SciLMM: Computing heritability with millions of individuals

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

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

[*]Correspondence: tal.shor@myheritage.com; 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. 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, LMMs cannot naturally scale to large pedigrees spanning millions of individuals, owing to their steep computational and storage requirements. Here we propose a novel modeling framework called **S**parse **C**holesky factor**I**zation LMM (SciLMM), that alleviates these difficulties by exploiting the sparsity patterns found in large pedigree data. The proposed framework can construct 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 SciLMM 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 SciLMM framework enables the analysis of extremely large pedigrees that was not previously possible. SciLMM 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., 2017) and to digitization of large cohorts collected by healthcare providers, spanning up to millions of individuals (Cáceres, 2013). For example, users of the online genealogy platform MyHeritage have collaborated to construct a pedigree spanning over a million individuals (MyHeritage Ltd, 2018). The accelerating digitization of medical records suggests that even larger pedigrees, spanning a large fraction of the population of developed nations, may soon come into existence.

Large pedigrees records hold the potential to provide answer to epidemiological questions on a scale that is orders of magnitude larger than existing studies. For example, such pedigrees can be employed for understanding the mode of transmission of genetic traits and diseases, or to tease apart genetic from non-genetic modes of transmission. However, the analysis of such data requires modeling complex covariance structures between trillions of pairs of individuals.

LMMs are often employed for analysis of pedigree data, as they can naturally encode dependencies between individuals due to inheritance or shared environment (Kruuk and Hadfield, 2007; Yu et al., 2006). At their core, LMMs encode the assumption that individuals who are more genetically or environmentally similar are more likely to exhibit similar phenotypes. In recent years, LMMs gained considerable popularity in the analysis of genome wide association studies (GWAS), as they can capture genetic similarities in order 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). However, existing LMM solutions for large samples require SNP data (Loh et al., 2015) and are thus not suitable for pedigree data with millions of individuals, necessitating development of novel approaches.

Here we propose SciLMM, a flexible statistical framework for analyzing pedigrees with millions of individuals. The underlying idea behind our approach is that pedigree data can be encoded in a sparse manner, which enables using state of the art numerical techniques to store and analyze such data in a very efficient manner (Chen et al., 2008). In particular, a pair of individuals with no known common ancestor can be regarded as having no genetic similarity. Consequently, pedigree analysis can use sparse matrices, which enables analyzing pedigree data with millions of individuals and dozens of explanatory variables.

Our contributions are two-fold: First, we provide an efficient algorithm for constructing very large sparse matrices that encode genetic and non-genetic similarity, based on pedigree records that link individuals with their parents. 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 hundreds of thousands 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., 2017). Our framework enables analysis of large pedigree records that was not previously possible.

Results

Overview of Proposed Framework

Here we provide a brief description of our proposed approach, with a longer description available in the Methods section.

Linear Mixed Models

LMMs are common statistical models for modeling covariance structures (see Methods for a formal description). LMMs assume that the phenotypes of a sample of individuals follow a multivariate normal distribution, whose mean is affected by covariates such as age and sex, and whose covariance matrix is affected by covariance structures such as familial relations or geographical proximity. Consequently, LMMs are parameterized by fixed effects, which describe the effect of the covariates on the mean, and by variance components, which describe the contribution of each covariance structure to the overall covariance.

The relative magnitudes of the variance components often serve as an indicator for the relative importance of different covariance structures, which can provide insights into the underlying mechanism of the studied phenotype. 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 infection.

Covariance Matrices

In this work, we used the following types of covariance matrices:

1. Identical by descent (IBD) matrix: The IBD coefficient between a pair of individuals is defined as the expected fraction of the genome inherited from common ancestors (Speed and Balding, 2015). To compute IBD from pedigree records efficiently, we combined the technique of Henderson (Henderson, 1976; Meuwissen and Luo, 1991) which computes the Cholesky decomposition of this matrix via dynamic programming, with efficient sparse matrix routines (Methods).
2. Dominance and epistasis matrices: Two other matrices commonly used in pedigree analysis are the dominance matrix, which captures shared genetic contents under a dominant mode of inheritance, and the pairwise epistatic matrix, which captures shared genetic contents under a mode of inheritance that allows for multiplicative interactions between pairs of genetic variants. Both matrices can be trivially constructed given the IBD matrix (Henderson, 1985) (Methods).

Principal Components

Principal components capture the major linear sources of variation in a data set. The inclusion of top principal components as covariates with fixed effects can often control for sources of confounding such as population structure (Price et al., 2006) and thus improve the estimation accuracy. Although the running time of standard principal component computation scales cubically with the sample size, sparse matrices enable efficient computation of top principal components (Sorensen, 1997). We therefore included the top principal components of the IBD matrix when estimating the heritability of longevity.

Estimating LMM Parameters

We used two different methods to estimate LMM parameters: Haseman-Elston (HE) regression (Haseman and Elston, 1972) and REML (Gilmour et al., 1995).

HE regression is a moment-based estimator, which assumes that the correlation between the phenotypes of every pair of individuals is a linear function of their covariance matrices entries. When all the matrices are sparse, this estimator can be computed efficiently in the presence of millions of individuals, via sparse matrix routines (Methods).

REML is similar to maximum likelihood estimation, but corrects for bias introduced due to fixed effects. REML is more statistically efficient than HE regression, but has a substantially larger computational cost. Nevertheless, restricted likelihood and its gradient can be computed extremely efficiently in the presence of sparse covariance matrices by exploiting sparse matrix techniques (Methods).

Simulation Studies

To evaluate the capabilities of our approach, we generated large synthetic pedigrees spanning 20 to 40 generations and various family structures, under a wide variety of settings. Briefly, we (1) created IBD, epistasis and dominance matrices for each pedigree; (2) generated a normally distributed phenotype, using a covariance matrix given by a weighted combination of these matrices; and (3) estimated the variance component and evaluated the estimation accuracy. Unless otherwise stated, the sparsity factor (the fraction of non-zero entries in each matrix) was 0.001. A full description is provided in the Methods.

Variance Component Estimation Accuracy

Our first experiment verified that both REML and HE yield empirically unbiased estimators in the presence of multiple variance components. To this end, we generated 130 data sets varying in size, estimated the variance components, and then compared the estimated and

true values. We repeated the above experiments both with and without using dominance and epistasis matrices.

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 1a-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. HE outperformed REML for samples with up to 100,000 individuals, whereas REML outperformed HE for larger samples. These results possibly indicate that REML convergence is difficult in the presence of sparse covariance matrices with limited sample sizes.

We also verified that 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 1d-e).

Runtime

When estimating heritability, SciLMM first creates IBD, dominance and pairwise-epistasis covariance matrices from pedigree data. We evaluated the IBD matrix computation time, since creation of the other matrices is trivial given the IBD matrix (Methods). The results indicate that IBD matrix construction for one million individuals with a sparsity factor of 0.001 requires less than 4 hours using a Linux workstation with a 2GHz Xeon E5 processor (Figure 2a). We additionally found that the runtime is approximately linear in the number of non-zero relationships (Figure 2b). Finally, we verified that REML estimation for samples with 500,000 individuals (representing 250 million covariance entries) can be performed in less than 24 hours (Figure 2c). These results demonstrate that the SciLMM framework is scalable to extremely large pedigrees.

Analysis of Real Data

We used SciLMM to estimate the heritability of longevity based on large pedigree records obtained from the GENI genealogical website, as described in (Kaplanis et al., 2017). After stringent quality control (described in (Kaplanis et al., 2017)) 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 this matrix.

To compute the IBD matrix, we included all shared ancestors of the 441,000 core individuals, yielding a cohort of 1.5 million individuals. The corresponding IBD matrix, which contains over 3 billion nonzero entries, was created in 10 hours by a single Linux workstation with a 2GHz Xeon E5 processor and 256 GB of memory. 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 results). The data set included 9.7 million pairs of individuals with a kinship coefficient corresponding to a >=20-degree relationship (Figure 3a).

Next, we estimated the heritability of longevity by fitting an LMM 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 3b).

Methods

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 kinship coefficient of two individuals, denoted as a_{ij}, is the probability that a randomly selected allele in an autosomal locus is identical by descent between individuals i and j (Speed and Balding, 2015; Wright, 1922), and is given by:

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

Where 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 4). 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 SciLMM up to cohorts with millions of individuals. A full description of the algorithm is provided in the Supplementary Material.

Computing a Dominance Relationship Matrix

A dominance matrix can model the effects of inherited alleles with a dominant mode of inheritance. Denote the dominance relationship matrix as **D**, IBD matrix as **A**, and f_k, m_k as the parents of individual k. The dominance relationship between individuals i and j can be formulated as (Henderson, 1985) :

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

It is easy to see that only nonzero IBD relationships might have nonzero dominance relationships, which enables computing the dominance matrix rapidly (Supplementary Material).

Computing a Pairwise Epistatic Relationship Matrix

Pairwise epistatic covariance encodes the assumption that pairs of variants interact multiplicatively to affect a given phenotype, and is given by the square of the corresponding IBD coefficient (up to a multiplicative constant) (Kempthorne, 1954). Therefore, epistatic covariance matrix can be computed in a straightforward manner given the IBD matrix.

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. A full description is provided in the Supplementary Material.

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).

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. A full description is provided in the Supplementary Material.

Generating Simulated data

We generated synthetic data sets with 20-40 generations with 50,000-1,000,000 individuals, and with 20 covariates. Briefly, we ran an iterative algorithm which generates one generation at a time. A subset of the individuals in each generation consists of children of individuals from one or several previous generations. To simulate patterns observed in real data sets, we allowed each individual to have zero, one or two recorded parents, according to the desired sparsity factor. We allowed each individual to have children with several spouses, where the

majority had a single partner. Most spouses are from the same generation and a small subset are one generation apart.

After creating a pedigree structure, we sampled random values for the covariates and the variance components, and then generated phenotype vectors according to the LMM. A full description of the simulation procedure is provided in the Supplementary Material.

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.

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 infections and other types of 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, the online genealogy platform MyHeritage has recently begun offering its users the ability to upload their genetic information and link it with their genealogical profile (MyHeritage, n.d.) 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.

References

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

Cáceres, S.B. (2013). Electronic health records: beyond the digitization of medical files. Clinics *68*, 1077–1078.

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.

Kaplanis, J., Gordon, A., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, G., Bhatia, G., MacArthur, D.G., Price, A., et al. (2017). Quantitative analysis of population-scale family trees using millions of relatives. BioRxiv.

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.

MyHeritage Ltd (2018). MyHeritage.com.

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.

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

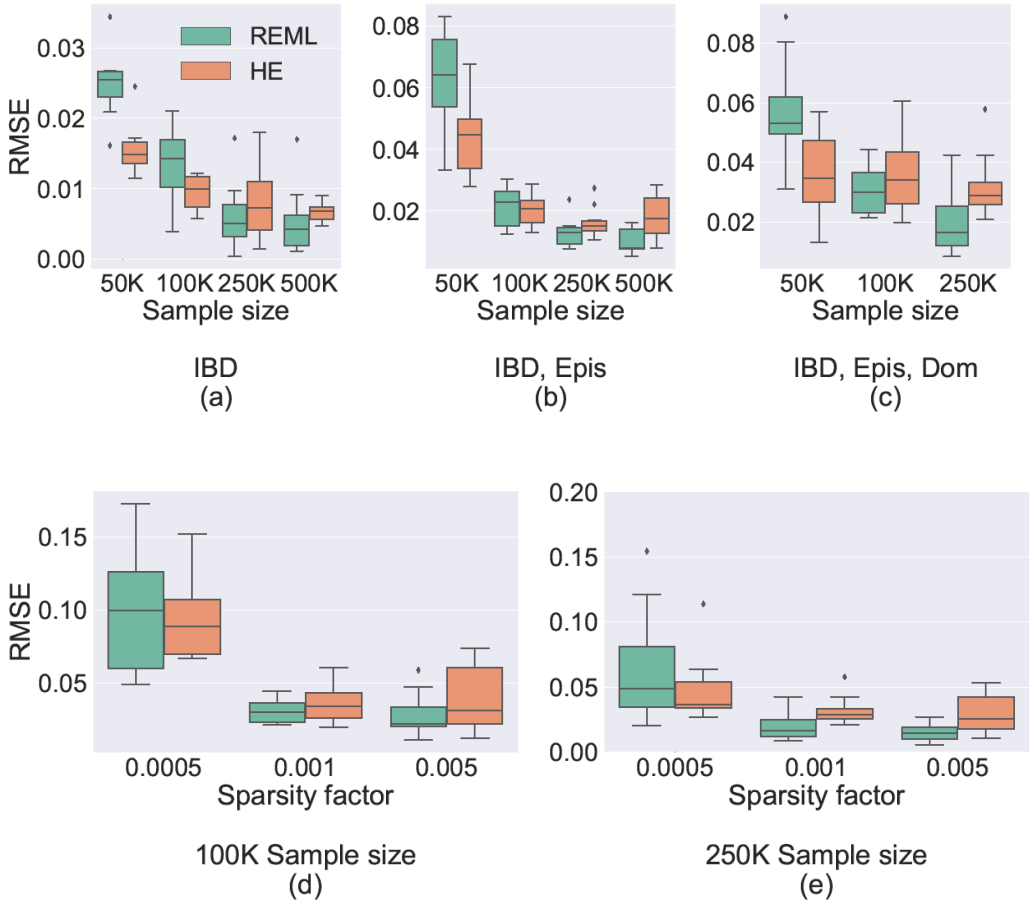

Figure 1: Evaluating the estimation accuracy of SciLMM. (**a-c**) Box plots comparing REML and HE estimation accuracy across simulated data sets (each box represents 10 experiments), under varying sample sizes, using **(a)** only IBD, **(b)** IBD and epistasis, or **(c)** IBD, epistasis and dominance variance components. 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.

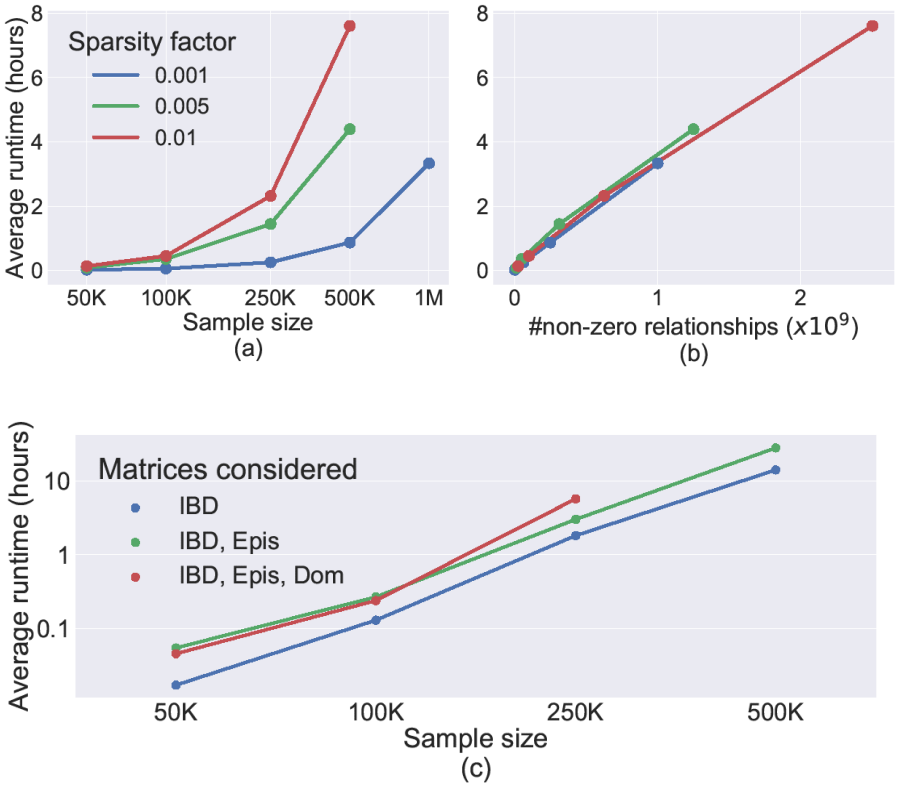

Figure 2: Analysis of SciLMM computation time. **(a)** Computation time required to compute an IBD matrix from pedigree data under different sparsity factors as a function of sample size. **(b)** Computation time required to compute an IBD matrix from pedigree data as a function of the number of nonzero 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.

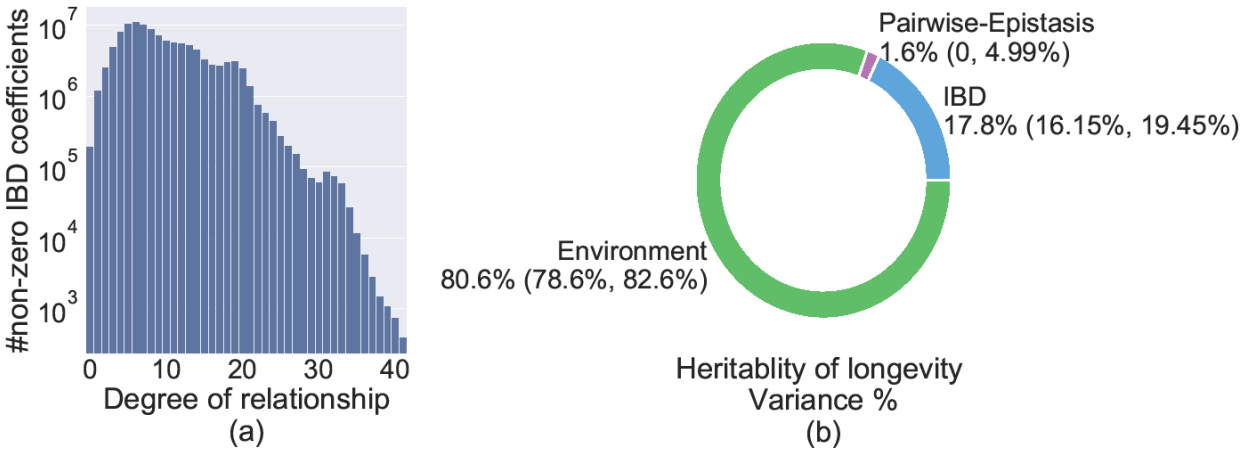

Figure 3: Results of analysis of a real pedigree with 441,000 individuals. **(a)** A histogram of genetic similarity across 441,000 individuals, using only the closest relationship between every pair of individuals. The data set includes approximately 9.7 million pairs of individuals whose least common ancestor lived at least 10 generations earlier. **(b)** The estimated fraction of longevity variance attributed to different variance components (and their 95% CI).

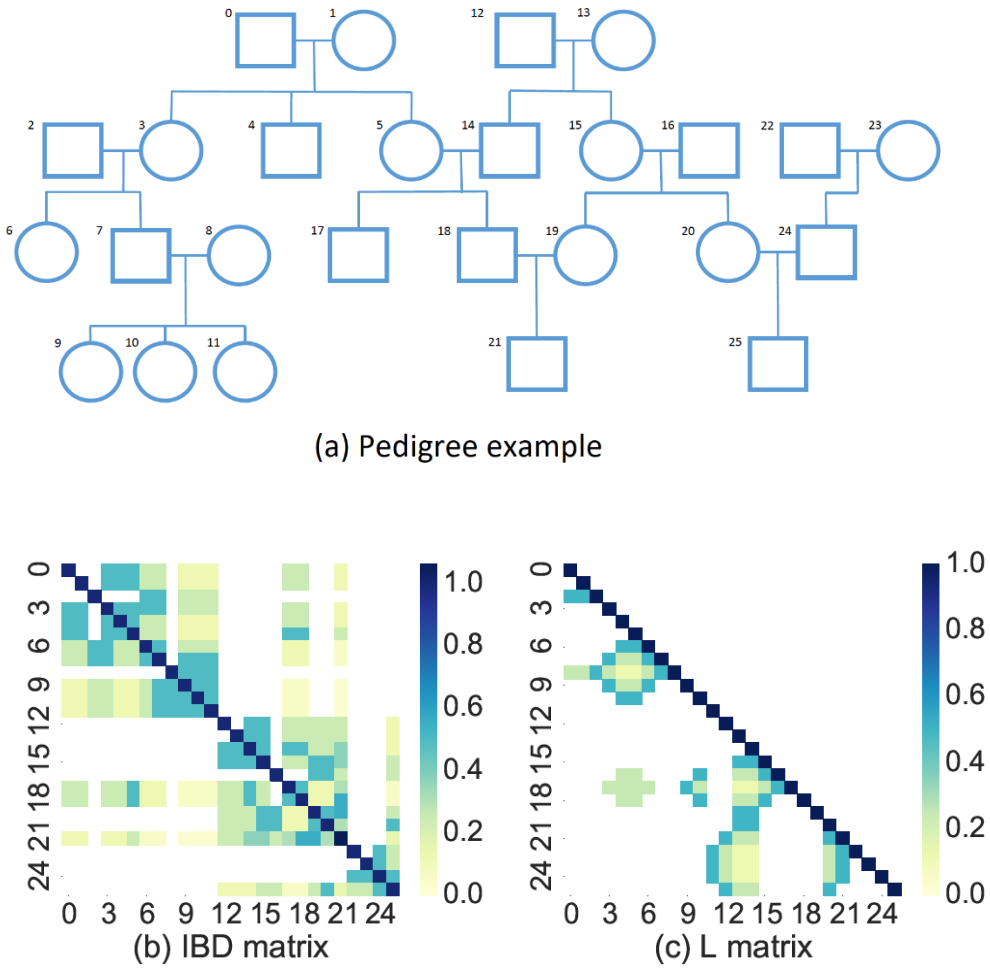

Figure 4: An example of a pedigree, its IBD matrix and its decomposition. (**a**) An example pedigree with 26 individuals. (**b**) A heat-map representing the IBD matrix, where zero elements are white to emphasize sparsity. (**c**) A heat-map representing the lower Cholesky factorization of the IBD matrix (i.e. the matrix L in the factorization $\boldsymbol{A} = \boldsymbol{LDL}^T$, 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.