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 *Fac*tored *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 ($\lambda = 1.135$).

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.

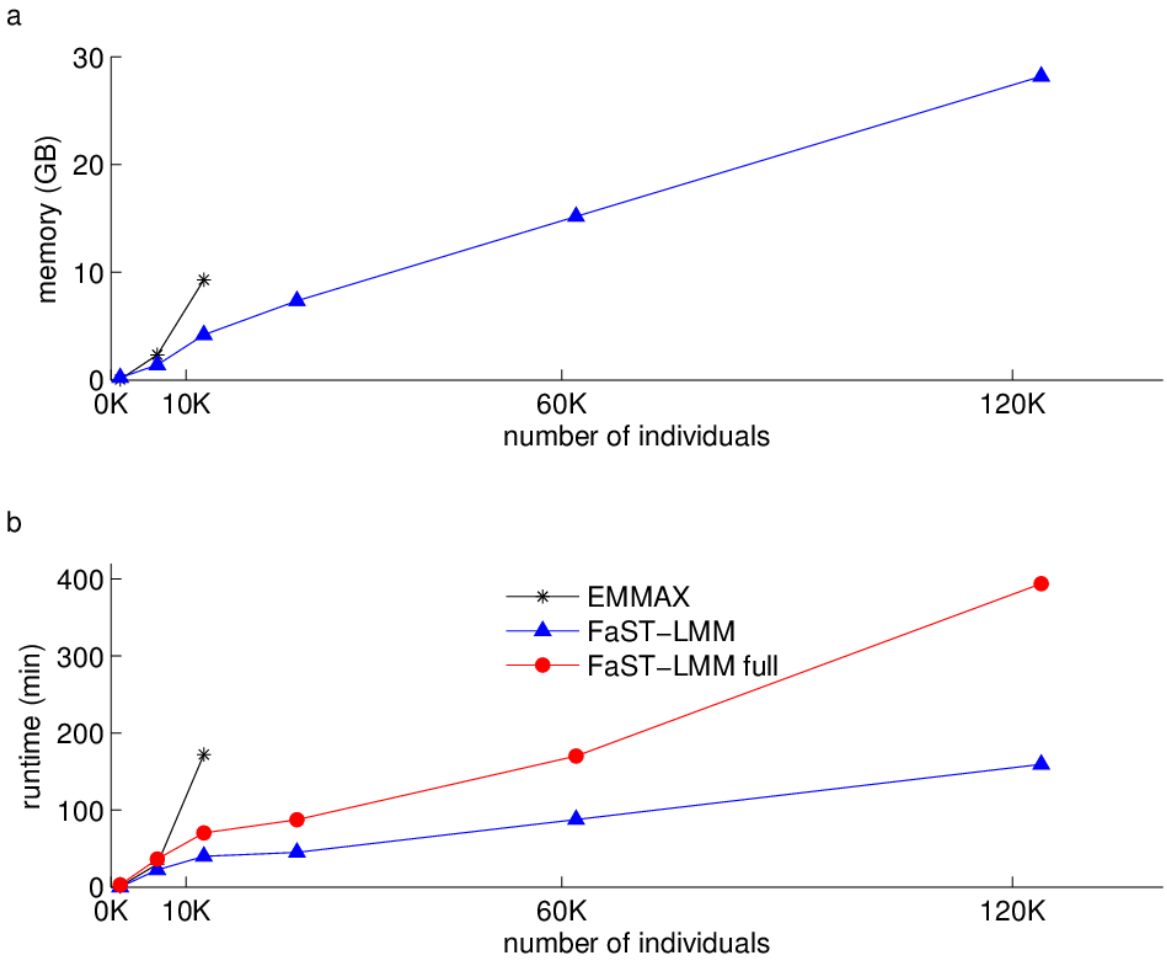


Figure 1. Memory footprint (**a**) and runtime (**b**) of FaST-LMM running on a single processor as a function of the number of individuals in synthetic datasets based on GAW14. In each run, we used 7,579 SNPs both to estimate genetic similarity (RRM for FaST-LMM and IBS for EMMAX) and to test for association. 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.

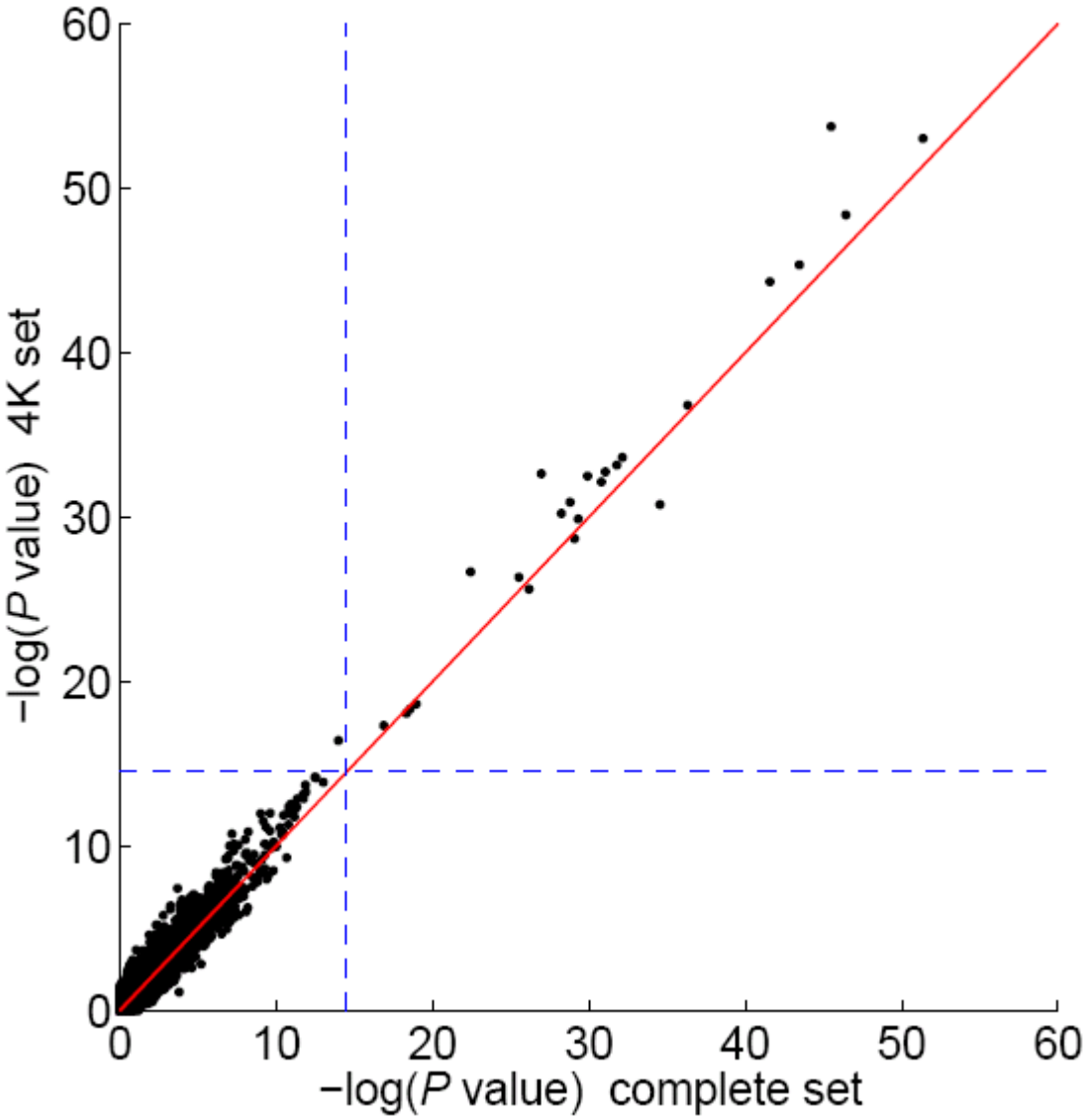


Figure 2. Accuracy of association *P* values resulting from SNP sampling on WTCCC data for the CD phenotype. Each point in the plot shows the negative log *P* values of association for a particular SNP from a LMM using a 4K SNP sample (*y*-axis) and all SNPs (*x*-axis) to compute the RRM. The complete 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

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

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; 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

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

where |**K**| denotes the determinant of matrix **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 **U** from the covariance term so that it now acts as a rotation matrix on the inputs (**X**) and targets (**y**), we obtain

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

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

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

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

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

To determine the values of δ, 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 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 **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 (**S**) and eigenvectors (**U**) of **K**, $O(n^2s)$ for rotating the phenotype vector **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 **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 **K** and **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 **K** is of low rank—that is, k, the rank of **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 $\delta>0$):

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

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).

Nature Methods

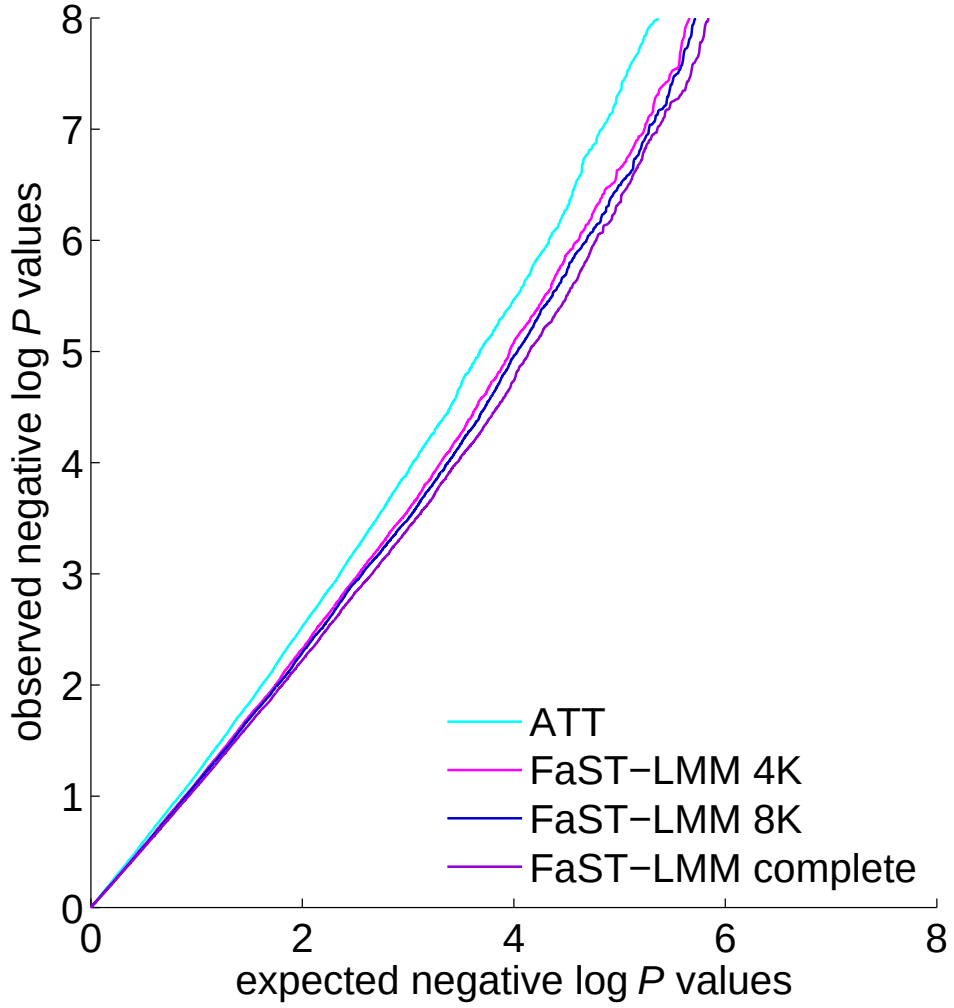
FaST linear mixed models for genome-wide association studies

Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson & David Heckerman

Supplementary Figure 1	Q-Q plots for the WTCCC data.
Supplementary Note 1	The FaST-LMM algorithm.
Supplementary Note 2	Null-model contamination.

Nature Methods:doi10.1038/nmeth.1681

Supplementary Figure 1



Q-Q plots for the WTCCC data. Shown are observed versus expected negative log P values for the association analyses on the CD phenotype described in the main paper. We used FaST-LMM to test all SNPs on chromosome 1, and SNP sets of various sizes from all but this chromosome—the complete set (340K), 8K, and 4K—to compute the RRM. We also used ATT to compute P values.

Nature Methods:doi10.1038/nmeth.1681

Supplementary Note 1: The FaST-LMM Algorithm

Here we describe our approach called FaST-LMM, which stands for *Fa*ctored *S*pectrally *T*ransformed *L*inear *M*ixed *M*odels. We derive formulas that allow for efficient evaluation of the likelihood as well as the maximum likelihood (ML) and restricted maximum likelihood (REML) parameters. We consider the cases where the genetic similarity matrix has full and low rank separately. The following notation will be used.

- n denotes the cohort size (the number of individuals represented in the data set).
- s denotes the total number of SNPs to be tested.
- d denotes the number of fixed effects in a single model, including the offset, the covariates, and in the case of an alternative model, the SNP to be tested. Although we use only one SNP at a time in our work, all equations follow regardless of the number of SNPs fixed effects.
- k denotes the rank of the genetic similarity matrix.
- s_c denotes the number of SNPs used to construct the genetic similarity matrix.
- $\mathbf{X} \in \mathbb{R}^{n \times d}$ denotes the matrix of fixed effects. This matrix includes the column of 1s corresponding to the offset, the covariates, and the SNP to be tested.
- $\mathbf{y} \in \mathbb{R}^{n \times 1}$ denotes the vector of phenotype measurements.
- $\mathbf{K} \in \mathbb{R}^{n \times n}$ denotes the symmetric positive (semi)-definite genetic similarity matrix.
- $\mathbf{I}_a$ denotes the identity matrix of dimension a. If no subscript a is given, the dimensionality is implied by the context.
- σ_g^2 denotes the magnitude of the genetic variance.
- σ_e^2 denotes the magnitude of the residual variance.
- $\delta \equiv \frac{\sigma_e^2}{\sigma_g^2}$ denotes the fraction of genetic variance and residual variance.
- $\boldsymbol{\beta} \in \mathbb{R}^{d \times 1}$ denotes the vector of fixed effect weights corresponding to $\mathbf{X} \in \mathbb{R}^{n \times d}$.
- $\mathbf{S} \in \mathbb{R}^{n \times n}$ denotes the diagonal matrix containing the eigenvalues of $\mathbf{K}$ ordered by their magnitude from large to small as diagonal elements.
- $\mathbf{U} \in \mathbb{R}^{n \times n}$ denotes to the matrix of eigenvectors of $\mathbf{K}$, in the order of the corresponding eigenvalues.
- $\mathbf{U}\mathbf{S}\mathbf{U}^\mathsf{T} = \mathbf{K}$ is the spectral decomposition of $\mathbf{K}$.
- $|\mathbf{A}|$ denotes the determinant of matrix $\mathbf{A}$.
- $\mathbf{A}^\mathsf{T}$ denotes the transpose of matrix $\mathbf{A}$.

Nature Methods:doi10.1038/nmeth.1681

- $\mathbf{A}^{-1}$ denotes the inverse of matrix $\mathbf{A}$.
- $\mathbf{A}^{-\mathrm{T}}$ denotes the transposed inverse (or inverse of the transpose) of matrix $\mathbf{A}$.
- $[\mathbf{A}]_{ij}$ denotes the element of matrix $\mathbf{A}$ in the i^{th} row and j^{th} column.
- $[\mathbf{A}]_{i:}$ denotes the i^{th} row of matrix $\mathbf{A}$.
- $[\mathbf{a}]_i$ denotes the i^{th} entry of vector $\mathbf{a}$.
- $\mathbf{0}$ denotes a matrix where every entry is zero.
- $[\mathbf{A}, \mathbf{B}]$ denotes the concatenation of matrices $\mathbf{A}$ and $\mathbf{B}$.

1 LMMs with a full rank genetic similarity

We first consider the case where the genetic similarity matrix is of full rank (*i.e.*, the rank is equal to the cohort size).

1.1 Linear-time evaluation of the log likelihood

The log likelihood is parameterized by a weight vector $\boldsymbol{\beta}$ and the variances of the random components, σ_e^2 and σ_g^2:

$$LL\left(\sigma_e^2, \sigma_g^2, \boldsymbol{\beta}\right) = \log \mathcal{N}\left(\mathbf{y} | \mathbf{X}\boldsymbol{\beta}; \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right). \tag{1.1}$$

Introducing $\delta \equiv \frac{\sigma_e^2}{\sigma_g^2}$, the covariance matrix becomes $\sigma_g^2 (\mathbf{K} + \delta \mathbf{I})$, and the likelihood becomes a function of $\boldsymbol{\beta}$, δ and σ_g^2 [1]:

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

Using the formula for the n-variate Normal distribution, we obtain

$$LL\left(\delta, \sigma_g^2, \boldsymbol{\beta}\right) = -\frac{1}{2}\left(n \log\left(2\pi\sigma_g^2\right) + \log\left(|(\mathbf{K} + \delta \mathbf{I})|\right) + \frac{1}{\sigma_g^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{K} + \delta \mathbf{I})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right). \tag{1.2}$$

Letting $\mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}} = \mathbf{K}$ be the spectral decomposition of $\mathbf{K}$, and noting that $\mathbf{I} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$, Equation 1.2 becomes

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

Next, we factor out $\mathbf{U}$ and $\mathbf{U}^{\mathrm{T}}$ from the covariance of the Normal, so that it becomes the diagonal matrix $(\mathbf{S} + \delta \mathbf{I})$, obtaining

$$-\frac{1}{2}\left(n \log\left(2\pi\sigma_g^2\right) + \log\left(\left|\mathbf{U}(\mathbf{S} + \delta \mathbf{I})\mathbf{U}^{\mathrm{T}}\right|\right) + \frac{1}{\sigma_g^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \left(\mathbf{U}(\mathbf{S} + \delta \mathbf{I})\mathbf{U}^{\mathrm{T}}\right)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right). \tag{1.3}$$

The determinant of the genetic similarity matrix, $|\mathbf{U}(\mathbf{S} + \delta \mathbf{I})\mathbf{U}^{\mathrm{T}}|$ can be written as $|(\mathbf{S} + \delta \mathbf{I})|$ using the properties that $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$, and that $|\mathbf{U}| = |\mathbf{U}^{\mathrm{T}}| = 1$. The inverse of the genetic similarity

matrix can be rewritten as $\mathbf{U}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\mathbf{U}^{\mathsf{T}}$ using the properties that $(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$, that $\mathbf{U}^{-1}=\mathbf{U}^{\mathsf{T}}$, and that $\mathbf{U}^{-\mathsf{T}}=\mathbf{U}$. 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

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

$$=-\frac{1}{2}\left(n\log\left(2\pi\sigma_g^2\right)+\log\left(\left|\left(\mathbf{S}+\delta\mathbf{I}\right)\right|\right)+\frac{1}{\sigma_g^2}\left(\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\left(\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\boldsymbol{\beta}\right)\right). \tag{1.4}$$

The "Fa" in FaST-LMM gets its name from these factorizations. 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

$$-\frac{1}{2}\left(n\log\left(2\pi\sigma_g^2\right)+\sum_{i=1}^{n}\log\left(\left[\mathbf{S}\right]_{ii}+\delta\right)+\frac{1}{\sigma_g^2}\sum_{i=1}^{n}\frac{\left(\left[\mathbf{U}^{\mathsf{T}}\mathbf{y}\right]_i-\left[\mathbf{U}^{\mathsf{T}}\mathbf{X}\right]_{i:}\boldsymbol{\beta}\right)^2}{\left[\mathbf{S}\right]_{ii}+\delta}\right). \tag{1.5}$$

Note that this expression is equal to the product of n single-variate Normal distributions, on the data transformed by $\mathbf{U}^{\mathsf{T}}$, yielding the equation

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

Having pre-computed the spectral decomposition of $\mathbf{K}$, we can rotate the phenotype and all SNPs once to get $\mathbf{UX}$ and $\mathbf{Uy}$. Given the parameters δ,σ_g^2 and $\boldsymbol{\beta}$ each evaluation of the likelihood is now linear in the cohort size n, as compared to cubic for direct evaluation of Equation 1.1.

1.2 Finding the maximum likelihood fixed effect weights efficiently

We take the gradient of the log likelihood in Equation 1.4 with respect to $\boldsymbol{\beta}$ and set it to zero, giving

$$\mathbf{0}=\frac{1}{\sigma_g^2}\Big(\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}\Big).$$

Multiplying both sides by σ_g^2 and then bringing the part involving $\hat{\boldsymbol{\beta}}$ to one side, we get

$$\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}=\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right).$$

Multiplying both sides by the inverse of the factor on the left side, we obtain

$$\hat{\boldsymbol{\beta}}=\left[\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\right]^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right).$$

As $(\mathbf{S}+\delta\mathbf{I})$ is a diagonal matrix, the matrix products again can be written as a sum over n independent terms, yielding

$$\hat{\boldsymbol{\beta}}=\left[\sum_{i=1}^{n}\frac{1}{\left[\mathbf{S}\right]_{ii}+\delta}\left[\mathbf{U}^{\mathsf{T}}\mathbf{X}\right]_{i:}^{\mathsf{T}}\left[\mathbf{U}^{\mathsf{T}}\mathbf{X}\right]_{i:}\right]^{-1}\left[\sum_{i=1}^{n}\frac{1}{\left[\mathbf{S}\right]_{ii}+\delta}\left[\mathbf{U}^{\mathsf{T}}\mathbf{X}\right]_{i:}^{\mathsf{T}}\left[\mathbf{U}^{\mathsf{T}}\mathbf{y}\right]_i\right],$$

analogous to linear regression estimates for $\hat{\boldsymbol{\beta}}$ on the rotated data. Assuming that all the terms involving the spectral decomposition of $\mathbf{K}$ are precomputed, this equation can be evaluated in $O(n)$.

1.3 Finding the maximum likelihood genetic variance efficiently

We start by substituting $\hat{\boldsymbol{\beta}}$ from the previous section into the log likelihood, Equation 1.5, and set the derivative with respect to σ_g^2 to zero, giving

$$0 = -\frac{1}{2}\left(\frac{n}{\hat{\sigma}_g^2} - \frac{1}{\hat{\sigma}_g^4}\sum_{i=1}^{n}\frac{\left([\mathbf{U}^\mathrm{T}\mathbf{y}]_i - [\mathbf{U}^\mathrm{T}\mathbf{X}]_{i:}\hat{\boldsymbol{\beta}}\right)^2}{[\mathbf{S}]_{ii}+\delta}\right).$$

Multiplying both sides by $2\hat{\sigma}_g^4$ and solving for $\hat{\sigma}_g^2$, we get

$$\hat{\sigma}_g^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\left([\mathbf{U}^\mathrm{T}\mathbf{y}]_i - [\mathbf{U}^\mathrm{T}\mathbf{X}]_{i:}\hat{\boldsymbol{\beta}}\right)^2}{[\mathbf{S}]_{ii}+\delta}.$$

This equation also can be evaluated in $O(n)$.

1.4 Efficient evaluation of the maximum likelihood

Plugging in $\hat{\sigma}_g^2$ and $\hat{\boldsymbol{\beta}}$ into Equation 1.5, the log likelihood becomes a function only of δ, $LL(\delta, \hat{\sigma}_g^2(\delta), \hat{\boldsymbol{\beta}}(\delta)) = LL(\delta)$:

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

As described next, we optimize this function of δ using a one-dimensional numerical optimizer to find the maximum likelihood value of δ, from which the maximum likelihood values of all the parameters can be directly computed.

1.5 Optimization of δ

As we've just shown, finding the maximum log likelihood of our model ($LL\left(\sigma_e^2, \sigma_g^2, \boldsymbol{\beta}\right)$) is equivalent to finding the value of δ that maximizes $LL(\delta)$, a non-convex optimization problem. To avoid local maxima in FaST-LMM, a quasi-exhaustive one dimensional optimization scheme similar to the one proposed in [1] is applied. In order to bracket local minima, we evaluate the maximum of the log likelihood for 100 equidistant values of $\log(\delta)$, ranging -10 to 10. Then, we apply Brent's method (a 1D numerical optimization algorithm) to find the locally optimal δ in each bracket where the middle log likelihood is higher than the log likelihoods of the neighboring evaluations.

To speed-up a full GWAS scan, one can find the maximum likelihood setting for δ for just the null-model, re-using the same δ for all alternative models. This speedup was described in [2] and is used in all of our experiments unless otherwise noted.

2 Relationship between spectral decomposition and singular value decomposition for the RRM and other factored genetic similarity matrices

Before we discuss the low-rank version of FaST-LMM, it will be useful to review the relationship between spectral decomposition and singular value decomposition (SVD) for matrices, for which the factorization $\mathbf{K} = \mathbf{W}\mathbf{W}^\mathsf{T}$ is known, such as the RRM or the Eigenstrat covariance matrix [3]. In this section, we shall refer to a matrix $\mathbf{K}$ that has this form as being *factored*.

The spectral decomposition of the genetic similarity matrix, $\mathbf{K}$, given by $\mathbf{U}\mathbf{S}\mathbf{U}^\mathsf{T} = \mathbf{K}$, yields the eigenvectors ($\mathbf{U}$) and eigenvalues ($\mathbf{S}$) of $\mathbf{K}$. In general, this decomposition can be determined by first computing the genetic similarity matrix ($\mathbf{K}$), and then taking the spectral decomposition of it. For many measures of genetic similarity, including RRM, the time complexity of computing $\mathbf{K}$ is $O(n^2 s_\mathrm{c})$, where s_c is the number of SNPs used to compute $\mathbf{K}$. Given the genetic similarity matrix, the eigenvalues and eigenvectors of $\mathbf{K}$ can then be found solving the spectral decomposition at a time complexity of $O(n^3)$ and space complexity of $O(n^2)$. If only the first k eigenvectors are desired, the computation can be achieved with other algorithms that have time complexity of $O(n^2 k)$ and a space complexity of $O(n^2)$.

When $\mathbf{K}$ if factored, however, one can bypass explicit computation of $\mathbf{K}$, obtaining the required eigenvectors and eigenvalues by direct application of an SVD to the $n \times s_\mathrm{c}$ data matrix of SNP markers at a time complexity of $O(n s_\mathrm{c}^2)$ (or $O(n s_\mathrm{c} k)$ for only the top k eigenvectors using, for example, [4]) and space complexity of $O(n s_\mathrm{c})$. Construction of $\mathbf{K}$ can be bypassed because (1) the eigenvectors (equivalently, singular vectors) of the factored matrix are the same as the singular vectors of the data matrix, and (2) the eigenvalues (equivalently singular values) of the factored matrix are the square of the singular values of the data matrix. This relationship is widely-known (*e.g.*, [5]) and is demonstrated below. In our experiments, FaST-LMM bypasses computation of the factored matrix to obtain the required spectral decomposition whenever $s_\mathrm{c} < n$.

Note that, when the rank of $\mathbf{K}$ is less than the cohort size n (such as occurs when the data matrix used to compute the factored genetic similarity matrix represents fewer SNPs than individuals), the SVD with time cost $O(n s_\mathrm{c}^2)$ is actually an *economy* SVD, that is, it yields only the first s_c eigenvectors. This set of eigenvectors is denoted $\mathbf{U}_1$ in Section 3 and referred to as the k-spectral decomposition in the main paper.

We now demonstrate the relationship just noted. Let $\mathbf{W} \in \mathbb{R}^{n \times s_\mathrm{c}}$ [6] be the matrix containing the set of SNPs used to compute the factored matrix, $\mathbf{K}$, defined as

$$\mathbf{K} \equiv \mathbf{W}\mathbf{W}^\mathsf{T}. \tag{2.1}$$

Let $\mathbf{U}\tilde{\mathbf{S}}\mathbf{V}^\mathsf{T}$ be the SVD of $\mathbf{W}$. Then Equation 2.1 can be rewritten as

$$\mathbf{K} = \left(\mathbf{U}\tilde{\mathbf{S}}\mathbf{V}^\mathsf{T}\right)\left(\mathbf{U}\tilde{\mathbf{S}}\mathbf{V}^\mathsf{T}\right)^\mathsf{T} = \mathbf{U}\tilde{\mathbf{S}}\mathbf{V}^\mathsf{T}\mathbf{V}\tilde{\mathbf{S}}\mathbf{U}^\mathsf{T}.$$

Because $\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$, we obtain

$$\mathbf{K} = \mathbf{U}\tilde{\mathbf{S}}\tilde{\mathbf{S}}\mathbf{U}^\mathsf{T} = \mathbf{U}\mathbf{S}\mathbf{U}^\mathsf{T},$$

where $\mathbf{S}_{ii} \equiv \tilde{\mathbf{S}}_{ii}\tilde{\mathbf{S}}_{ii}$. By definition, $\mathbf{U}$ consists of the eigenvectors of $\mathbf{K}$ (because it satisfies the properties of a spectral decomposition of $\mathbf{K}$, namely that $\mathbf{K} = \mathbf{USU}^\mathsf{T}$ where $\mathbf{S}$ is diagonal and $\mathbf{U}$ contains orthonormal vectors). Furthermore, the eigenvalues of $\mathbf{K}$ are given by $\tilde{\mathbf{S}}_{ii}^2$. Consequently, we can obtain the spectral decomposition of $\mathbf{K}$ by computing the SVD of $\mathbf{W}$, which has time cost $O(ns_c^2)$.

3 LMMs with a low rank genetic similarity matrix

Now we consider the evaluation of the likelihood when the rank of $\mathbf{K}$, k, is low ($k < n$) (*i.e.*, $\mathbf{K}$ is not full rank). This condition will occur when the RRM is used and the number of SNPs used to estimate it, $s_c = k$, is smaller than n. It will also occur if we reduce the rank of $\mathbf{K}$ to $k \leq \min(n, s_c)$ by eliminating the eigenvectors with the lowest eigenvalues as described in Discussion of the main paper. We address both possibilities in this section.

Let $\mathbf{USU}^\mathsf{T} = \mathbf{K}$ be the complete spectral decomposition of $\mathbf{K}$. Thus, $\mathbf{S}$ is 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, and $\mathbf{U}$ is an $n \times n$ matrix of eigenvectors. Now, write the full $n \times n$ orthonormal matrix $\mathbf{U}$ as $\mathbf{U} \equiv [\mathbf{U}_1, \mathbf{U}_2]$, where $\mathbf{U}_1 \in \mathbb{R}^{n \times k}$ contains the eigenvectors corresponding to non-zero eigenvalues, and $\mathbf{U}_2 \in \mathbb{R}^{n \times n-k}$ contains the eigenvectors corresponding to zero eigenvalues. Thus, we have

$$\mathbf{K} = \mathbf{USU}^\mathsf{T} = [\mathbf{U}_1, \mathbf{U}_2] \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{bmatrix} [\mathbf{U}_1, \mathbf{U}_2]^\mathsf{T} = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^\mathsf{T} + \mathbf{U}_2\mathbf{S}_2\mathbf{U}_2^\mathsf{T}.$$

As $\mathbf{S}_2 = [\mathbf{0}]$, $\mathbf{K}$ can be recovered by the k-spectral decomposition of $\mathbf{K}$:

$$\mathbf{K} = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^\mathsf{T}.$$

The expression $(\mathbf{K} + \delta\mathbf{I})$, however, is always of full rank (because $\delta > 0$):

$$\mathbf{K} + \delta\mathbf{I} = \mathbf{U}\left(\mathbf{S} + \delta\mathbf{I}\right)\mathbf{U}^\mathsf{T} = \mathbf{U}\begin{bmatrix} \mathbf{S}_1 + \delta\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \delta\mathbf{I}_{n-k} \end{bmatrix}\mathbf{U}^\mathsf{T}.$$

Therefore, it is not possible to simply ignore $\mathbf{U}_2$ while using our previous approach (as in in Section 1), as $\mathbf{U}_2$ enters the expression for the log likelihood. Furthermore, directly computing the complete spectral decomposition does not exploit the low rank of $\mathbf{K}$. Thus, we use algebraic manipulations to rewrite the likelihood in terms not involving $\mathbf{U}_2$, as explained next. As a result, we incur only the computational complexity of computing $\mathbf{U}_1$ rather than $\mathbf{U}$.

3.1 Linear time evaluation of the likelihood

To exploit the low rank of $\mathbf{K}$ to evaluate the log likelihood efficiently, one possible approach would be to augment the spectrum using $n-k$ vectors that are orthogonal to the first k. Unfortunately, this strategy has a time complexity of $O((n-k)n^2)$. Consequently, we take the following alternative approach.

We begin with Equation 1.2:

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

The two terms involving $\mathbf{K}+\delta\mathbf{I}$ are highlighted in color and will be treated separately in the following.

As in Equation 1.5, the log-determinant of the genetic similarity matrix can be efficiently computed using the economy SVD of $\mathbf{X}$ to obtain the spectral decomposition of $\mathbf{K}$:

$$\log\left|(\mathbf{K}+\delta\mathbf{I})\right| = \sum_{i=1}^{n}\log\left([\mathbf{S}]_{ii}+\delta\right) = \sum_{i=1}^{k}\log\left([\mathbf{S}]_{ii}+\delta\right) + (n-k)\left(\log\delta\right), \tag{3.1}$$

where we use the fact that the last $n-k$ singular values are zero.

Also, as we show in Section 3.3, the residual quadratic form can be evaluated using the low-rank decomposition:

$$\begin{aligned}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{K}+\delta\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) &= \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{y}-\mathbf{U}_1^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{S}_1+\delta\mathbf{I}_k\right)^{-1}\left(\mathbf{U}_1^{\mathsf{T}}\mathbf{y}-\mathbf{U}_1^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}\right)\\ &+\frac{1}{\delta}\left(\left(\mathbf{I}_n-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{y}-\left(\mathbf{I}_n-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\left(\mathbf{I}_n-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{y}-\left(\mathbf{I}_n-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{X}\boldsymbol{\beta}\right).\end{aligned} \tag{3.2}$$

Furthermore, both terms in the expression on the right can be written as sums, leading to

$$\begin{aligned}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{K}+\delta\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) &= \sum_{i=1}^{k}\frac{\left([\mathbf{U}_1^{\mathsf{T}}\mathbf{y}]_i-[\mathbf{U}_1^{\mathsf{T}}\mathbf{X}]_{i:}\boldsymbol{\beta}\right)^2}{[\mathbf{S}]_{ii}+\delta} +\\ &\frac{1}{\delta}\sum_{i=1}^{n}\left(\left[\mathbf{y}-\mathbf{U}_1\left(\mathbf{U}_1^{\mathsf{T}}\mathbf{y}\right)\right]_i-\left[\mathbf{X}-\mathbf{U}_1\left(\mathbf{U}_1^{\mathsf{T}}\mathbf{X}\right)\right]_i\boldsymbol{\beta}\right)^2.\end{aligned} \tag{3.3}$$

3.2 Finding the maximum likelihood and parameters efficiently

Plugging both the determinant (Equation 3.1) and the quadratic form (Equation 3.3) into the log likelihood, we obtain

$$\begin{aligned}LL(\delta, \sigma_g^2, \boldsymbol{\beta}) = &-\frac{1}{2}\left(n\log\left(2\pi\sigma_g^2\right)+\sum_{i=1}^{k}\log\left([\mathbf{S}]_{ii}+\delta\right)+(n-k)\left(\log\delta\right)\right)\\ &-\frac{1}{2\sigma_g^2}\left(\sum_{i=1}^{k}\frac{\left([\mathbf{U}_1^{\mathsf{T}}\mathbf{y}]_i-[\mathbf{U}_1^{\mathsf{T}}\mathbf{X}]_{i:}\boldsymbol{\beta}\right)^2}{[\mathbf{S}]_{ii}+\delta}+\frac{1}{\delta}\sum_{i=1}^{n}\left(\left[\mathbf{y}-\mathbf{U}_1\left(\mathbf{U}_1^{\mathsf{T}}\mathbf{y}\right)\right]_i-\left[\mathbf{X}-\mathbf{U}_1\left(\mathbf{U}_1^{\mathsf{T}}\mathbf{X}\right)\right]_i\boldsymbol{\beta}\right)^2\right).\end{aligned} \tag{3.4}$$

Setting the gradient of $LL(\delta, \sigma_g^2, \boldsymbol{\beta})$ in Equation 3.4 with respect to $\boldsymbol{\beta}$ to zero, we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}} = &\left[\left(\sum_{i=1}^{k}\frac{1}{[\mathbf{S}]_{ii}+\delta}\left[\mathbf{U}_1^{\mathsf{T}}\mathbf{X}\right]_{i:}^{\mathsf{T}}\left[\mathbf{U}_1^{\mathsf{T}}\mathbf{X}\right]_{i:}\right)+\left(\frac{1}{\delta}\sum_{i=1}^{n}\left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{X}\right]_{i:}^{\mathsf{T}}\left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{X}\right]_{i:}\right)\right]^{-1}\\ &*\left[\left(\sum_{i=1}^{k}\frac{1}{[\mathbf{S}]_{ii}+\delta}\left[\mathbf{U}_1^{\mathsf{T}}\mathbf{X}\right]_{i:}^{\mathsf{T}}\left[\mathbf{U}_1^{\mathsf{T}}\mathbf{y}\right]_{i}\right)+\left(\frac{1}{\delta}\sum_{i=1}^{n}\left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{X}\right]_{i:}^{\mathsf{T}}\left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{y}\right]_{i}\right)\right].\end{aligned} \tag{3.5}$$

Nature Methods:doi10.1038/nmeth.1681

Plugging $\hat{\boldsymbol{\beta}}$ into the log likelihood and setting the derivative with respect to σ_g^2 to zero, we get

$$0 = -\frac{1}{2}\left(\frac{n}{\hat{\sigma}_g^2} - \frac{1}{\hat{\sigma}_g^4}\left(\sum_{i=1}^{k}\frac{\left(\left[\mathbf{U}_1^\mathsf{T}\mathbf{y}\right]_i - \left[\mathbf{U}_1^\mathsf{T}\mathbf{X}\right]_{i:}\hat{\boldsymbol{\beta}}\right)^2}{[\mathbf{S}]_{ii}+\delta} + \frac{1}{\delta}\sum_{i=1}^{n}\left(\left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\mathsf{T}\right)\mathbf{y}\right]_i - \left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\mathsf{T}\right)\mathbf{X}\right]_{i:}\hat{\boldsymbol{\beta}}\right)^2\right)\right).$$

Consequently,

$$\hat{\sigma}_g^2 = \frac{1}{n}\left(\sum_{i=1}^{k}\frac{\left(\left[\mathbf{U}_1^\mathsf{T}\mathbf{y}\right]_i - \left[\mathbf{U}_1^\mathsf{T}\mathbf{X}\right]_{i:}\hat{\boldsymbol{\beta}}\right)^2}{[\mathbf{S}]_{ii}+\delta} + \frac{1}{\delta}\sum_{i=1}^{n}\left(\left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\mathsf{T}\right)\mathbf{y}\right]_i - \left[\left(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\mathsf{T}\right)\mathbf{X}\right]_{i:}\hat{\boldsymbol{\beta}}\right)^2\right). \quad (3.6)$$

Plugging Equations 3.5 and 3.6 into 3.4 yields

$$\begin{aligned} LL(\delta,\hat{\sigma}_g^2,\hat{\boldsymbol{\beta}}) = &-\frac{1}{2}\left(n\log(2\pi) + \sum_{i=1}^{k}\log\left([\mathbf{S}]_{ii}+\delta\right) + (n-k)\left(\log\delta\right)\right) \\ &-\frac{1}{2}\left(n + n\log\frac{1}{n}\left(\sum_{i=1}^{k}\frac{\left(\left[\mathbf{U}_1^\mathsf{T}\mathbf{y}\right]_i - \left[\mathbf{U}_1^\mathsf{T}\mathbf{X}\right]_{i:}\hat{\boldsymbol{\beta}}\right)^2}{[\mathbf{S}]_{ii}+\delta} + \frac{1}{\delta}\sum_{i=1}^{n}\left(\left[\mathbf{y}-\mathbf{U}_1\left(\mathbf{U}_1^\mathsf{T}\mathbf{y}\right)\right]_i - \left[\mathbf{X}-\mathbf{U}_1\left(\mathbf{U}_1^\mathsf{T}\mathbf{X}\right)\right]_i\hat{\boldsymbol{\beta}}\right)^2\right)\right), \end{aligned} \quad (3.7)$$

which can be evaluated in $O(n+k)$.

3.3 Derivation of the low-rank quadratic form

Let $\mathbf{K}$ be a rank k genetic similarity matrix whose spectral decomposition can be written

$$\mathbf{K} = \mathbf{U}\mathbf{S}\mathbf{U}^\mathsf{T} = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^\mathsf{T} + \mathbf{U}_2\mathbf{S}_2\mathbf{U}_2^\mathsf{T} = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^\mathsf{T} + \mathbf{U}_2\left[\mathbf{0}\right]\mathbf{U}_2^\mathsf{T} = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^\mathsf{T},$$

where

$$\mathbf{U} = \left[\mathbf{U}_1, \mathbf{U}_2\right], \quad (3.8)$$

$\mathbf{U}_1 \in \mathbb{R}^{n\times k}$ contains the eigenvectors corresponding to non-zero eigenvalues, and $\mathbf{U}_2 \in \mathbb{R}^{n\times n-k}$. Using the fact that $\mathbf{U} \in \mathbb{R}_{n\times n}$ is a normal matrix, that is, $\mathbf{U}^{-1} = \mathbf{U}^\mathsf{T}$, we have

$$\mathbf{I}_n = \mathbf{U}\mathbf{U}^\mathsf{T} = \left[\mathbf{U}_1, \mathbf{U}_2\right]\left[\mathbf{U}_1, \mathbf{U}_2\right]^\mathsf{T} = \mathbf{U}_1\mathbf{U}_1^\mathsf{T} + \mathbf{U}_2\mathbf{U}_2^\mathsf{T}. \quad (3.9)$$

Solving Equation 3.9 for $\mathbf{U}_2\mathbf{U}_2^\mathsf{T}$, we get

$$\mathbf{U}_2\mathbf{U}_2^\mathsf{T} = \mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^\mathsf{T}. \quad (3.10)$$

Further, because the columns of $\mathbf{U}$ are orthonormal, it follows that

$$\mathbf{I}_n = \mathbf{U}^\mathsf{T}\mathbf{U},$$

$$\mathbf{I}_k = \mathbf{U}_1^\mathsf{T}\mathbf{U}_1,$$

$$\mathbf{I}_{n-k} = \mathbf{U}_2^{\mathsf{T}}\mathbf{U}_2. \tag{3.11}$$

Let $\mathbf{a} \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Our goal is to efficiently evaluate $\mathbf{a}^{\mathsf{T}} (\mathbf{K} + \delta\mathbf{I})^{-1} \mathbf{a}$. Substituting the spectral decomposition for $\mathbf{K}$ into this expression, we have

$$\mathbf{a}^{\mathsf{T}} (\mathbf{K} + \delta\mathbf{I})^{-1} \mathbf{a} = \left(\mathbf{U}^{\mathsf{T}}\mathbf{a}\right)^{\mathsf{T}} (\mathbf{S} + \delta\mathbf{I})^{-1} \left(\mathbf{U}^{\mathsf{T}}\mathbf{a}\right). \tag{3.12}$$

Using Equation 3.8, we can stack the matrix product in blocks involving $\mathbf{U}_1$ and $\mathbf{U}_2$ to re-write this expression as

$$\begin{bmatrix}\mathbf{U}_1^{\mathsf{T}}\mathbf{a} & \mathbf{U}_2^{\mathsf{T}}\mathbf{a}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix} (\mathbf{S}_1 + \delta\mathbf{I}_k)^{-1} & \mathbf{0} \\ \mathbf{0} & (\delta\mathbf{I}_{n-k})^{-1} \end{bmatrix} \begin{bmatrix}\mathbf{U}_1^{\mathsf{T}}\mathbf{a} & \mathbf{U}_2^{\mathsf{T}}\mathbf{a}\end{bmatrix}. \tag{3.13}$$

As the off-diagonal blocks of the central matrix are equal to zero, the quadratic form reduces to the sum of two terms, namely

$$\left(\mathbf{U}_1^{\mathsf{T}}\mathbf{a}\right)^{\mathsf{T}} (\mathbf{S}_1 + \delta\mathbf{I}_k)^{-1} \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{a}\right) + \left(\mathbf{U}_2^{\mathsf{T}}\mathbf{a}\right)^{\mathsf{T}} (\delta\mathbf{I}_{n-k})^{-1} \left(\mathbf{U}_2^{\mathsf{T}}\mathbf{a}\right). \tag{3.14}$$

Substituting $\mathbf{U}_2^{\mathsf{T}}\mathbf{U}_2$ for $\mathbf{I}_{n-k}$ (using Equation 3.11), the second term becomes

$$\left(\mathbf{U}_2^{\mathsf{T}}\mathbf{a}\right)^{\mathsf{T}} (\delta\mathbf{I}_{n-k})^{-1} \left(\mathbf{U}_2^{\mathsf{T}}\mathbf{a}\right) = \frac{1}{\delta}\mathbf{a}^{\mathsf{T}}\mathbf{U}_2\mathbf{I}_{n-k}\mathbf{U}_2^{\mathsf{T}}\mathbf{a} = \frac{1}{\delta}\mathbf{a}^{\mathsf{T}}\mathbf{U}_2 \left(\mathbf{U}_2^{\mathsf{T}}\mathbf{U}_2\right) \mathbf{U}_2^{\mathsf{T}}\mathbf{a}. \tag{3.15}$$

Finally, using Equation 3.10, we can eliminate $\mathbf{U}_2$ to obtain

$$\frac{1}{\delta} \left(\mathbf{U}_2\mathbf{U}_2^{\mathsf{T}}\mathbf{a}\right)^{\mathsf{T}} \left(\mathbf{U}_2\mathbf{U}_2^{\mathsf{T}}\mathbf{a}\right) = \frac{1}{\delta} \left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{a}\right)^{\mathsf{T}} \left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{a}\right). \tag{3.16}$$

Substituting (3.16) into (3.14), we obtain

$$\mathbf{a}^{\mathsf{T}} (\mathbf{K} + \delta\mathbf{I})^{-1} \mathbf{a} = \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{a}\right)^{\mathsf{T}} (\mathbf{S}_1 + \delta\mathbf{I}_k)^{-1} \left(\mathbf{U}_1^{\mathsf{T}}\mathbf{a}\right) + \frac{1}{\delta} \left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{a}\right)^{\mathsf{T}} \left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\mathbf{a}\right). \tag{3.17}$$

Substituting $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ for $\mathbf{a}$, we obtain Equation 3.2.

4 Restricted maximum likelihood

So far the derivations have been limited to maximum likelihood parameter estimation. However, it is straightforward to extend these results to the restricted log likelihood, which comprises the log likelihood (with $\hat{\boldsymbol{\beta}}$ plugged in), plus three additional terms [1]:

$$REMLL_{\mathrm{R}}\left(\sigma_e^2, \sigma_g^2\right) = LL\left(\sigma_e^2, \sigma_g^2, \hat{\boldsymbol{\beta}}\right) + \frac{1}{2}\left(d\log\left(2\pi\sigma_g^2\right) + \log\left|\mathbf{X}^{\mathsf{T}}\mathbf{X}\right| - \log\left|\mathbf{X}^{\mathsf{T}} (\mathbf{K} + \delta\mathbf{I})^{-1} \mathbf{X}\right|\right).$$

Again, using the spectral decomposition of $\mathbf{K}$, the restricted log likelihood becomes

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

Neglecting the cubic dependence on d for computing the determinants, these additional terms can be evaluated in time complexity $O(n)$. If $\mathbf{K}$ has rank $k < n$, we can evaluate the additional terms in

$O(n+k)$, using the k-spectral decomposition $\mathbf{K} = \mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathrm{T}}$. For this purpose, we re-use the results from Section 3.3, substituting $\mathbf{X}$ for $\mathbf{a}$, to get

$$REMLL_{\mathrm{R}}\left(\sigma_e^2,\sigma_g^2\right) = LL\left(\sigma_e^2,\sigma_g^2,\hat{\boldsymbol{\beta}}\right) + \frac{1}{2}\left(d\log\left(2\pi\sigma_g^2\right) + \log\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}\right|\right)$$
$$+\frac{1}{2}\left(-\log\left|\left(\mathbf{U}_1^{\mathrm{T}}\mathbf{X}\right)^{\mathrm{T}}\left(\mathbf{S}_1+\delta\mathbf{I}_k\right)^{-1}\left(\mathbf{U}_1^{\mathrm{T}}\mathbf{X}\right) + \frac{1}{\delta}\left(\left(\mathbf{I}_n-\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right)\mathbf{X}\right)^{\mathrm{T}}\left(\left(\mathbf{I}_n-\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right)\mathbf{X}\right)\right|\right).$$

The restricted maximum likelihood (*REML*) variance component estimate is given by

$$\hat{\sigma}_g^2 = \frac{1}{n-d}\sum_{i=1}^{n}\frac{\left([\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i - [\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\hat{\boldsymbol{\beta}}\right)^2}{[\mathbf{S}]_{ii}+\delta}.$$

The formulas for the remaining parameters remain unchanged. The space requirements for REML are the same as those for ML.

5 FaST-LMM for groups of genetically identical individuals and for compression

FaST-LMM can be made even more efficient when multiple individuals share the same genotype or when the LMM is compressed (as in compressed mixed linear models (CMLM) [7]). In either case, the likelihood can be written as

$$LL\left(\sigma_e^2,\sigma_g^2,\boldsymbol{\beta}\right) = \log\mathcal{N}\left(\mathbf{y}|\mathbf{X}\boldsymbol{\beta};\sigma_g^2\mathbf{Z}\mathbf{K}\mathbf{Z}^{\mathrm{T}}+\sigma_e^2\mathbf{I}\right), \tag{5.1}$$

where $\mathbf{Z}$ is an $n\times g$ binary indicator matrix, that assigns the data for each of n individuals to exactly one of the g groups, and $\mathbf{K}$ is a $g\times g$ between group genetic similarity matrix. The individuals in each group may have the same genotype, or merely a similar genotype as in the case of compression.

In the spirit of FaST-LMM, we look for an efficient way of computing the spectral decomposition of $\mathbf{Z}\mathbf{K}\mathbf{Z}^{\mathrm{T}}$. This spectral decomposition can then be plugged into Formulas 3.4-3.7 as a means to evaluate Equation 5.1, in run time and memory that are linear in the cohort size n. In Section 5.1, we consider the case where genetic similarity is defined by an RRM, given by $\mathbf{Z}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}$. We show that, given a $g\times s_{\mathrm{c}}$ matrix $\boldsymbol{\Phi}$ of s_{c} SNPs (in the case of compression obtained, e.g., by averaging the SNP data for individuals over the members of each group), the spectral decomposition of the RRM can be computed from the SVD of the $g\times s_{\mathrm{c}}$ matrix $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{1/2}\boldsymbol{\Phi}$ in $O(\min(g,s_{\mathrm{c}})gs_{\mathrm{c}})$ time and $O(gs_{\mathrm{c}})$ memory. (In the case of compression, the same $\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$ would be obtained if instead we used a group-wise average of the $n\times n$ RRM.) In Section 5.2, we consider arbitrary genetic similarity. We prove that, given any $g\times g$ positive semi-definite group similarity matrix $\mathbf{K}$, the spectral decomposition of the $n\times n$ matrix $\mathbf{Z}\mathbf{K}\mathbf{Z}^{\mathrm{T}}$ can be computed from the spectral decomposition of the much smaller $g\times g$ matrix $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{1/2}\mathbf{K}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{1/2}$ using $O(g^3)$ time and $O(g^2)$ memory.

5.1 Spectral decomposition of $\mathbf{Z\Phi\Phi^T Z^T}$

Let $\mathbf{\Phi}$ be the $g \times s_c$ matrix of SNP data. Let $\mathbf{Z}$ be the $n \times g$ group indicator matrix that assigns data for each of n individuals to exactly one group. Then the genetic similarity matrix becomes $\mathbf{Z\Phi\Phi^T Z^T}$.

For our argument, we use the fact that, given a matrix $\mathbf{A}$, both $\mathbf{AA^T}$ as well as $\mathbf{A^TA}$ share the same eigenvalues, and that these eigenvalues are given by the square of the singular values of $\mathbf{A}$. The eigenvectors of $\mathbf{AA^T}$ are given by the left singular vectors of $\mathbf{A}$; and the eigenvectors of $\mathbf{A^TA}$ are given by the right singular vectors of $\mathbf{A}$. So the eigenvalues of $\mathbf{Z\Phi\Phi^T Z^T}$ are the same as the eigenvalues of $\mathbf{\Phi^T Z^T Z\Phi} = \mathbf{\Phi^T}(\mathbf{Z^TZ})^{1/2}(\mathbf{Z^TZ})^{1/2}\mathbf{\Phi}$. Using the same argument, the latter matrix has the same eigenvalues as $(\mathbf{Z^TZ})^{1/2}\mathbf{\Phi\Phi^T}(\mathbf{Z^TZ})^{1/2}$. These eigenvalues are given by the square of the singular values of $(\mathbf{Z^TZ})^{1/2}\mathbf{\Phi}$, where $(\mathbf{Z^TZ})^{1/2}$ is a $g \times g$ diagonal matrix holding the square root of the number of members of each group on the diagonal. Because $(\mathbf{Z^TZ})^{1/2}$ is diagonal, multiplication can be done in $O(gs_c)$ time.

Let $\tilde{\mathbf{U}}\tilde{\mathbf{S}}\mathbf{V^T}$ be the SVD of $(\mathbf{Z^TZ})^{1/2}\mathbf{\Phi}$. Then the following holds:

$$\mathbf{Z\Phi\Phi^T Z^T} = \mathbf{Z}(\mathbf{Z^TZ})^{-1/2}(\mathbf{Z^TZ})^{1/2}\mathbf{\Phi\Phi^T}(\mathbf{Z^TZ})^{1/2}(\mathbf{Z^TZ})^{-1/2}\mathbf{Z^T}, \tag{5.2}$$

where $(\mathbf{Z^TZ})^{-1/2}$ is a $g \times g$ diagonal matrix, holding one over the square root of the number of members of each group on its diagonal. Substituting $(\mathbf{Z^TZ})^{1/2}\mathbf{\Phi}$ by its SVD, we get

$$\mathbf{Z}(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}}\tilde{\mathbf{S}}\mathbf{V^T}\mathbf{V}\tilde{\mathbf{S}}\tilde{\mathbf{U}}^{\mathbf{T}}(\mathbf{Z^TZ})^{-1/2}\mathbf{Z^T}.$$

Finally, by orthonormality of $\mathbf{V}$, this expression simplifies to

$$\mathbf{Z}(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}}\mathbf{S}\tilde{\mathbf{U}}^{\mathbf{T}}(\mathbf{Z^TZ})^{-1/2}\mathbf{Z^T}, \tag{5.3}$$

where $\mathbf{S} = \tilde{\mathbf{S}}^2$ is a diagonal matrix, holding the non-zero eigenvalues of $\mathbf{\Phi\Phi^T}$ on its diagonal. The columns of $\mathbf{Z}(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}}$ are orthonormal, as can be seen by

$$\tilde{\mathbf{U}}^{\mathbf{T}}(\mathbf{Z^TZ})^{-1/2}\mathbf{Z^TZ}(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}} = \tilde{\mathbf{U}}^{\mathbf{T}}(\mathbf{Z^TZ})(\mathbf{Z^TZ})^{-1}\tilde{\mathbf{U}} = \tilde{\mathbf{U}}^{\mathbf{T}}\tilde{\mathbf{U}} = \mathbf{I}_g. \tag{5.4}$$

where we have once again used the fact that $(\mathbf{Z^TZ})$ is diagonal. It follows that $\mathbf{Z}(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}}$ holds the eigenvectors of $\mathbf{Z\Phi\Phi^T Z^T}$, completing the spectral decomposition of $\mathbf{Z\Phi\Phi^T Z}$. Note that the rotation of the data by $(\mathbf{Z}(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}})^{\mathbf{T}}$ can be done efficiently by multiplying the data by the transpose of the rows of $(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}}$ belonging to the respective group.

5.2 Spectral decomposition of $\mathbf{ZKZ^T}$, when the factors are not known

Here we extend the arguments in Section 5.1 to any positive semi-definite $g \times g$ group genetic similarity matrix $\mathbf{K}$. In this case, the spectral decomposition of $\mathbf{ZKZ^T} = \mathbf{USU^T}$ can also be computed efficiently, namely from the spectral decomposition of $(\mathbf{Z^TZ})^{1/2}\mathbf{K}(\mathbf{Z^TZ})^{1/2} = \tilde{\mathbf{U}}\mathbf{S}\tilde{\mathbf{U}}^{\mathbf{T}}$, which can be computed in $O(g^3)$ run time. As $\mathbf{K}$ is positive semi-definite, there always exists some square root $\mathbf{\Phi}$ of $\mathbf{K}$, such that $\mathbf{K} = \mathbf{\Phi\Phi^T}$. In Section 5.1, we have shown, that $\mathbf{ZKZ^T}$ and $(\mathbf{Z^TZ})^{1/2}\mathbf{K}(\mathbf{Z^TZ})^{1/2}$ have the same eigenvalues. Consequently, we can compute the eigenvalues $\mathbf{S}$ of $\mathbf{ZKZ^T}$ from the spectral decomposition $\tilde{\mathbf{U}}\mathbf{S}\tilde{\mathbf{U}}^{\mathbf{T}}$. Analogous to the derivation in Equations 5.2-5.3, it follows that the eigenvectors of $\mathbf{ZKZ^T}$ are $\mathbf{Z}(\mathbf{Z^TZ})^{-1/2}\tilde{\mathbf{U}}$, where by Equation 5.4, the columns are orthonormal.

References

[1] Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **107** (2008).

[2] Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

[3] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904–909 (2006).

[4] Tipping, M. & Bishop, C. M. Probabilistic principal component analysis. *J.R. Statistical Society, Series B* **61**, 6111–622 (1999).

[5] Wall, M. E., Rechtsteiner, A. & Rocha, L. M. *A Practical Approach to Microarray Data Analysis* (Kluwer, Norwell, MA, 2003).

[6] Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

[7] Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).

Supplementary Note 2: Null-Model Contamination

In our experiments measuring the accuracy of association P values in the main paper, the SNPs being tested and the SNPs used to estimate genetic similarity were deliberately made disjoint. Here, we discuss the reason for this approach.

As discussed in the main paper, a 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 linear weights (i.e., SNP effects) integrated over independent Normal distributions having the same variance [1]. So, by using a LMM with an RRM to test a given SNP for an association with the phenotype, we are in effect adjusting for *background SNPs,* precisely those used to construct the RMM. Thus, when testing a given SNP, using that SNP in the computation of the RRM would be equivalent to using that SNP as a regressor in the null model, making the log likelihood of the null-model higher than it should be, thus making the P value higher than it should be. We call this phenomenon *null-model contamination.* A weaker form of this phenomenon could exist due to linkage disequilibrium.

In this note, we show that, on the WTCCC data for the CD phenotype with non-white individuals and close family members included, this effect produces substantially deflated P values as measured by the λ statistic, and quantify the degree to which LD plays a role. In an ideal experiment, we would compare the λ statistic for two association tests of each available SNP, one where the RRM is constructed from all SNPs, and one where the RRM is constructed from all SNPs but the SNP being tested (and those nearby having at least a certain amount of LD with it). Unfortunately, such a comparison is computationally infeasible, as it would require the construction of many thousands of RRMs and their corresponding spectral decompositions.

Instead, we used an approach where the SNPs used to construct the RRM were chosen to be systematically further and further away from a set of test SNPs, while holding the number of SNPs used to construct the RRM (i.e., the number of background SNPs in the equivalent linear regression) constant. In particular, after ordering SNPs by their position, we used every thirty-second SNP starting from the i^{th} SNP in each chromosome to form a set of test SNPs. In addition, we created six sets of SNPs to construct RRMs, each set lying further away from the set of test SNPs. In a given set, we included every thirty-second SNP starting at the $i + j^{\text{th}}$ SNP in each chromosome, $j = 0, 1, 2, 4, 8$, and 16. This experiment was performed for $i = 1, 2, 3, 4$, and 5. Each set of SNPs contained approximately 11K SNPs. As shown in **Fig. 1**, λ generally increased with j for $j \leq 8$, beyond which LD presumably had little effect. Note that the values for λ for the experiments having the greatest amount of null contamination ($j = 0$) were quite similar to those when all 367K SNPs were used to construct the RRM (differences were less than 0.027 over all values of i), suggesting that our experiment did not deviate substantially from the idealized one.

These experiments show that null-model contamination can be a substantial effect. Consequently, when using a LMM to test whether a given SNP is associated with the phenotype, the RRM should be computed from all SNPs except for those in close proximity to the test SNP. As this approach is again computationally infeasible, in our experiments in the main paper evaluating the accuracy of association P values, we tested SNPs on chromosome 1 and constructed the "gold-standard" RRM from all SNPs on all but chromosome 1. We used chromosome 1 because it has a large number of

Nature Methods:doi10.1038/nmeth.1681

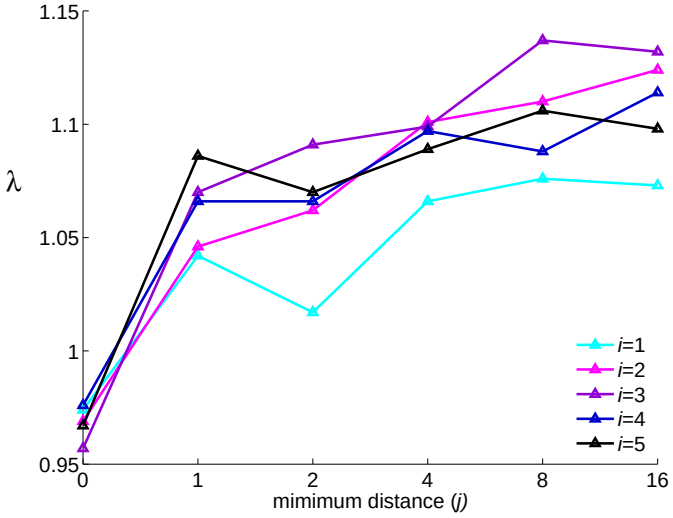


Figure 1: The λ statistic as a function of the minimum distance between a SNP in the test set and a SNP in the set used to construct the RRM. Each set of SNPs was selected by incorporating every thirty-second SNP along each chromosome starting at position i.

SNPs for testing and because there were enough genome-wide significant SNPs to assess the effects of sampling on calls of significance.

References

[1] Goddard, M. E., Wray, N., Verbyla, K. & Visscher, P. M. Estimating effects and making predictions from genome-wide marker data. *Statis. Sci* **24**, 517–529 (2009).

Nature Methods:doi10.1038/nmeth.1681