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], and David Heckerman[1,3]

[1]Microsoft Research
Los Angeles, CA

[2]Max Plank 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), and
D.H. (heckerma@microsoft.com).

Abstract

We describe *Fa*ctored *S*pectrally *T*ransformed *L*inear *M*ixed *M*odels (FaST-LMM), an algorithm for genome-wide association studies that scales linearly in the number of individuals in both runtime and memory use. On Wellcome Trust data with 15,000 individuals, FaST-LMM runs an order of magnitude faster than current efficient algorithms. Our algorithm can analyze data for 120,000 individuals in just a few hours, whereas the current algorithms fail at 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 the other methods, LMMs have been shown capable of capturing 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 runtime and memory footprint required by these models scale as the cube and square of the number of individuals 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 share causative alleles. Such measures include those based on IBD[10,11] and the realized relationship matrix (RRM)[9,10,12], and have been estimated with a small sample of markers (200-2000 in number)[2,4]. Herein, we take advantage of such sampling to make LMM analysis applicable to extremely large datasets. In particular, we introduce a reformulation of LMMs, called FaST-LMM for *Fa*ctored *S*pectrally *T*ransformed *L*inear *M*ixed Models. We show that, provided (1) the number of SNPs used to estimate genetic similarity between pairs of individuals is less than the number of individuals in the dataset (regardless of how many SNPs are to be tested) and (2) the RRM is used to determine these similarities, then FaST-LMM produces exactly the same results as a standard LMM, but with a runtime and memory footprint that is only linear in the number of individuals. 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 (ML)—or alternatively, the restricted maximum likelihood (REML)—of a 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 ML (or REML) parameters becomes an optimization problem over δ only. The algorithm *E*fficient *M*ixed *M*odel *A*ssociation (EMMA)[3] speeds up the evaluation of the log likelihood for any value of δ, which is ordinarily cubic in the number of individuals, by clever use of spectral decompositions. However, the approach requires a new spectral decomposition for each SNP tested (a cubic operation). The algorithms *E*fficient *M*ixed *M*odel *A*ssociation e*X*pedited (EMMAX) and *P*opulation *P*arameters *P*reviously *D*etermined (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 this rotated data becomes uncorrelated and hence amenable to analysis with a linear regression model, which has a runtime and memory footprint linear in the number of individuals.

In general, the size (the number of entries) of the required rotation matrix is quadratic in the number of individuals, and computing this matrix by way of a spectral decomposition has cubic runtime in the number of individuals. When the number of SNPs used to construct the genetic similarity matrix is less than the number of individuals, however, the size of the matrix required to perform the rotations is linear in the number of individuals (and linear in this number of SNPs), and the time required to compute it is linear in the number of individuals (and quadratic in this number of SNPs). Intuitively, these savings can be achieved because the intrinsic dimensionality of the space spanned by the individuals and SNPs used to construct the similarity matrix can never be higher than the smaller of these two values. Thus, we can always choose to perform operations in the smaller space without any loss of information, while the computations remain exact. This basic idea has been exploited previously[8,14], but when applied to GWAS, would require expensive computations per SNP, making these approaches far less efficient than FaST-LMM.

To achieve our linear runtime 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 given in Methods.

We compared memory footprint and runtime for non-parallelized implementations of the FaST-LMM and EMMAX/P3D algorithms (**Fig. 1**). For the latter, we used the EMMAX implementation, which was no less efficient than P3D (in TASSEL) in terms of runtime and memory use. In the comparison, we used GAW14 data to construct synthetic datasets having roughly 1, 5, 10, 20, 50 and 100 times as many individuals and always the same number of SNPs (approximately 8K) as the original data. The largest such dataset contained 123,800 individuals. We tested all SNPs and used them all to estimate genetic similarity. EMMAX would not run on the 20x, 50x, or 100x 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. Runtime results highlight the linear dependence of the computations on the number of individuals when the numbers of individuals exceeds the 8K SNPs used to construct the RRM. Furthermore, computations remain practical within our approach even when the variance parameters are re-estimated 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 having 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 WTCCC data for the CD phenotype. Specifically, we tested all SNPs on chromosome 1 while using SNP sets of various sizes from all but this chromosome—the complete set (340K) and uniformly distributed samples of size 8K and 4K—to compute the RRM (**Supplementary Note 2**). The P values resulting from the complete and sampled sets were similar (**Fig. 2**). More important, the two algorithms made nearly identical calls of significance, using the genome-wide significance threshold of 5×10^{-7}. Namely, 24 SNPs were called significant when the complete set was used, whereas the 8K and 4K 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, 8K, and 4K analyses—1.132, 1.173, 1.203, respectively—in contrast to λ = 1.333 for the ATT. Corresponding Q-Q plots are shown in **Supplementary Fig. 1**. Finally, using these SNP samples to construct genetic similarity, FaST-LMM ran an order of magnitude faster than

EMMAX: 23 and 53 minutes for the 4K and 8K FaST-LMM analyses, and 260 and 290 minutes 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 in strong association with the phenotype. On the WTCCC data, we found that using the 200 most strongly associated SNPs according to ATT outperformed the 8K sample (λ = 1.135).

There are 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 analyses, including those using regularized estimation methods, can be achieved in time linear in the number of individuals with no additional spectral decompositions or rotations. In addition, the time complexity of FaST-LMM can be further 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. Furthermore, the ideas behind FaST-LMM and Compressed Mixed Linear Models[4] can be combined (**Supplementary Note 1**). Finally, the identification of associations between genetic markers and gene expression—eQTL analyses—can be thought of as multiple applications of GWAS[15], making our FaST-LMM approach applicable to such analyses.

Software updates for FaST-LMM, including source code and executables, are available from http://mscompbio.codeplex.com.

Acknowledgments

We thank E. Renshaw for help with implementation of Brent's method and the χ^2 distribution, 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. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. The GAW14 data were provided by the Collaborative Study on the Genetics of Alcoholism (U10 AA008401).

Author Contributions

C.L., J.L., and D.H contributed equally to this work. They designed research, 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

C.L., J.L., Y.L., C.K., R.D., and D.H. performed research related to this manuscript while employed by Microsoft.

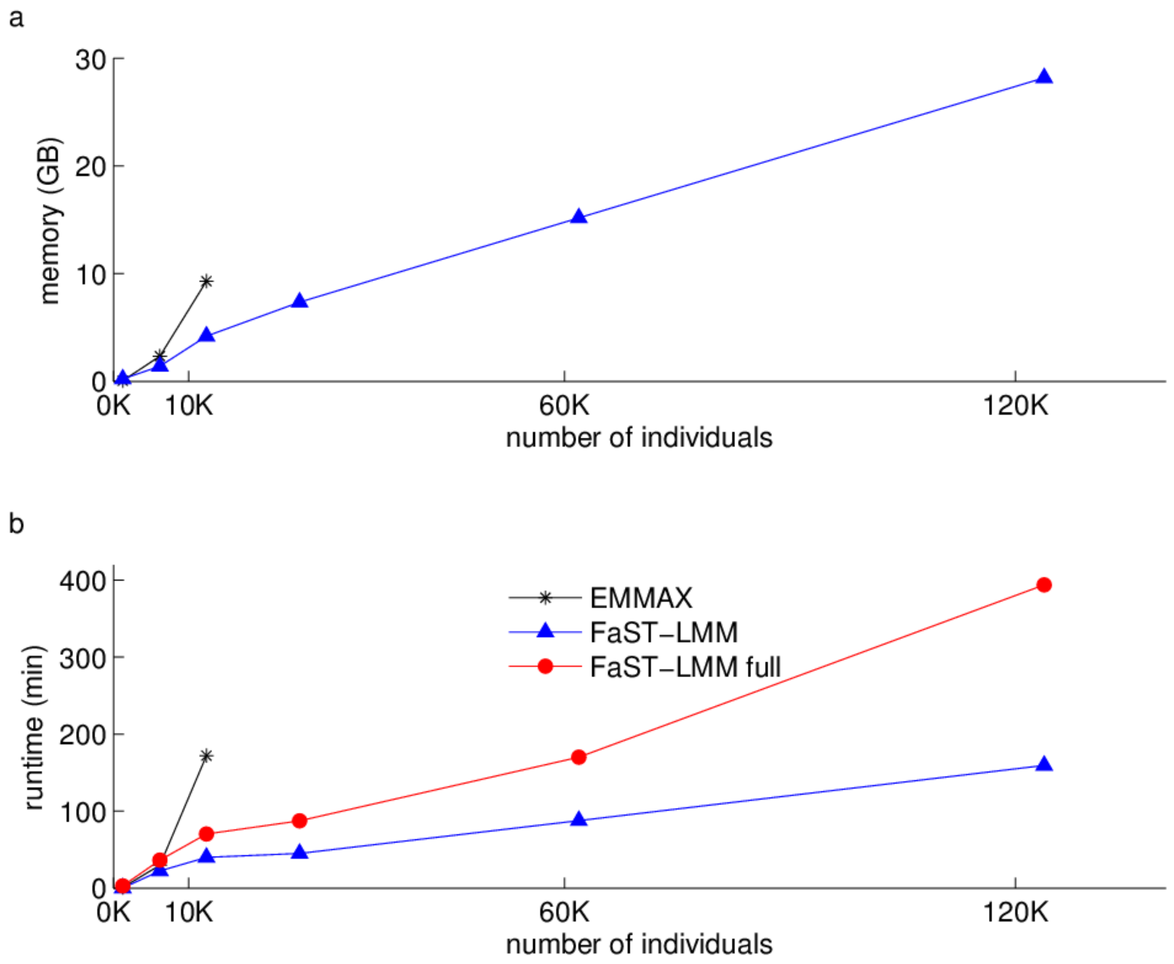

Figure 1. Memory footprint (**a**) and runtime (**b**) of FaST-LMM running on a single processor as a function of the number of individuals in synthetic datasets based on GAW14. In each run, we used 7,579 SNPs both to estimate genetic similarity (RRM for FaST-LMM and IBS for EMMAX) and to test for association. FaST-LMM full refers to an analysis where the variance parameters were re-estimated for each test, whereas FaST-LMM refers to estimating these parameters only once for the null model, as in EMMAX/P3D. 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 number of individuals in the GAW14 data, because the memory required to store the large matrices exceeded the 32 GB available.

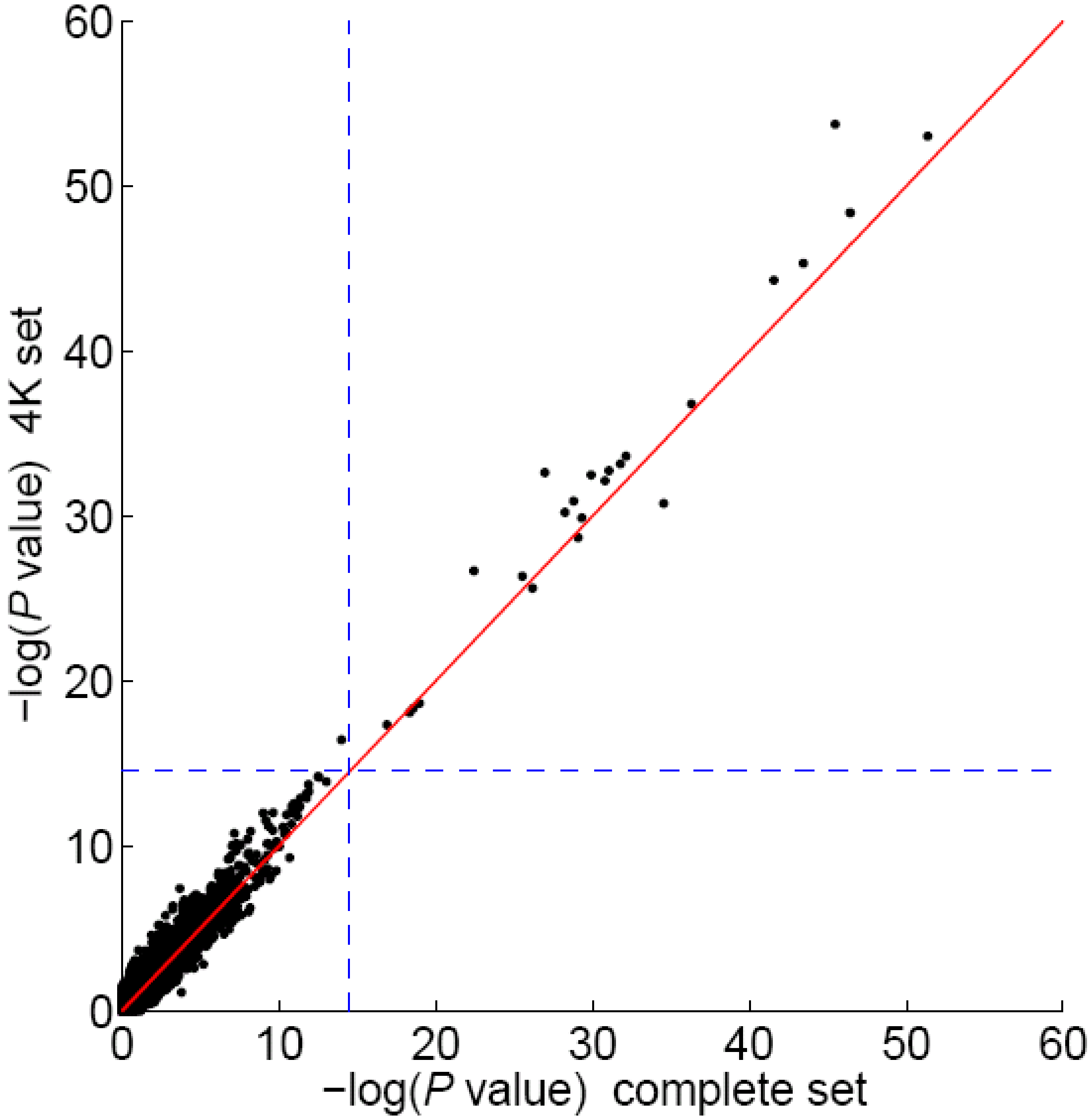

Figure 2. Accuracy of association *P* values resulting from SNP sampling on WTCCC data for the CD phenotype. Each point in the plot shows the negative log *P* values of association for a particular SNP from a LMM using a 4K SNP sample (*y*-axis) and all SNPs (*x*-axis) to compute the RRM. The complete set used all 340K SNPs from all but chromosome 1, whereas the 4K sample used equally spaced SNPs from these chromosomes. All 28K 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. A corresponding plot for an 8K sample looks essentially the same (correlation 0.98).

References

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** (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, 1984).
9. Goddard, M. E., Wray, N., Verbyla, K. & Visscher, P. M. *Statis. Sci* **24**, 517–529 (2009).
10. Hayes, B. J., Visscher, P. M. & Goddard, M. E. *Genet Res* (Camb) **91**, 47–60 (2009).
11. Fisher, R. *Trans. Roy. 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, 2004).
15. Listgarten, J., Kadie, C., Schadt, E. E. & Heckerman, D. *Proc. Natl. Acad. Sci*. U.S.A. **107**, 16465–16470 (2010).

Online Methods

Experimental Details

The calibration of P values was assessed using the λ statistic, also known as the inflation factor from genomic control[1,18]. 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 Genetic Analysis Workshop (GAW) 14 data[16] 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 (1) its minor allele frequency was less than 0.05, (2) its values were missing in more than 5% of the population, or its allele frequencies were not in Hardy-Weinberg equilibrium ($P < 0.0001$). In addition, we excluded an individual with more than 10% SNP values missing. After filtering, there were 7,579 SNPs across 1,261 individuals. The data consisted of multiple races and numerous close family members—1,034 individuals in the dataset had parents, children, or siblings also in the dataset.

We used the GAW14 data as the basis for creating large synthetic datasets to evaluate runtimes 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 data (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 parents, 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 (1) the phenotype frequency in the real and synthetic data were almost the same, and (2) the genetic variance parameter of a 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 runtimes and memory footprints.

The Wellcome Trust Case Control Consortium (WTCCC) 1 data consisted of the SNP and phenotype data for seven common diseases: bipolar disorder (BP), coronary artery disease (CAD), hypertension (HT), Chron's disease (CD), rheumatoid arthritis (RA), type-I diabetes (T1D), and type-II diabetes (T2D)[17]. Each phenotype group contained about 1,900 individuals. In addition, the data included a set of approximately 1,500 controls from the UK 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 individuals and SNPs as previously described[17]. The difference between values of λ from an (uncorrected) analysis using ATT, and the ATT values from the original analysis[17] averaged 0.02 across the phenotypes with a standard deviation 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 non-varying SNPs or single-nucleotide constants (SNCs). In addition, we found (not surprisingly) that SNPs with very low minor-allele frequencies led to skewed *P* value distributions. Consequently, we employed a more conservative SNP filter, also described by the WTCCC[17], wherein a SNP was excluded if either its minor-allele frequency was less than 1%, or it was missing in greater than 1% of individuals. 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 CD phenotype, because it was the only one that had appreciable apparent inflation according to ATT *P* values. We created the 8K and 4K 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 IBS kinship matrix. Missing SNP data was mean imputed. A likelihood ratio test was used to compute *P* values for FaST-LMM. Runtimes were measured on a dual AMD six-core Opteron machine with a 2.6GHz clock and 32 GB of RAM. Only one core was used. FaST-LMM used the AMD Core Math Library.

FaST-LMM

In this section, 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 given 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, the covariates, and the column of ones corresponding to the bias (offset), can be written as

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

where $N(\boldsymbol{r}|\boldsymbol{m}; \Sigma)$ denotes a Normal distribution in $\boldsymbol{r}$ with mean $\boldsymbol{m}$ and covariance matrix Σ; $\mathbf{K}$ (dimension $n \times n$) is the genetic similarity matrix; $\mathbf{I}$ is the identity matrix; e (scalar) is the magnitude of the residual variance; $\sigma_g{}^2$ (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 $\boldsymbol{\beta}$, $\sigma_g{}^2$ and $\sigma_e{}^2$ and the log likelihood at those values, we can factor Equation 1. In particular, we let δ be $\sigma_e{}^2/\sigma_g{}^2$ 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

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

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

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

The "Fa" in FaST-LMM gets its name from 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

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

where $[\mathbf{X}]_{\mathrm{I:}}$ 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

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

To determine the values of δ, $\sigma_g{}^2$, 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 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 runtime 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. For a more general exposition, wherein **K** is of low rank for other reasons—for example, by forcing some eigenvalues to zero—see **Supplementary Note 1**.

In the complete spectral decomposition of **K** given by $\mathbf{USU}^T$, we let **S** be an $n \times n$ diagonal matrix containing the k non-zero 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 **U** as $[\mathbf{U}_1, \mathbf{U}_2]$, where $\mathbf{U}_1$ (of dimension $n \times k$) contains the eigenvectors corresponding to non-zero eigenvalues, and $\mathbf{U}_2$ (of dimension $n \times n - k$) contains the eigenvectors corresponding to zero eigenvalues. Thus, **K** is given by $\mathbf{USU}^T = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^T + \mathbf{U}_2\mathbf{S}_2\mathbf{U}_2^T$. Furthermore, as $\mathbf{S}_2$ is $[\mathbf{0}]$, **K** becomes $\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^T$, the k-spectral decomposition of **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 **K**. Consequently, we use an algebraic trick involving the identity $\mathbf{U}_2\,\mathbf{U}_2^T = \mathbf{I} - \mathbf{U}_1\,\mathbf{U}_1^T$ to rewrite the likelihood in terms not involving $\mathbf{U}_2$ (see 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 **U**.

Given the k-spectral decomposition of **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 $O(n^2 s_c)$ and space complexity of $O(n^2)$, and then finding its first k eigenvalues and eigenvectors at a time complexity of $O(n^2 k)$. When the RRM is used, however, the k-spectral decomposition can be performed more efficiently by circumventing the construction of **K**, because the singular vectors of the data matrix are the same as the eigenvectors of the RRM constructed from that data (**Supplementary Note 1**). In particular, the k-spectral decomposition of **K** can be obtained from the singular value decomposition of the $n \times s_c$ SNP matrix directly, which is an $O(ns_c k)$ operation. Therefore, the total time complexity of low-rank FaST-LMM using δ from the null model is $O(ns_c k + nsk + C(n + k))$. 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 runtime 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 speedups.

References

16. Edenberg, H. J. *et al. BMC Genet*. **6 Suppl 1**, S2 (2005).
17. Burton, P. R. *et al. Nature* **447**, 661–678 (2007).
18. Devlin, B. & Roeder, K. *Biometrics* **55**, 997–1004 (1999).
19. Wall, M. E., Rechtsteiner, A. & Rocha, L. M. A Practical Approach to Microarray Data Analysis (Kluwer, Norwell, MA, 2003).