Sci-LMM is an efficient strategy for inferring genetic variance components using population scale family trees

Tal Shor,[1,2,*] Dan Geiger,[1] Yaniv Erlich[2,3,4] and Omer Weissbrod[1,5,**]

[1] Computer Science Department, Technion - Israel Institute of Technology, Haifa, Israel

[2] MyHeritage Ltd., Or Yehuda, Israel

[3] The New York Genome Center, New York, NY, USA

[4] Department of Computer Science, Fu School of Engineering, Columbia University, NY, USA

[5] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[*]Correspondence: tal.shor@myheritage.com
**Correspondence: oweissbrod@hsph.harvard.edu

Abstract

The rapid digitization of genealogical and medical records enables the assembly of extremely large pedigree records spanning millions of individuals and trillions of pairs of relatives. Such pedigrees provide the opportunity to answer genetic and epidemiological questions in scales much larger than previously possible. Linear mixed models (LMMs) are often used for analysis of pedigree data. However, traditional LMMs do not scale well due to steep computational and storage requirements. Here, we propose a novel modeling framework called **S**parse **C**holesky factor**I**zation LMM (Sci-LMM), that alleviates these difficulties by exploiting the sparsity patterns found in population-scale family-trees. The proposed framework constructs a matrix of genetic relationships between trillions of pairs of individuals in several hours, and can fit the corresponding LMM in several days. We demonstrate the capabilities of Sci-LMM via simulation studies and by estimating the heritability of longevity in a very large pedigree spanning millions of individuals and over five centuries of human history. The Sci-LMM framework enables the analysis of extremely large pedigrees that was not previously possible. Sci-LMM is available at https://github.com/TalShor/SciLMM.

Introduction

Pedigrees records can reflect social and cultural structures, and serve as a record of the flow of genetic material throughout human history. In recent years, very large pedigree records have come into existence, owing to collaborative digitization of large genealogical records (Kaplanis et al., 2018) and to digitization of large cohorts collected by healthcare providers, spanning up to millions of individuals (Huang et al., 2018; Polubriaginof et al., 2017; Smoller, 2017). Such population-scale pedigrees hold the potential to answer genetic epidemiological questions on a scale that is orders of magnitude larger than existing studies. However, the analysis of such data requires modeling complex covariance structures between trillions of pairs of individuals.

Studies with traditional pedigree data often employ LMM-based approaches to decompose the phenotypic variation among individuals into a series of variance components such as additive genetic effects and shared environment (Kruuk and Hadfield, 2007; Yu et al., 2006). In recent years, LMM-based approaches also gained considerable popularity in the analysis of DNA data to estimate heritability (Yang et al., 2010), perform phenotype prediction (Golan and Rosset, 2014; de Los Campos et al., 2013) or model population structure (Price et al., 2010). Previous studies suggested effective solutions for large-scale LMM using genetic similarity matrices based on SNP data (Loh et al., 2015). However, these approaches are largely unsuitable for pedigree data, because they require knowing the decomposition of the matrices encoding the phenotypic covariance between individuals beforehand. Hence, novel approaches are required for analysis of population-scale pedigrees.

Here, we propose a flexible statistical framework for analyzing pedigrees with trillions of relative pairs, called Sci-LMM. The underlying idea behind our approach is that pedigree data, unlike SNP-based genetic similarity, gives rise to a sparse data structure, which enables state of the art numerical techniques to store and analyze such data efficiently (Chen et al., 2008). In particular, a pair of individuals with no known common ancestor can be regarded as having no genetic similarity. Consequently, these pairs induce a zero entry in the genetic similarity matrix. However, as our pedigrees encompass millions of pairs with genealogical paths between them, the genetic similarity matrix still holds extensive information about relatedness, despite its sparsity.

Our contributions are two-fold: First, we provide an efficient algorithm for constructing very large sparse matrices that encode pedigree records. Second, we provide a unified framework for fitting an LMM in the presence of sparse covariance matrices, using either a moments-based estimation (Haseman and Elston, 1972) or restricted maximum likelihood (REML) (Gilmour et al., 1995). Our framework therefore provides a comprehensive solution for LMM-based pedigree analysis.

To demonstrate the capabilities of our approach, we carry out an extensive analysis of simulated data with millions of individuals, which we complete in a manner of hours. We additionally estimate the heritability of longevity, using a large cohort spanning millions of genealogical records and several centuries of human history (Kaplanis et al., 2018). Our framework enables analysis of large pedigree records that was not previously possible.

Material and Methods

In the following paragraphs, we present the basic elements of each one of our algorithms. For brevity, the text describes the most important aspects necessary to understand our approach. The Supplementary Material presents the full technical details.

Linear Mixed Models

Consider a sample of n individuals with observed phenotypes $y_1, \dots, y_n$, and covariates vectors $\boldsymbol{C}_1, \dots, \boldsymbol{C}_n$, and consider a set of $n \times n$ covariance matrices $M^1, \dots, M^d$, where $M^k_{i,j}$ encodes the covariance between the phenotypes of individuals i and j according to the kth covariance structure, up to a scaling constant. Under LMMs, the vector $\boldsymbol{y} = [y_1, \dots, y_n]^T$ follows a multivariate normal distribution:

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{C\beta}; \sum_{k=1}^{d} \sigma_k^2 M^k\right).$$

Here, $\boldsymbol{C} = [\boldsymbol{C}_1, \dots, \boldsymbol{C}_n]^T$ is an $n \times c$ matrix of covariates (including an intercept), $\boldsymbol{\beta}$ is a $c \times 1$ vector of fixed effects, and σ_k^2 is the kth variance component. Typically, one of the covariance matrices is the identity matrix, which encodes all sources of variance that is not captured by the other matrices.

Computing an Identical by Descent Matrix

The IBD (identity by descent) kinship coefficient of two individuals, denoted as a_{ij}, is the probability that a randomly selected allele in an autosomal locus was originated from the same chromosome of a shared ancestor between individuals i and j (Speed and Balding, 2015; Wright, 1922), and is given by:

$$a_{i,j} = \begin{cases} 1 + f_i, & i = j \\ r_{ij}\sqrt{a_{ii}a_{jj}}, & i \neq j \end{cases}$$

Here, f_i is the inbreeding coefficient, defined as half of the IBD coefficient of the parents of individual i (Speed and Balding, 2015), and r_{ij} is the coefficient of relationships, defined as:

$$r_{ij} = \frac{\sum_{path} \frac{1 + f_A}{2^{|path|+1}}}{\sqrt{(1 + f_i)(1 + f_j)}}$$

The quantities in the above equation are defined as follows: A is a least common ancestor of individual i and j in the pedigree graph (a graph where every node is an individual connected to her parents and children); the summation is performed over every path connecting individuals i, j in the pedigree graph, culminating at some ancestor A, such that the path does not contain the same individual twice; and $|\text{path}|$ is the number of individuals on a path.

To efficiently compute the IBD matrix we first construct the matrices $\boldsymbol{L}$ and $\boldsymbol{H}$ of its decomposition $\boldsymbol{A} = \boldsymbol{LHL^T}$, where $\boldsymbol{L}$ is a lower triangular matrix such that $\boldsymbol{L_{ij}}$ contains the

fraction of genome shared between individuals i and her ancestor j, $\boldsymbol{H}$ is diagonal, and the matrices are ordered such that ancestors precede their descendants (Henderson, 1976) (Figure 1a-c). The matrices $\boldsymbol{L}$ and $\boldsymbol{H}$ can be computed efficiently via dynamic programming (Meuwissen and Luo, 1991). We used sparse matrix routines (Chen et al., 2008) to scale Sci-LMM up to cohorts with trillions of pairs of individuals.

Computing a Dominance Relationship Matrix

Dominancy represents the genetic variance due to co-ancestry of two alleles rather than one as in regular genetic similarity. The dominance relationship between individuals i and j, $\boldsymbol{D}_{ij}$, can be formulated as (Henderson, 1985):

$$\boldsymbol{D}_{ij} = \frac{1}{4}\left(\boldsymbol{A}_{f_i,f_j} \cdot \boldsymbol{A}_{m_i,m_j} + \boldsymbol{A}_{f_i,m_j} \cdot \boldsymbol{A}_{m_i,f_j}\right),$$

where $A_{k,l}$ is the IBD coefficient of individuals k, l, and f_k, m_k are the parents of individual k. It is easy to see that a necessary condition for nonzero dominancy entry is a nonzero IBD relationship. We used this condition for rapid computation of the dominance matrix.

Computing an Epistatic Relationship Matrix

Epistatic covariance encodes the assumption that variants interact multiplicatively to affect a given phenotype, and is proportional to the exponent of the corresponding IBD coefficient, i.e., $\left(A_{k,l}\right)^2$ for two-loci epistasis, $\left(A_{k,l}\right)^3$ for three-loci epistasis and so on (Kempthorne, 1954). Therefore, an epistatic covariance matrix can be computed in a straightforward manner given the IBD matrix.

Pruning of uninformative individuals

Population scale pedigree data typically presents heterogeneity of the completeness of records. For example, some individuals have records with high quality phenotypes, whereas other individuals are part of the pedigree but do not provide any phenotypic information. However, these individuals with missing data may still be required for IBD computation purposes. For example, consider a small pedigree of two siblings with phenotypic data and two parents and one uncle without phenotypic data. The parents are important to the IBD computation of the siblings despite their absence of phenotypic information, but the uncle is non-informative.

We devised an efficient algorithm for finding non-informative individuals whose removal does not affect heritability estimation. Briefly, we defined required individuals as individuals who have phenotypic and explanatory variables data, or individuals who appear in a lineage path connecting two individuals with such data with one of their least common ancestors. In our previous studies with population scale family trees, this algorithm reduced the computational operations of the matrix construction stage by several orders of magnitude.

Haseman-Elston Regression

HE regression estimates variance components by fitting the following linear regression model:

$$(y_i - \boldsymbol{C}_i\boldsymbol{\beta})(y_j - \boldsymbol{C}_j\boldsymbol{\beta}) = \sum_{k=1}^{d} \sigma_k^2 M_{ij}^k + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma_e^2)$ encodes residuals, as in standard linear regression, y is standardized to have a unit variance, and the estimation is performed over all pairs of individuals i, j. The estimation can thus be regarded as a linear regression problem. Typically, the fixed effects $\boldsymbol{\beta}$ are first estimated without considering the covariance matrices (which yields a consistent estimator under mild regularity conditions (McCulloch et al., 2003)), and are then incorporated into the moment estimator above. The least squares estimator of $\sigma_1^2, \dots, \sigma_d^2$ can be computed efficiently via sparse matrix routines, despite the fact that the linear regression is performed over possibly trillions of pairs of individuals. One disadvantage of HE regression is that it does not provide standard error estimates, because covariates and variance components are estimated separately in a heuristic manner.

Restricted Maximum Likelihood Estimator (REML)

The LMM restricted log likelihood (after discarding constant terms) is given by:

$$\ell(\boldsymbol{\beta}, \sigma_1^2, \dots, \sigma_d^2) = -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} + \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}|_+,$$

where $\boldsymbol{\Sigma}^{-1}$ is the inverse of the aggregated covariance matrix after projecting out the covariates (Supplementary Material), and $|\boldsymbol{\Sigma}|_+$ is the product of its non-zero eigenvalues (Lippert et al., 2014). This expression, as well as its partial derivative according to each variance component, can be computed efficiently via sparse matrix routines, enabling rapid REML estimation. Specifically, we can estimate $\boldsymbol{\Sigma}^{-1}\mathbf{y}$ and $\log|\boldsymbol{\Sigma}^{-1}|_+$ by first computing the sparse Cholesky factorization of $\boldsymbol{\Sigma}$, via sparse matrix routines, based on the CHOLMOD package (Chen et al., 2008). Importantly, the resulting lower diagonal matrix is itself sparse because CHOLMOD permutes the covariance matrix prior to its factorization, which allows us to store it in memory. This enables us to estimate REML for matrices with hundreds of thousands of individuals within a few minutes. Exact gradient estimation is more computationally challenging, but can be approximated efficiently via approximations proposed in (Loh et al., 2015). REML also provides analytical formulas for standard errors, which can be used to compute confidence intervals (Supplementary Material).

Since the REML optimization procedure is sensitive to initial values, we first obtain initial estimates via the moment estimator, and use these as the initial values for the REML procedure.

Computing environment

All experiments were conducted using a Linux workstation with a 2GHz Xeon E5 processor and 256Gb of RAM.

Results

Overview of Proposed Framework

Sci-LMM allows decomposing a range of factors contributing to phenotypic variance

We developed Sci-LMM to analyze population scale pedigrees in a flexible manner that takes various types of covariates and components of variance into account. Our approach, like traditional LMMs, assumes that phenotypes follow a multivariate normal distribution, whose mean is affected by fixed effects associated with covariates such as age and sex, and whose covariance is affected by variance components associated with dependency structures such as familial relations or geographical proximity. The aim is to identify the contribution of these fixed effects and variance components to decipher their relative contribution. For example, when modeling both geographical proximity and familiar relationships, the magnitudes of the variance components can indicate whether the studied trait is more strongly transmitted via inheritance or via environmental factors.

Currently, Sci-LMM supports the efficient computation of the following covariance matrices: additive genetic component (based on IBD), dominancy, and pair-wise epistasis. Sci-LMM facilitates the computation by computing the Cholesky decomposition of this matrix via dynamic programming

In addition to the covariance matrices, Sci-LMM also includes adjustments using principal components with fixed effects. These adjustments can capture major linear sources of variation in a dataset such as confounding due to population structure (Price et al., 2006) and thus improve the estimation accuracy. The running time of standard principal component computation scales cubically with the sample size. To address this, we exploited again the sparsity of our data to efficiently compute the top principal components (Sorensen, 1997).

To estimate the contribution of all of these fixed and random factors, we used REML. REML is similar to maximum likelihood but corrects for bias introduced due to fixed effects. While REML comes with substantial computational costs, we exploited the sparsity of the data to substantially speed up the calculation of the restricted likelihood and its gradient (Methods).

Simulation Studies

To evaluate the capabilities of Sci-LMM, we generated large synthetic pedigrees spanning 20 to 40 generations and various family structures, under a wide variety of settings. The pedigrees had 50,000-2,000,000 individuals, amounting to trillions of pairs of relatives. A subset of the individuals in each generation consists of children of individuals from either the previous generation, or from two generations in the past. To simulate patterns observed in real datasets, the simulations also included half-siblings, and individuals with less than two recorded parents (Supplementary Material).

Next, we used the pedigree structure to compute the additive, epistasis and dominance matrices. We then generated a normally distributed phenotype, using ten covariates and a covariance matrix given by a weighted combination of the above matrices. Unless otherwise stated, the sparsity factor (the fraction of non-zero entries in each matrix) was 0.001. We generated ten different datasets for each combination of sample size and sparsity factor that we investigated (Supplementary Material).

In all conditions, Sci-LMM yielded empirically unbiased estimates of the variance components compared to the true parameters used in the simulations. As expected, estimation accuracy substantially increased as the sample size grew, though the estimators became slightly less accurate when increasing the number of variance components, (Figure 2a-c). Specifically, the root mean square error (RMSE) was < 0.03 for all methods under all settings with more than 250,000 individuals, indicating <3% average error. To further validate our approach, we also analyzed the data with HE regression. We found that the average RMSE of HE was slightly smaller than the average RMSE of REML under relatively small sample sizes of up to 100,000 individuals (Figure 2a-c). These results possibly indicate that REML convergence is difficult in the presence of sparse covariance matrices with limited sample sizes. However, REML estimation was more accurate HE regression under larger samples, as requested for our method for population scale family trees. We also found that the estimation accuracy increased as the covariance matrices became less sparse, indicating that the estimators efficiently exploit the information found in non-zero covariance entries (Figure 2d-e).

Runtime

Next, we evaluated the matrix construction time. While Sci-LMM supports several variance components, the rate limiting factor is the IBD matrix construction. We found that the Sci-LMM run-time is heavily dependent on the sparsity of the matrix (Figure 3a) and scales linearly with the number of non-zero entries in the matrix (Figure 3b). For example, Sci-LMM takes less than 4 hours to construct an IBD matrix with 5×10^{11} pairs of possible relatives and a sparsity factor of 0.001.

Next, we investigated the variance component estimation run-time. We found that the Sci-LMM REML estimation for samples with 500,000 individuals (representing 250 million covariance entries) can be performed in less than 24 hours (Figure 3c), and that HE estimation for the same data can be performed in 16 seconds (Figure 3d). Overall, our results demonstrate that the Sci-LMM framework is scalable to extremely large pedigrees.

Analysis of Real Data

Encouraged by our simulation results, we used Sci-LMM to estimate the heritability of longevity based on large-scale pedigree records obtained from the Geni genealogical website (Kaplanis et al., 2018). After stringent quality control, we extracted approximately 441,000 individuals with birth and death dates. We first computed the IBD, dominance and epistasis matrices of these individuals, and then estimated the heritability of longevity using these matrices.

The corresponding IBD matrix of the population scale tree contained over 3 billion nonzero entries. It included the 441,000 core individuals and their informative ancestors, yielding a total of 1.6 million individuals. The submatrix consisting of only the core individuals included 251 million non-zero IBD coefficients (yielding a sparsity factor of ~0.001, in correspondence with the simulation studies). The dataset included 9.7 million pairs of individuals with a kinship coefficient corresponding to a >=20-degree relationship (Figure 4a). Sci-LMM constructed this matrix in 10 hours.

Next, we used Sci-LMM to estimate the heritability of longevity with covariates encoding gender and the top 20 principal components of the IBD matrix, and with covariance matrices encoding IBD and pairwise epistasis (Methods). A dominance matrix was not included because the analysis included a relatively small number of parent-child relationships, rendering this matrix almost completely equivalent to the identity matrix (Supplementary Material). The REML estimates (and their 95% confidence intervals) were: IBD: 17.8% (16.15%, 19.45%); pair-wise epistasis: 1.6% (0%, 4.99%); environmental effects: 80.6% (78.6%, 82.6%). The HE estimates were very similar, with IBD explaining 17.7% of the variance, pair-wise epistasis explaining 1.5%, and environmental effects explaining 80.8%. These results suggest that the narrow sense heritability of longevity is roughly equal to 17.8% with a high degree of certainty (Figure 4b).

Discussion

We have described a statistical framework for analysis of large pedigree records spanning millions of individuals. Our framework includes methodologies for constructing large sparse matrices given raw pedigree data, as well as methodologies for LMM analysis in the presence of the dependencies encoded by these matrices. Taken together, the proposed solution enables an end-to-end analysis of large pedigree records on a scale that was not previously possible.

We evaluated two methods for variance components estimation: HE regression and REML. HE regression is fast but provides only point estimates, whereas REML is more accurate and can also provide confidence intervals, but is slower. Hence, the two methods are complementary in terms of their strengths and weaknesses. In practice, we found that it is difficult to scale REML to datasets with more than 500,000 individuals due to technical issues in the sparse Choleksy factorization routines we used. In the future we intend to replace these routines with ones that enable scaling REML to larger datasets. Our recommendation is to use REML when possible, and HE regression otherwise.

Our work serves to demonstrate of the feasibility of large pedigree analysis. Here, we focused on dependency due to additive genetics (represented via IBD coefficients), epistasis and dominance. However, our framework is flexible and can be extended in various directions. For example, LMMs are often used in the geostatistics literature to model transmissible phenomena (Buhmann, 2001; Gaspari and Cohn, 1999; Gneiting, 1999, 2002; Sansò and Schuh, 1987; Wendland, 1995). A common idea in this domain is to generate sparse matrices by thresholding the covariance between individuals whose distance exceeds a certain

threshold. It is therefore possible to combine such models with a pedigree-based model and exploit the sparsity patterns in both domains.

A potential caveat of pedigree based analysis is that it is approximate by definition, because in practice every pair of individuals share a common ancestor, even if the ancestor is not recorded in the pedigree. This approximation can be motivated by the fact that the expected proportion of genome sharing decays exponentially with the degree of relatedness, indicating that distant individuals can be regarded as having zero shared genetic similarity with a negligible loss of accuracy.

In this work we focused on an LMM-based analysis of large pedigree records with no measured genotypes. In recent years, LMMs have been used extensively in GWAS, which typically consist of unrelated samples with measured genotypes (Yang et al., 2014). We anticipate that cohorts combining both types of data will become increasingly prevalent in the near future. For example, we and other online genealogy platforms allow users to upload their genetic information and link it with their genealogical profile. The analysis of such combined cohorts will likely require approaches that combine the pedigree analysis techniques described in this work with state of the art approaches for LMM-based analysis of GWAS samples, and remains a potential avenue for future work.

Supplemental Data

Supplemental Data include mathematical derivations, and can be found online.

Acknowledgements

We thank Regev Schweiger, Elior Rahmani, Eran Halperin and Saharon Rosset for fruitful discussion. This study was supported by a generous gift from Andria and Paul Heafy (Y.E) and the Burroughs Wellcome Fund Career Awards at the Scientific Interface (Y.E.).

Web Resources

Sci-LMM, https://github.com/TalShor/SciLMM

References

Buhmann, M.D. (2001). A New Class of Radial Basis Functions with Compact Support. Math Comput *70*, 307–318.

Chen, Y., Davis, T.A., Hager, W.W., and Rajamanickam, S. (2008). Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. ACM Trans. Math. Softw. TOMS *35*, 22.

Gaspari, G., and Cohn, S.E. (1999). Construction of correlation functions in two and three dimensions. Q. J. R. Meteorol. Soc. *125*, 723–757.

Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics *51*, 1440–1450.

Gneiting, T. (1999). Correlation functions for atmospheric data analysis. Q. J. R. Meteorol. Soc. *125*, 2449–2464.

Gneiting, T. (2002). Compactly supported correlation functions. J. Multivar. Anal. *83*, 493–508.

Golan, D., and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. Am J Hum Genet *95*, 383–393.

Haseman, J., and Elston, R. (1972). The investigation of linkage between a quantitative trait and a marker locus. Behav Genet *2*, 3–19.

Henderson, C. (1985). Best linear unbiased prediction of nonadditive genetic merits. J Anim Sci *60*, 111–117.

Henderson, C.R. (1976). A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics *32*, 69–83.

Huang, X., Elston, R.C., Rosa, G.J., Mayer, J., Ye, Z., Kitchner, T., Brilliant, M.H., Page, D., and Hebbring, S.J. (2018). Applying family analyses to electronic health records to facilitate genetic research. Bioinformatics *34*, 635–642.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., et al. (2018). Quantitative analysis of population-scale family trees using millions of relatives. Science, in press.

Kempthorne, O. (1954). The correlation between relatives in a random mating population. Proc. R. Soc. Lond. B Biol. Sci. *143*, 103–113.

Kruuk, L., and Hadfield, J. (2007). How to separate genetic and environmental causes of similarity between relatives. J. Evol. Biol. *20*, 1890–1903.

Lippert, C., Xiang, J., Horta, D., Widmer, C., Kadie, C., Heckerman, D., and Listgarten, J. (2014). Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. Bioinformatics *30*, 3206–3214.

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., Schizophrenia Working Group of the Psychiatric Genomics, C., de Candia, T.R., Lee, S.H., Wray, N.R., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet *47*, 1385–1392.

de Los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics *193*, 327–345.

McCulloch, C.E., Sciences, C.B. of the M., and Foundation (U.S.), N.S. (2003). Generalized Linear Mixed Models (Institute of Mathematical Statistics).

Meuwissen, T., and Luo, Z. (1991). Computing inbreeding coefficients in large populations. Genet. Sel. Evol. GSE *24*, 305–313.

Polubriaginof, F., Vanguri, R., Quinnies, K., Belbin, G., Yahi, A., Salmasian, H., Lorberbaum, T., Nwankwo, V., Li, L., Shervey, M., et al. (2017). Estimate of disease heritability using 7.4 million familial relationships inferred from electronic health records. BioRxiv.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet *38*, 904–909.

Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat Rev Genet *11*, 459–463.

Sansò, F., and Schuh, W.-D. (1987). Finite covariance functions. Bull. Géod. *61*, 331–347.

Smoller, J.W. (2017). The use of electronic health records for psychiatric phenotyping and genomics. Am. J. Med. Genet. B Neuropsychiatr. Genet. 9999:1–12.

Sorensen, D.C. (1997). Implicitly restarted Arnoldi/Lanczos methods for large scale eigenvalue calculations. In Parallel Numerical Algorithms, (Springer), pp. 119–165.

Speed, D., and Balding, D.J. (2015). Relatedness in the post-genomic era: is it still useful? Nat. Rev. Genet. *16*, 33–44.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Adv. Comput. Math. *4*, 389–396.

Wright, S. (1922). Coefficients of inbreeding and relationship. Am. Nat. *56*, 330–338.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet *42*, 565–569.

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat Genet *46*, 100–106.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet *38*, 203–208.

Figures

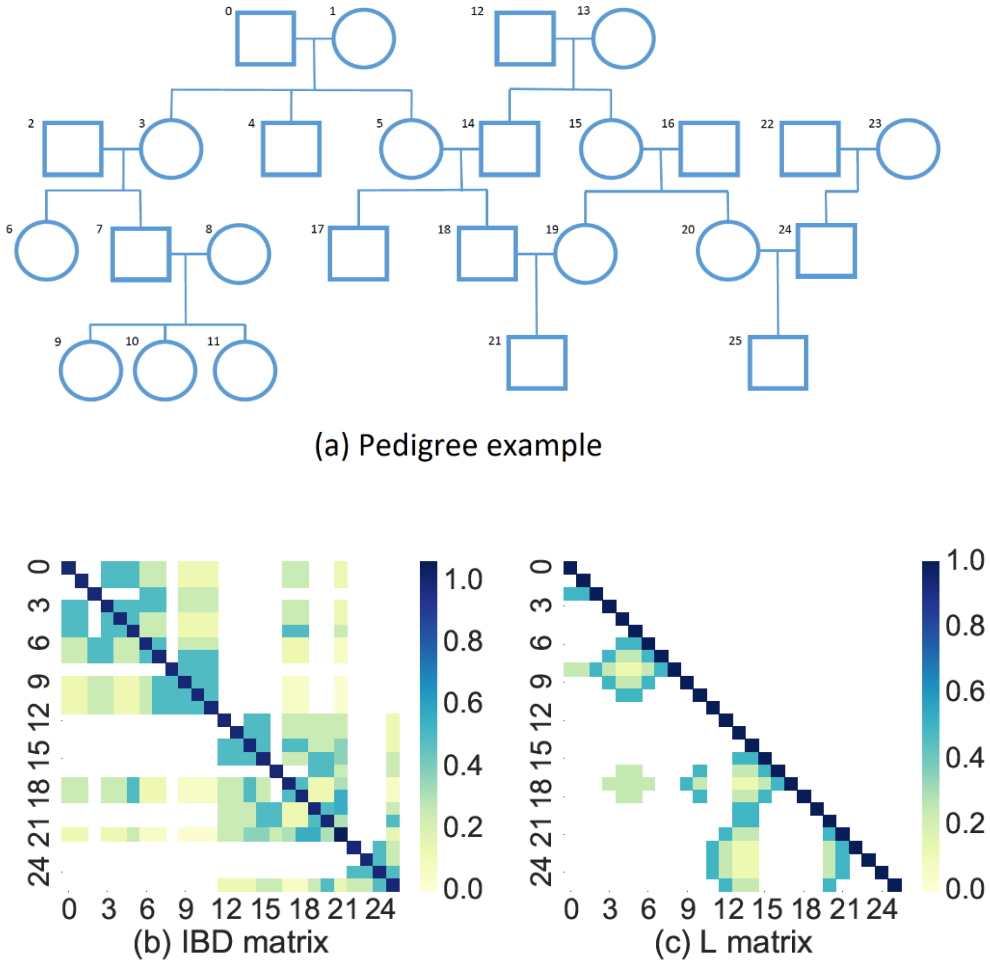

Figure 1: A demonstration of the Sci-LMM IBD matrix construction algorithm. (**a**) An example pedigree with 26 individuals. (**b**) A heat-map representing the IBD matrix, where zero elements are white to emphasize sparsity. (**c**) A heat-map representing the lower Cholesky factorization of the IBD matrix (i.e. the matrix L in the factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$, where $\boldsymbol{A}$ is the IBD matrix). The value of entry i, j is the expected fraction of the genome that is shared between individual i and her ancestor j.

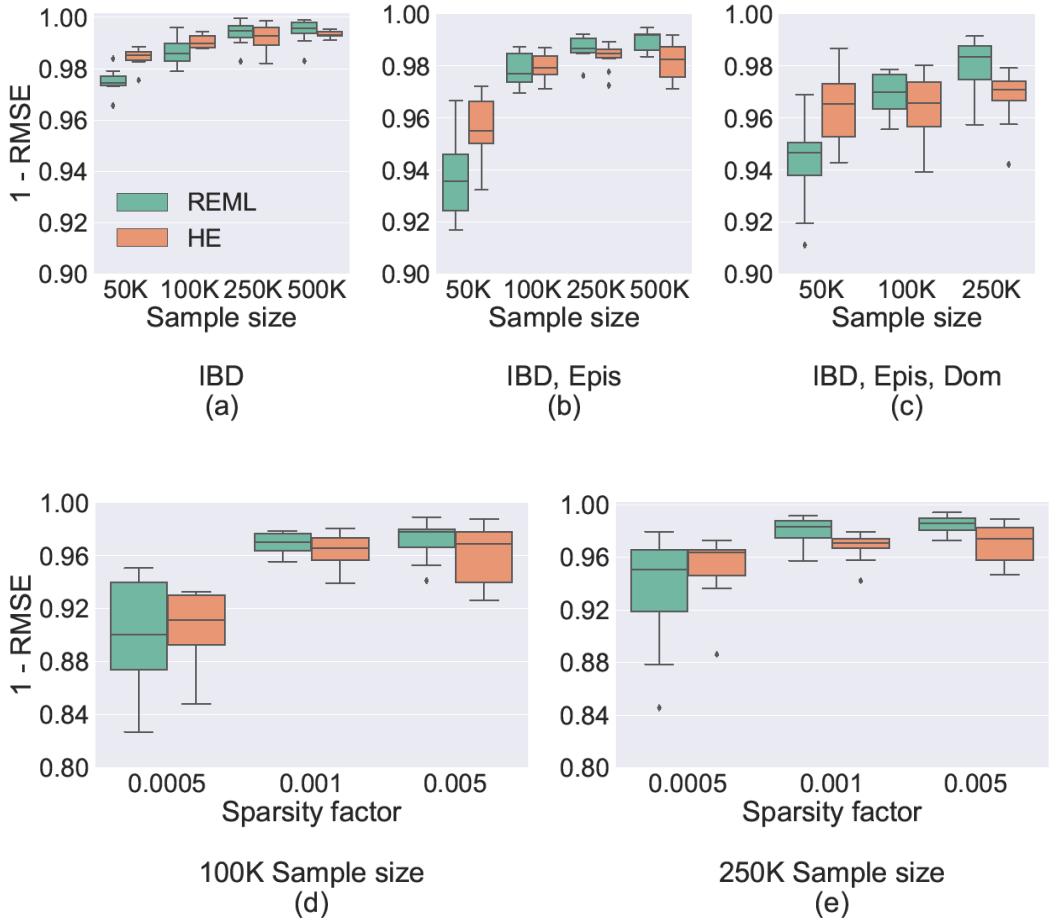

Figure 2: Evaluating the estimation accuracy of Sci-LMM. (**a-c**) Box plots comparing REML and HE estimation accuracy across simulated datasets (each box represents 10 experiments), under varying sample sizes, using **(a)** only IBD, **(b)** IBD and epistasis, or **(c)** IBD, epistasis and dominance variance components. The accuracy is measured via 1-RMSE (higher is better). HE is more accurate than REML for smaller sample sizes, but REML outperforms HE as the sample size increases. Results for analyses with three matrices and 500,000 individuals are omitted due to excessive required computational time. (**d-e**) Comparing REML and HE estimation accuracy when using IBD, epistasis and dominance matrices under various sparsity factors (the fraction of non-zero matrix entries) with either **(d)** 100,000 individuals, or (**e**) 250,000 individuals. The estimation accuracy of both REML and HE increases with the number of non-zero entries, for both REML and HE.

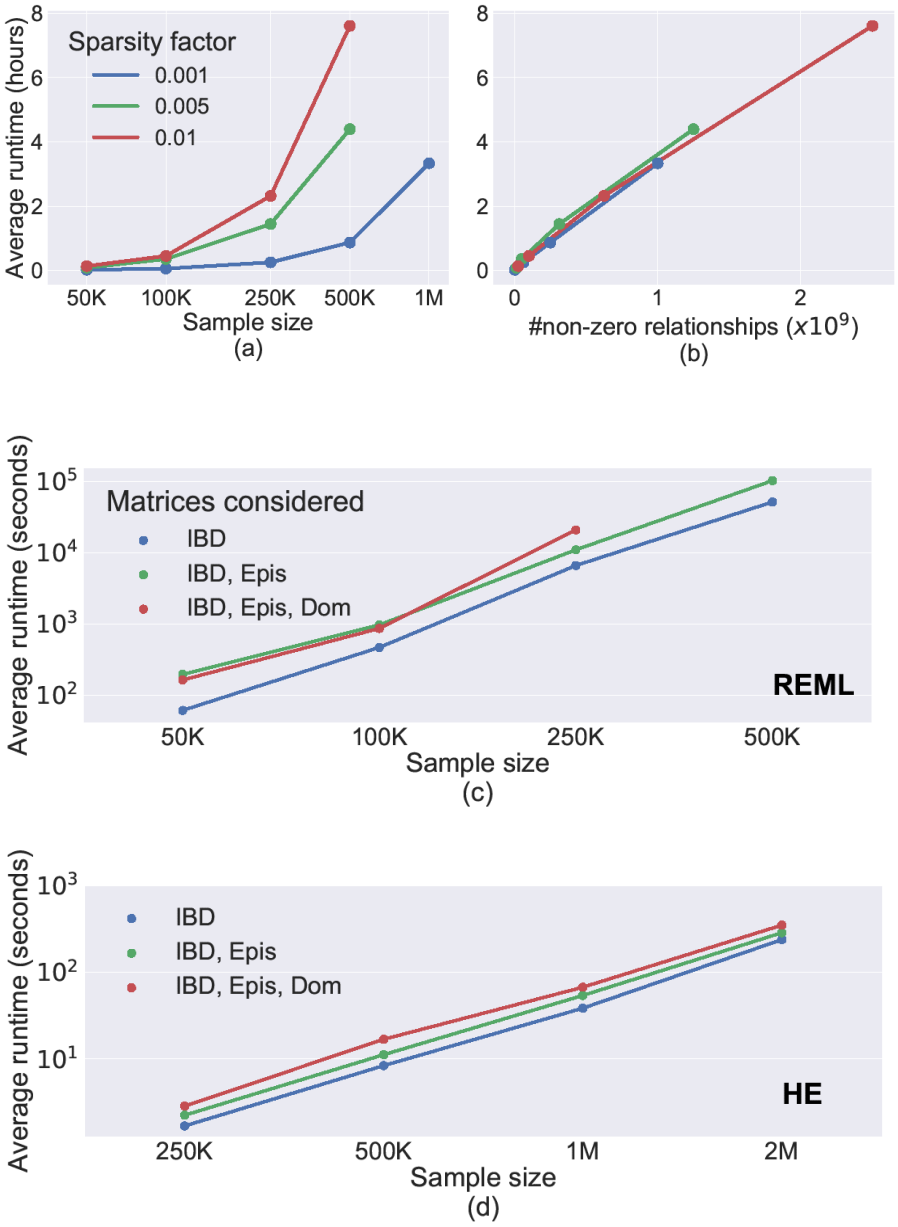

Figure 3: Analysis of Sci-LMM computation time. **(a)** Computation time required to compute an IBD matrix from pedigree data under different sparsity factors as a function of sample size. **(b)** Computation time required to compute an IBD matrix from pedigree data as a function of the number of nonzero relationships, demonstrating a linear relationship. The maximal number of evaluated non-zero relationships increases with the sparsity cutoff, because we only generated matrices with up to a million individuals. **(c)** Variance component estimation time (using REML), as a function of sample size, when using different combinations of covariance matrices. Epis – Epistasis; Dom - dominance **(d)** same as (c), but for HE regression instead of REML estimation. Here we evaluated datasets with up to 2 million individuals that were not investigated in (c), owing to technical limitations of the sparse matrix factorization routines used in our REML implementation.

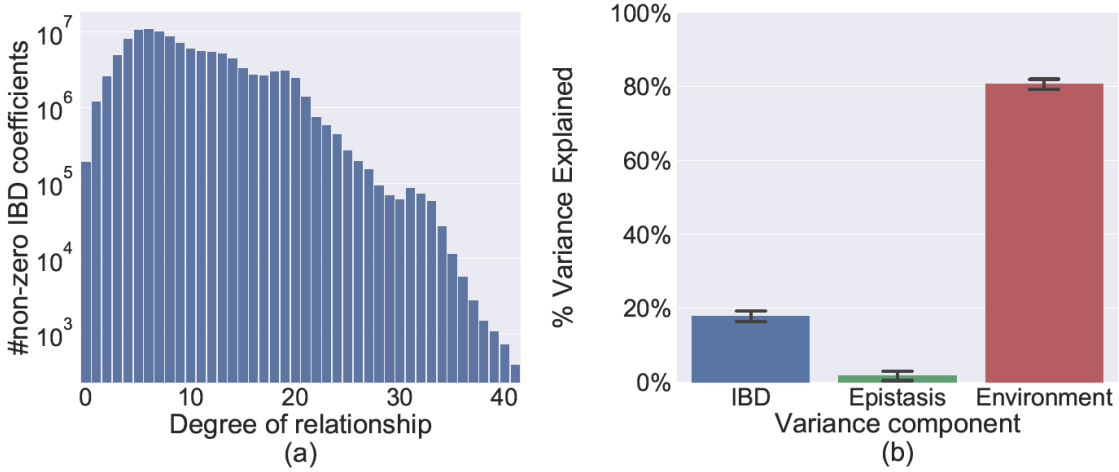

Figure 4: Results of analysis of a real pedigree with 441,000 individuals. **(a)** A histogram of genetic similarity across 441,000 individuals, using only the closest relationship between every pair of individuals. The dataset includes approximately 9.7 million pairs of individuals whose least common ancestor lived at least 10 generations earlier. **(b)** The estimated fraction of longevity variance attributed to different variance components (and their 95% CI).