix must be computable without the explicit computation of the matrix itself. The RRM has this property as do other matrices (**Supplementary Note 1**). A more formal description of FaST-LMM is available in Online Methods.

We compared memory footprint and run time for non-parallelized implementations of the FaST-LMM and EMMAX algorithms (**Fig. 1**). (The EMMAX implementation was no less efficient in terms of run time and memory use than that of P3D in the 'trait analysis by association, evolution and linkage' (TASSEL) package). In the comparison, we used Genetic Analysis Workshop 14 data (GAW14 data; Online Methods) to construct synthetic datasets with the same number of SNPs (~8,000 SNPs) and roughly 1, 5, 10, 20, 50 and 100 times the cohort size of the original data. The largest such dataset contained data for 123,800 individuals. We tested all SNPs and used them all to estimate genetic similarity. EMMAX would not run on the 20×, 50× or 100× datasets because the memory required to store the large matrices exceeded the 32 gigabytes (GB) available. In contrast, FaST-LMM, which did not require these matrices (because it bypassed their computation, using them only implicitly), completed the analyses using 28 GB of memory on the largest dataset. Run-time results highlight the linear dependence of the computations on the cohort size when that size exceeded the 8,000 SNPs used to construct the RRM. Also, computations remained practical using our approach even when we re-estimated the variance parameters for each test.

It is known that the LMM with no fixed effects using an RRM constructed from a set of SNPs is equivalent to a linear regression of the SNPs on the phenotype, with weights integrated over independent normal distributions with the same variance[9,10]. In this view, sampling SNPs for construction of the RRM can be seen as the omission of regressors and hence an approximation. Nonetheless, SNPs could be sampled uniformly across the genome so that linkage disequilibrium would diminish the effects of sampling. To examine this issue, we compared association *P* values with and without sampling on the Wellcome Trust Case Control Consortium (WTCCC) data for Crohn's disease. Specifically, we tested all SNPs on chromosome 1 while using SNP sets of various sizes from all but this chromosome (the complete set (~340,000 SNPs) and uniformly distributed samples of ~8,000 SNPs and ~4,000 SNPs) to compute the RRM (**Supplementary Note 2**). The *P* values resulting from the complete and sampled sets were similar (**Fig. 2**). The different SNP sets led to nearly identical calls of significance, using the genome-wide significance threshold of 5×10^{-7}. When we used the complete set, the algorithm called 24 SNPs significant, and the 8,000-SNP and 4,000-SNP analyses labeled only one additional SNP significant and missed none. By comparison, the Armitage trend test (ATT) labeled seven additional SNPs significant and missed none. Furthermore, the λ statistic was similar for the complete, 8,000-SNP and 4,000-SNP analyses (1.132, 1.173 and 1.203, respectively) in contrast to $\lambda = 1.333$ for the ATT. We show corresponding quantile-quantile (Q-Q) plots in **Supplementary Figure 1**. Finally, using these SNP samples to construct genetic similarity, FaST-LMM ran an order of magnitude faster than EMMAX: 23 min and 53 min for the 4,000-SNP and 8,000-SNP FaST-LMM analyses compared with 260 min and 290 min for the respective EMMAX analyses.

With respect to selecting SNPs to estimate genetic similarity, an alternative to uniformly distributed sampling would be to choose SNPs with a strong association to phenotype. On the WTCCC data, we found that using the 200 most strongly associated SNPs according to ATT performed at least as well as the 8,000-SNP sample, making the same calls of significance as the analysis with the complete set and yielding a λ statistic of 1.135.

We envision several future directions. One is to apply FaST-LMM to multivariate analyses. Once the rotations have been applied to the SNPs, covariates and phenotype, then multivariate additive

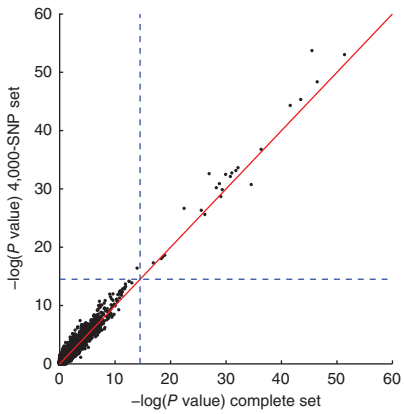

Figure 2 | Accuracy of association *P* values resulting from SNP sampling on WTCCC data for the Crohn's disease phenotype. Each point in the plot shows the negative log *P* values of association for a particular SNP from an LMM using a 4,000-SNP sample and all SNPs to compute the RRM. The complete set used all 340,000 SNPs from all but chromosome 1, whereas the 4,000-SNP sample used equally spaced SNPs from these chromosomes. All 28,000 SNPs in chromosome 1 were tested. Dashed lines show the genome-wide significance threshold (5×10^{-7}). The correlation for the points in the plot is 0.97.

analyses, including those using regularized estimation methods, can be achieved in time that is linear in the cohort size with no additional spectral decompositions or rotations. Also, the time complexity of FaST-LMM can be additionally reduced by using only the top eigenvectors of the spectral decomposition to rotate the data (those with the largest eigenvalues). On the WTCCC data, use of fewer than 200 eigenvectors yielded univariate *P* values comparable to those obtained from many thousands of eigenvectors. FaST-LMM can be made even more efficient when multiple individuals have the same genotype in common or when the LMM is compressed (as in compressed mixed linear models[4]) (**Supplementary Note 1**). Finally, the identification of associations between genetic markers and gene expression ('expression quantitative trait loci' analyses) can be thought of as multiple applications of GWAS[15], making our FaST-LMM approach applicable to such analyses.

FaST-LMM software is available as **Supplementary Software** and at http://mscompbio.codeplex.com/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank E. Renshaw for help with implementation of Brent's method and the χ^2 distribution function, J. Carlson for help with tools used to manage the data and deploy runs on our computer cluster, and N. Pfeifer for an implementation of the ATT. A full list of the investigators who contributed to the generation of the Wellcome Trust Case-Control Consortium data we used in this study is available from http://www.wtccc.org.uk/. Funding for the project was provided by the Wellcome Trust (076113 and 085475). The GAW14 data were provided by the members of the Collaborative Study on the Genetics of Alcoholism (US National Institutes of Health grant U10 AA008401).

AUTHOR CONTRIBUTIONS

C.L., J.L. and D.H. designed and performed research, contributed analytic tools, analyzed data and wrote the paper. Y.L. designed and performed research. C.M.K. and R.I.D contributed analytic tools.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Balding, D.J. *Nat. Rev. Genet.* **7**, 781–791 (2006).
2. Yu, J. *et al. Nat. Genet.* **38**, 203–208 (2006).
3. Kang, H.M. *et al. Genetics* **107**, 1709–1723 (2008).
4. Zhang, Z. *et al. Nat. Genet.* **42**, 355–360 (2010).
5. Kang, H.M. *et al. Nat. Genet.* **42**, 348–354 (2010).
6. Zhao, K. *et al. PLoS Genet.* **3**, e4 (2007).
7. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. *Nat. Rev. Genet.* **11**, 459–463 (2010).
8. Henderson, C.R. *Applications of Linear Models in Animal Breeding* (University of Guelph, Guelph, Ontario, Canada, 1984).
9. Goddard, M.E., Wray, N., Verbyla, K. & Visscher, P.M. *Stat. Sci.* **24**, 517–529 (2009).
10. Hayes, B.J., Visscher, P.M. & Goddard, M.E. *Genet. Res.* **91**, 47–60 (2009).
11. Fisher, R. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
12. Yang, J. *et al. Nat. Genet.* **42**, 565–569 (2010).
13. Welham, S. & Thompson, R. *J. R. Stat. Soc. B* **59**, 701–714 (1997).
14. Demidenko, E. *Mixed Models Theory and Applications* (Wiley, Hoboken, New Jersey, USA, 2004).
15. Listgarten, J., Kadie, C., Schadt, E.E. & Heckerman, D. *Proc. Natl. Acad. Sci. USA* **107**, 16465–16470 (2010).

ONLINE METHODS

Software. FaST-LMM is available as **Supplementary Software**, and updates to source code and software are available from http://mscompbio.codeplex.com/.

Experimental details. The calibration of *P* values was assessed using the λ statistic, also known as the inflation factor from the genomic control[1,16]. The value λ is defined as the ratio of the median observed to median theoretical test statistic. Values of λ substantially greater than (less than) 1.0 are indicative of inflation (deflation).

The GAW14 data[17] consisted of autosomal SNP data from an Affymetrix SNP panel and a phenotype indicating whether an individual smoked a pack of cigarettes a day or more for six months or more. In addition to the curation provided by GAW, we excluded a SNP when either (i) its minor allele frequency was less than 0.05, (ii) its values were missing in more than 5% of the population or (iii) its allele frequencies were not in Hardy-Weinberg equilibrium ($P < 0.0001$). In addition, we excluded an individual with more than 10% of SNP values missing. After filtering, there were 7,579 SNPs across 1,261 individuals. The data relected individuals of multiple races and many close family members: 1,034 individuals represented in the dataset also had parents, children or siblings represented in the dataset.

We used the GAW14 data as the basis for creating large synthetic datasets to evaluate run times and memory use. Datasets GAW14.*x*, with x = 1, 5, 10, 20, 50 and 100 were generated. Roughly, we constructed the synthetic GAW14.*x* dataset by 'copying' the original dataset x times. For each 'white', 'black' and Hispanic individual in the original dataset (1,238 individuals), we created x individuals in the copy. Similarly, we copied the family relationships among these individuals from the pedigree on the real data. For each individual with no parent represented in the dataset, we sampled data for each SNP using the race-based marginal frequency of that SNP in the original dataset. We determined the SNPs for the remaining individuals from the parental SNPs assuming a rate of 38 recombination events per genome. We then sampled a phenotype for each individual from a generalized linear mixed model (GLMM) with a logistic link function whose parameters were adjusted to mimic that of the real data. In particular, we adjusted the offset and genetic-variance parameters of the GLMM so that (i) the phenotype frequency in the real and synthetic data were almost the same, and (ii) the genetic variance parameter of an LMM fit to the real and synthetic data were comparable. We assumed that there were no fixed effects. Analysis of GAW14 and that of GAW14.1 had almost identical run times and memory footprints. The GAW14.x datasets are available at http://www.gaworkshop.org/about/dh_simulation_ms.html.

The WTCCC 1 data consisted of the SNP and phenotype data for seven common diseases: bipolar disorder, coronary artery disease, hypertension, Crohn's disease, rheumatoid arthritis, type-I diabetes and type-II diabetes[18]. Each phenotype group contained information for ~1,900 individuals. In addition, the data included ~1,500 controls from the UK National Blood Service Control Group (NBS). The data did not include a second control group from the 1958 British Birth Cohort (58C), as permissions for it precluded use by a commercial organization. Our analysis for a given disease phenotype used data from the NBS group and the remaining six phenotypes as controls. In our initial analysis, we excluded data for individuals and SNPs as previously described[18]. The difference between values of λ from an (uncorrected) analysis using ATT and the ATT values from the original analysis[18] averaged 0.02 across the phenotypes with an s.d. of 0.01, indicating that the absence of the 58C data in our analysis had little effect on inflation or deflation. In these initial analyses, we found a substantial over-representation of *P* values equal to one and traced this to the existence of thousands of nonvarying SNPs or single-nucleotide constants. In addition, we found that SNPs with very low minor-allele frequencies led to skewed *P*-value distributions. Consequently, we used a more conservative SNP filter, also described by the WTCCC[18], in which a SNP was excluded if either its minor-allele frequency was less than 1% or it was missing in greater than 1% of the individuals represented in the dataset. After filtering, 368,584 SNPs remained.

In the sampling and timing experiments, we included non-white individuals and close family members to increase the potential for confounding and thereby better exercise the LMM. In total, there were 14,925 individuals across the seven phenotypes and control. We used only the Crohn's disease phenotype because it was the only one that had appreciable apparent inflation according to ATT *P* values. We created the 8,000-SNP and 4,000-SNP sets used to estimate genetic similarity from all but chromosome 1 by including every forty-second and every eighty-fourth SNP, respectively, along each chromosome.

All analyses assumed a single additive effect of a SNP on the phenotype, using a 0-1-2 encoding for each SNP. The FaST-LMM runs used the RRM, whereas the EMMAX runs used the identity by state kinship matrix. Missing SNP data was mean imputed. A likelihood ratio test was used to compute *P* values for FaST-LMM. Run times were measured on a dual AMD six-core Opteron machine with a 2.6-gigahertz (GHz) clock and 32 GB of RAM. Only one core was used. FaST-LMM used the AMD Core Math library.

FaST-LMM. Here we highlight important points in the development of the maximum likelihood version of FaST-LMM. A complete description, including minor modifications needed for the REML version, is available in **Supplementary Note 1**.

The LMM log likelihood of the phenotype data, $\mathbf{y}$ (dimension $n \times 1$), given fixed effects $\mathbf{X}$ (dimension $n \times d$), which include the SNP to be tested, the covariates and the column of ones corresponding to the bias (offset), can be written as

$$LL\left(\sigma_e^2,\sigma_g^2,\boldsymbol{\beta}\right) = \log N\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta};\sigma_g^2\mathbf{K}+\sigma_e^2\mathbf{I}\right) \tag{1}$$

where $N(\mathbf{r}|\mathbf{m};\,\Sigma)$ denotes a normal distribution in variable $\mathbf{r}$ with mean $\mathbf{m}$ and covariance matrix Σ; $\mathbf{K}$ (dimension $n \times n$) is the genetic similarity matrix; $\mathbf{I}$ is the identity matrix; σ_e^2 (scalar) is the magnitude of the residual variance; σ_g^2 (scalar) is the magnitude of the genetic variance; and $\boldsymbol{\beta}$ (dimension $d \times 1$) are the fixed-effect weights.

To efficiently estimate the parameters $\boldsymbol{\beta}$, σ_g^2 and σ_e^2 and the log likelihood at those values, we can factor equation (1). In particular, we let δ be σ_e^2/σ_g^2 and $\mathbf{USU}^{\mathrm{T}}$ be the spectral decomposition of $\mathbf{K}$ (where $\mathbf{U}^{\mathrm{T}}$ denotes the transpose of $\mathbf{U}$), so that equation (1) becomes

$$LL\left(\delta,\sigma_g^2,\boldsymbol{\beta}\right) = -\frac{1}{2}\left(n\log\left(2\pi\sigma_g^2\right)+\log\left(\left|\mathbf{U}(\mathbf{S}+\delta\mathbf{I})\mathbf{U}^{\mathrm{T}}\right|\right)\right.$$
$$\left.+\frac{1}{\sigma_g^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\left(\mathbf{U}(\mathbf{S}+\delta\mathbf{I})\mathbf{U}^{\mathrm{T}}\right)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right),$$

doi:10.1038/nmeth.1681

where $|\mathbf{K}|$ denotes the determinant of matrix $\mathbf{K}$. The determinant of the genetic similarity matrix, $|\mathbf{U}(\mathbf{S}+\delta\mathbf{I})\mathbf{U}^{\mathrm{T}}|$ can be written as $|\mathbf{S}+\delta\mathbf{I}|$. The inverse of the genetic similarity matrix can be rewritten as $\mathbf{U}(\mathbf{S}+\delta\mathbf{I})^{-1}\mathbf{U}^{\mathrm{T}}$. Thus, after additionally moving out $\mathbf{U}$ from the covariance term so that it now acts as a rotation matrix on the inputs ($\mathbf{X}$) and targets ($\mathbf{y}$), we obtain

$$LL\left(\delta,\sigma_g^2,\boldsymbol{\beta}\right) = -\frac{1}{2}\left(n\log\left(2\pi\sigma_g^2\right)+\log\left(|\mathbf{S}+\delta\mathbf{I}|\right)\right.$$
$$\left.+\frac{1}{\sigma_g^2}\left(\left(\mathbf{U}^{\mathrm{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathrm{T}}\mathbf{X}\right)\boldsymbol{\beta}\right)^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\left(\left(\mathbf{U}^{\mathrm{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathrm{T}}\mathbf{X}\right)\boldsymbol{\beta}\right)\right).$$

The 'Fa' in FaST-LMM stands for this factorization. As the covariance matrix of the normal distribution is now a diagonal matrix $\mathbf{S}+\delta\mathbf{I}$, the log likelihood can be rewritten as the sum over n terms, yielding

$$LL\left(\delta,\sigma_g^2,\boldsymbol{\beta}\right)=-\frac{1}{2}\left(n\log\left(2\pi\sigma_g^2\right)+\sum_{i=1}^{n}\log\left([\mathbf{S}]_{ii}+\delta\right)\right.$$
$$\left.+\frac{1}{\sigma_g^2}\sum_{i=1}^{n}\frac{\left(\left[\mathbf{U}^T\mathbf{y}\right]_i-\left[\mathbf{U}^T\mathbf{X}\right]_{i:}\boldsymbol{\beta}\right)^2}{[\mathbf{S}]_{ii}+\delta}\right) \quad (2)$$

where $[\mathbf{U}^T\mathbf{X}]_{i:}$ denotes the i^{th} row of $\mathbf{X}$. Note that this expression is equal to the product of n univariate normal distributions on the rotated data, yielding the linear regression equation

$$LL\left(\delta,\sigma_g^2,\boldsymbol{\beta}\right)=\log\prod_{i=1}^{n}N\left(\left[\mathbf{U}^{\mathrm{T}}\mathbf{y}\right]_i\,|\left[\mathbf{U}^{\mathrm{T}}\mathbf{X}\right]_{i:}\boldsymbol{\beta};\sigma_g^2\left([\mathbf{S}]_{ii}+\delta\right)\right).$$

To determine the values of δ, σ_g^2, and $\boldsymbol{\beta}$ that maximize the log likelihood, we first differentiate equation (2) with respect to $\boldsymbol{\beta}$, set it to zero and analytically solve for the maximum likelihood (ML) value of $\boldsymbol{\beta}(\delta)$. We then substitute this expression in equation (2), differentiate the resulting expression with respect to σ_g^2, set it to zero and solve analytically for the ML value of $\sigma_g^2(\delta)$. Next, we plug in the ML values of $\sigma_g^2(\delta)$ and $\boldsymbol{\beta}(\delta)$ into equation (2) so that it is a function only of δ. Finally, we optimize this function of δ using a one-dimensional numerical optimizer based on Brent's method (**Supplementary Note 1**).

Note that, given δ and the spectral decomposition of $\mathbf{K}$, each evaluation of the likelihood has a run time that is linear in n. Consequently, when testing s SNPs, the time complexity is $O(n^3)$ for finding all eigenvalues ($\mathbf{S}$) and eigenvectors ($\mathbf{U}$) of $\mathbf{K}$, $O(n^2s)$ for rotating the phenotype vector $\mathbf{y}$, and all of the SNP and covariate data (that is, computing $\mathbf{U}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{U}^{\mathrm{T}}\mathbf{X}$), and $O(Cns)$ for performing C evaluations of the log likelihood during the one-dimensional optimization over δ. Therefore, the total time complexity of FaST-LMM, given $\mathbf{K}$, is $O(n^3+n^2s+Cns)$. By keeping δ fixed to its value from the null model (analogously to EMMAX/P3D), this complexity reduces to $O(n^3+n^2s+Cn)$. The size of both $\mathbf{K}$ and $\mathbf{U}$ is $O(n^2)$, which dominates the space complexity, as each SNP can be processed independently so that there is no need to load all SNP data into memory at once. In most applications, the number of fixed effects per test, d, is a single-digit integer and is omitted in these expressions because its contribution is negligible.

Next we consider the case where $\mathbf{K}$ is of low rank, that is, k, the rank of $\mathbf{K}$ is less than n, the number of individuals. This case will occur when the RRM is used and the number of (linearly independent) SNPs used to estimate it, $s_c = k$, is smaller than n. $\mathbf{K}$ can be of low rank for other reasons: for example, by forcing some eigenvalues to zero (**Supplementary Note 1**).

In the complete spectral decomposition of $\mathbf{K}$ given by $\mathbf{USU}^{\mathrm{T}}$, we let $\mathbf{S}$ be an $n \times n$ diagonal matrix containing the k nonzero eigenvalues on the top left of the diagonal, followed by $n-k$ zeros on the bottom right. In addition, we write the $n \times n$ orthonormal matrix $\mathbf{U}$ as $[\mathbf{U}_1, \mathbf{U}_2]$, where $\mathbf{U}_1$ (of dimension $n \times k$) contains the eigenvectors corresponding to nonzero eigenvalues, and $\mathbf{U}_2$ (of dimension $n \times n-k$)) contains the eigenvectors corresponding to zero eigenvalues. Thus, $\mathbf{K}$ is given by $\mathbf{USU}^{\mathrm{T}} = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathrm{T}} + \mathbf{U}_2\mathbf{S}_2\mathbf{U}_2^{\mathrm{T}}$. Furthermore, as $\mathbf{S}_2$ is $[\mathbf{0}]$, $\mathbf{K}$ becomes $\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathrm{T}}$, the k-spectral decomposition of $\mathbf{K}$, so-called because it contains only k eigenvectors and arises from taking the spectral decomposition of a matrix of rank k. The expression $\mathbf{K}+\delta\mathbf{I}$ appearing in the LMM likelihood, however, is always of full rank (because $\delta > 0$):

$$\mathbf{K}+\delta\mathbf{I} = \mathbf{U}(\mathbf{S}+\delta\mathbf{I})\mathbf{U}^{\mathrm{T}} = \mathbf{U}\begin{bmatrix}\mathbf{S}_1+\delta\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \delta\mathbf{I}\end{bmatrix}\mathbf{U}^{\mathrm{T}}.$$

Therefore, it is not possible to ignore $\mathbf{U}_2$ as it enters the expression for the log likelihood. Furthermore, directly computing the complete spectral decomposition does not exploit the low rank of $\mathbf{K}$. Consequently, we use an algebraic trick involving the identity $\mathbf{U}_2\,\mathbf{U}_2^{\mathrm{T}} = \mathbf{I} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}$ to rewrite the likelihood in terms not involving $\mathbf{U}_2$ (equation 3.4 in **Supplementary Note 1**). As a result, we incur only the time and space complexity of computing $\mathbf{U}_1$ rather than $\mathbf{U}$.

Given the k-spectral decomposition of $\mathbf{K}$, the maximum likelihood of the model can be evaluated with time complexity $O(nsk)$ for the required rotations and $O(C(n+k)s)$ for the C evaluations of the log likelihood during the one-dimensional optimizations over δ. By keeping δ fixed to its value from the null model, as in EMMAX/P3D, $O(C(n+k)s)$ is reduced to $O(C(n+k))$. In general, the k-spectral decomposition can be computed by first constructing the genetic similarity matrix from k SNPs at a time complexity of $O(n^2s_c)$ and space complexity of $O(n^2)$, and then finding its first k eigenvalues and eigenvectors at a time complexity of $O(n^2k)$. When the RRM is used, however, the k-spectral decomposition can be performed more efficiently by circumventing the construction of $\mathbf{K}$ because the singular vectors of the data matrix are the same as the eigenvectors of the RRM constructed from those data (**Supplementary Note 1**). In particular, the k-spectral decomposition of $\mathbf{K}$ can be obtained from the singular value decomposition of the $n \times s_c$ SNP matrix directly, which is an $O(ns_ck)$ operation. Therefore, the total time complexity of low-rank FaST-LMM using δ from the null model is $O(ns_ck + nsk + C(n+k))$. Assuming SNPs to be tested are loaded into memory in small blocks, the total space complexity is $O(ns_c)$.

Finally, we note that for both the full and low-rank versions of FaST-LMM, the rotations (and, if performed, the search for δ for each test) are easily parallelized. Consequently, the run time of the LMM analysis is dominated by the spectral decomposition (or singular value decomposition for the low-rank version). Although parallel algorithms for singular-value decomposition exist, improvements to such algorithms should lead to even greater speedup.

16. Devlin, B. & Roeder, K. *Biometrics* **55**, 997–1004 (1999).
17. Edenberg, H.J. *et al. BMC Genet.* **6** (suppl. 1), S2 (2005).
18. Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).