Estimating variance components in 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

Short title: Estimating variance components in population scale family trees

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 investigate the sociological and epidemiological history of human populations in scales much larger than previously possible. Linear mixed models (LMMs) are routinely used to analyze extremely large animal and plant pedigrees for the purposes of selective breeding. However, LMMs have not been previously applied to analyze population-scale human family trees. Here, we present **S**parse **C**holesky factor**I**zation LMM (Sci-LMM), a modeling framework for studying population-scale family trees that combines techniques from the animal and plant breeding literature and from human genetics literature. The proposed framework can construct a matrix of relationships between trillions of pairs of individuals and fit the corresponding LMM in several hours. We demonstrate the capabilities of Sci-LMM via simulation studies and by estimating the heritability of longevity and of reproductive fitness (quantified via number of children) in a large pedigree spanning millions of individuals and over five centuries of human history. Sci-LMM provides a unified framework for investigating the epidemiological history of human populations via genealogical records.

Author Summary

The advent of online genealogy services allows the assembly of population-scale family trees, spanning millions of individuals and centuries of human history. Such datasets enable answering genetic epidemiology questions on unprecedented scales. Here we present Sci-LMM, a pedigree analysis framework that combines techniques from animal and plant breeding research and from human genetics research for large-scale pedigree analysis. We apply Sci-LMM to analyze population-scale human genealogical records, spanning trillions of relationships. We have made both Sci-LMM and an anonymized dataset of millions of individuals freely available to download, making the analysis of population-scale human family trees widely accessible to the research community. Together, these resources allow researchers to investigate genetic and epidemiological questions on an unprecedented scale.

Introduction

Genealogical records can reflect social and cultural structures, and record the flow of genetic material throughout history. In recent years, very large pedigree records have come into existence, owing to collaborative digitization of large genealogical records[1,2] and to digitization of large cohorts collected by healthcare providers, spanning up to millions of individuals[3–7]. Such population-scale pedigrees allow investigating the sociological and epidemiological history of human populations on a scale that is orders of magnitude larger than existing studies.

Traditional human pedigree studies collect a large number of independent families which are analyzed separately and then meta-analyzed. However, this approach is not suitable for population-scale pedigrees, because such pedigrees cannot be decomposed into mutually exclusive families[1]. Hence, the analysis of such pedigrees requires modeling complex covariance structures between trillions of pairs of individuals.

Pedigree studies often employ LMMs to decompose the phenotypic variation among individuals into variance components such as genetic effects and shared environment[8]. LMMs have been the statistical backbone of animal and plant breeding programs for almost six decades[9], and have been continuously developed over the years[10–21]. LMMs are routinely used nowadays to analyze pedigrees of millions of animals and plants[13,22], hundreds of thousands of which are often genotyped (e.g. [23,24]).

In recent years, LMMs and their extensions have also gained considerable popularity in human genetics studies for the purposes of estimating heritability[25–31] and genetic correlation[32–36], predicting phenotypes[37–40] and modeling sample relatedness[41–45]. Unlike classical animal and plant studies, human studies typically do not include pedigree data, but instead measure genetic relatedness via dense genotyping of single nucleotide polymorphisms (SNPs).

In recent years, animal and human studies used different techniques to scale LMMs to datasets with millions of individuals. Animal studies typically fit large-scale LMMs via restricted maximum likelihood (REML)[12], by exploiting the sparsity of pedigree data. Specifically, 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. Such sparse matrices can be stored and analyzed efficiently with suitable numerical techniques[21,46,47].

Human genotyping studies do not give rise to sparse data structures. Instead, human studies have managed to scale LMMs to large datasets via two approaches. The first approach applies REML, either via supercomputers with thousands of CPUs and terabytes of memory[45], or by approximating the restricted likelihood gradient via Monte-Carlo techniques[27]. However, the latter technique is only suitable for specific types of covariance matrices whose decomposition is known beforehand.

The second approach to scale LMMs uses the method of moments rather than REML, by solving a set of second moment matching equations[48–52]. Such approaches have become increasingly popular recently[29–34,53–57] owing to their computational tractability and their compatibility with privacy-preserving summary statistics[58]. Although moment

estimators are less statistically efficient compared to REML estimators, they have several advantages: the loss of efficiency has been found to often be small[55]; they are more robust to modeling violation because they make fewer distributional assumptions; and they are more flexible, which enables applying technique to limit confounding factors such as assortative mating (Methods). Moment estimators have also recently been explored in animal breeding studies[59–61] and were found to be faster than REML while providing similar accuracy, but they have not been widely adopted in animal studies to date.

Here we propose Sci-LMM, a statistical framework for analyzing population-size pedigrees that combines techniques from animal and human genetic studies. Sci-LMM uses sparse data structures as is common in animal studies, and supports both moment and REML estimators. The moment estimator is based on a common technique called Haseman-Elston (HE) regression[62,63] (Methods). Sci-LMM scales HE regression to population-sized pedigrees via sparse matrix tools[64]. The REML estimator combines a direct sparse REML solver [46] with Monte-Carlo gradient approximation[65]. Importantly, existing packages for pedigree-based REML[66–70] cannot handle the analyses performed in this paper because they require the inverse of the epistatic interactions matrix[46,71,72], which is extremely difficult to compute in large pedigrees[73]. Hence, Sci-LMM provides a comprehensive solution for LMM-based pedigree analysis.

To demonstrate the capabilities of Sci-LMM, we carry out an extensive analysis of simulated data with millions of individuals, which we complete within a few hours. We additionally estimate the heritability of longevity and of reproductive fitness (quantified via number of children), using a large cohort spanning millions of genealogical records and several centuries of human history. We estimate that both traits have a substantial heritable component, with an estimated 22.1% heritability for longevity and 34.4% for reproductive fitness. Sci-LMM enables analysis of large pedigree records that was not previously possible.

Results

Overview of Proposed Framework

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

Sci-LMM is a flexible framework to analyze population scale pedigrees with various types of covariates and variance components. Sci-LMM 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 and 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. Sci-LMM estimates these parameters using either REML or HE regression (Methods).

Following the classical genetic model originally proposed by Fisher[74] and developed by Cockerham[71] and Kempthorne[72], Sci-LMM supports the efficient computation of an

additive IBD-based genetic component, a dominancy, and an epistasic covariance matrix (Methods). These matrices can be constructed rapidly via dynamic programming[75–77].

In addition to covariance matrices, Sci-LMM also includes adjustments using principal components (PCs) with fixed effects, which can be computed efficiently using sparse matrix routines[78]. These adjustments can capture major linear sources of variation in a dataset that may not be captured well via the entire covariance matrix, such as founder effects. However, we note that unlike PCs computed from genetic data, PCs computed from IBD are not guaranteed to capture population structure[79].

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 included 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 consanguinity, 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 binary 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 settings, Sci-LMM yielded empirically unbiased estimates of the variance components, using both REML and HE regression. As expected, estimation accuracy increased with sample size, 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 (because the phenotype was standardized to have unit variance).

A comparison of the REML and the HE results shows that that HE was slightly more accurate in the presence of <100,000 individuals (Figure 1a-c), and REML was slightly more accurate otherwise. These results possibly indicate that REML convergence is difficult in the presence of sparse covariance matrices with limited sample sizes. We also found that estimation accuracy was anti-correlated with relatedness sparsity, indicating that the estimators efficiently exploit the information found in non-zero covariance entries (Figure 1d-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 scales linearly with the number of non-zero entries in the matrix (Figure 2a-b).

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 REML and HE estimation runtime. REML estimation for samples with 500,000 individuals (representing 250 million covariance entries) required less than 24 hours (Figure 2c), whereas HE estimation required 16 seconds (Figure 2d). Overall, our results demonstrate that the Sci-LMM framework is scalable to extremely large pedigrees.

Finally, we tried comparing Sci-LMM with pedigree-based REML software packages[66–69]. We could not invoke any of these packages with an epistatic covariance matrix because they require its inverse, which is not sufficiently sparse[73]. We verified this by trying to invert the matrix analytically via the software package nadiv[80] and numerically via sparse matrix libraries[81] and via the matrix inversion facilities of the software package WOMBAT[68], all of which ran out of memory on a 256GB machine. We additionally tried running the analysis via the *fitNullModel* function of the GENESIS package[82] and the *lmekin* function of the coxme package[83], both of which can work with sparse covariance matrices. However, both packages could not finish the analysis in 4 days, presumably because they do not use the approximate gradient approximation techniques of Sci-LMM.

When excluding the analysis to only an additive covariance matrix, WOMBAT was much faster than Sci-LMM, completing an analysis of 250,000 individuals in several minutes (Supplementary Table 1). This is because REML estimation with only an additive matrix can be extremely efficient when combining the analytical form of the inverse matrix[75–77] with a mixed model equations solver[46].

Estimating the heritability of longevity and reproductive fitness

We used Sci-LMM to estimate the heritability of longevity and reproductive fitness, based on large-scale pedigree records obtained from the Geni genealogical website[1] (see Web Resources). An initial description of the longevity analysis was reported in [1], but here we substantially refine and extend this analysis. We applied stringent quality control to minimize deaths due to non-natural reasons such as wars or natural disasters, by excluding pairs of individuals who died within 10 days of each other or within periods with significantly elevated death rates[1]. This filtering yielded 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 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 3a). Sci-LMM constructed this matrix in 10 hours.

Next, we used Sci-LMM to estimate the heritability of longevity with covariates encoding gender, year of birth (yob), yob^2, yob^3, and the top 10 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 equivalent to the identity matrix (Supplementary Material). The REML estimates were: IBD: 22.1% (s.e. 0.8%); pair-wise epistasis: 0.001% (s.e. 1.7%); environmental effects: 74.4% (s.e. 1.0%) (Figure 3b). The HE estimate for IBD was 24.3% (s.e. 0.5%) when excluding the epistatic interactions matrix (HE results with epistatic interactions were inconclusive due to large standard errors). We also tried to limit confounding due to shared environmental effects[2,84] by excluding ~399,000 pairs (~0.15%) of individuals with a shared household (spouses or parent-child pairs) from the analysis without excluding the individuals themselves, using HE (Methods). This led to a heritability estimate of 26.3% (s.e. 0.9%), indicating that shared household effects are unlikely to up-bias our estimates. Overall, our results suggest that the heritability of longevity is upper bounded by ~26%. However, the true heritability may be lower since our estimates may be confounded by other environmental factors[2,84] (see Discussion).

We also estimated the heritability of reproductive fitness, quantified by number of children. To limit confounding due to non-genetic factors, we applied stringent filtering of individuals. We removed individuals with less than two children records, individuals with a shared household (spouses and children) and individuals who are first- or second-degree relatives with another individual in the data set. The filtered dataset included ~45,000 individuals. We used the same covariates as before, applied a Box-cox transformation to induce normality for number of children, and excluded epistatic interactions from this reduced dataset, because they led to large standard errors. The REML and HE estimated heritability of reproductive fitness were 28.4% (s.e. 0.5%) and 34.4% (s.e. 1.2%), respectively. These results indicate a substantial genetic component for reproductive fitness, in line with previous work[85]. As before, we caution that this estimate may be upper-biased due to confounding by non-genetic factors[2,84] (see Discussion).

Finally, we performed a series of experiments to examine the robustness of our analysis to potential confounding factors. These included considering only individuals born after 1800, excluding first- and second-degree relatives from the analysis, excluding individuals that share or likely shared a household in the past, and various combinations thereof. The results indicate that our analysis is not upper biased due to inclusion of these potential confounding factors (Supplementary Table 2).

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, and methodologies for LMM analysis with random effects described by these matrices. Taken together, the proposed solution enables an end-to-end analysis of population scale human family trees.

In this work we focused on partitioning phenotypic variance into additive genetics, epistasis and dominance. However, the LMM framework is flexible and can be extended in various directions. For example, sparse LMMs are often used to model transmissible phenomena [86–

91], which enables combining pedigree-based and geography-based covariance structures. Both Sci-LMM and the data studied in this paper are freely available for download, which makes the analysis of population-scale human family trees widely accessible to the research community. Combined, these resources allow researchers to investigate genetic and epidemiological questions on unprecedented scales.

We evaluated two methods for variance components estimation: REML and HE regression. REML is more accurate than HE and provides a likelihood-based solution, which can be used for model comparison and hypothesis testing. HE estimates are less accurate but can be more robust to modeling violation. Importantly, HE regression can mitigate confounding due to environmental factors by zeroing selected entries in the covariance matrices, which may be especially suitable for studying human genealogical records (Methods). 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 with a sparsity factor of 0.001. Our recommendation is to use REML when it is feasible and all model assumptions hold, and HE regression otherwise. We note that REML estimation can be substantially faster when not fitting epistatic interactions by using a mixed model equations approach [46], which is implemented in several software packages[66–69].

Our work demonstrates the technical feasibility of studying population scale human family trees. However, the analysis of human genealogical records is challenging due to imperfect data and the difficulty of controlling for confounding factors. Potential issues include non-paternity, cryptic relatedness, missing or false genealogical records, genetic nurture[92,93], environmental bias[84,94], assortative mating[2] and correlation between additive and epistatic effects[17,18]. As such, our estimates should be considered as a first order approximation, and our heritability estimates are likely upper biased due to confounding. We expect that several recently proposed techniques to address these issues (e.g. [2,93,94]) could be integrated into the Sci-LMM framework in the future.

In this work we focused on analyzing large pedigree records with no measured genotypes. In recent years, the advent of biobank-sized datasets allows analyzing population-scale genotyped cohorts. The two study types are complementary because biobanks cannot be used to investigate longevity, traits with a late age of onset, or epidemiological and sociological questions on historical scales. We anticipate that cohorts combining both types of data will become increasingly common. For example, we and other online genealogy platforms allow users to upload their genetic information and link it with their genealogical profile. Such combined datasets have been extensively explored in the animal breeding literature[19,21,95–99]. However, privacy and logistical concerns limit public access to human genetic data, necessitating methods based on summary statistics[58]. Thus, approaches for analysis of such combined datasets will combine state of the art techniques from the animal breeding and the GWAS literature, and remain a potential avenue for future work.

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 [100,101], 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 [100], 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 [75] (Figure 4a-c). The matrices $\boldsymbol{L}$ and $\boldsymbol{H}$ can be computed efficiently via dynamic programming [77] with sparse matrix routines [81].

Computing a Dominance Relationship Matrix

Dominancy represents the genetic variance due to co-ancestry of two alleles, and can be approximated by $\frac{1}{4}\left(\boldsymbol{A}_{f_i,f_j} \cdot \boldsymbol{A}_{m_i,m_j} + \boldsymbol{A}_{f_i,m_j} \cdot \boldsymbol{A}_{m_i,f_j}\right)$, where $A_{k,l}$ is the IBD coefficient of individuals k, l, and f_k, m_k are the parents of individual k [10,102]. A necessary condition for nonzero dominancy entry is a nonzero IBD relationship, which enables 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 [72]. 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. However, individuals with missing data may still be required for IBD computation. For example, consider a pedigree of two siblings with phenotypic data, and two parents and one uncle without phenotypic data. The parents are important for the IBD computation of the siblings, but the uncle is non-informative.

Sci-LMM applies pedigree-pruning techniques to remove non-informative individuals, similarly to other REML packages for pedigree analysis[67–70]. 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. This algorithm reduces the matrix construction time by several hours.

Haseman-Elston Regression

HE regression estimates variance components via the method of moments, using:

$$\operatorname{cov}\left(y_i - \boldsymbol{C}_i\boldsymbol{\beta}, y_j - \boldsymbol{C}_j\boldsymbol{\beta}\right) = \sum_{k=1}^{d} \sigma_k^2 M_{ij}^k.$$

This allows fitting the variance components σ_k^2 via least square minimization, which can be solved via sparse matrix routines. Typically, the fixed effects $\boldsymbol{\beta}$ are first estimated without considering the covariance matrices (which yields a consistent estimator under mild regularity conditions[103]), and are then incorporated into the moment estimator above. The sampling variance can be efficiently approximated via Month-Carlo approximations, similar to REML gradients (Supplementary Material).

HE regression provides a simple technique for excluding specific pairs of individuals (e.g. spouses) from the analysis without excluding the individuals themselves. This can be useful when trying to limit confounding due to factors such as assortative mating. Excluded pairs can be omitted from the analysis by zeroing the covariance matrix entries of corresponding pairs. Importantly, this technique cannot be used in REML, since the resulting covariance matrices may not be positive definite. We note that another potential approach to capture environmental risk factors is by including shared effects with a suitable incidence matrix[10], but this approach requires additional assumptions and has not been used here.

Restricted Maximum Likelihood Estimator (REML)

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

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

where $\boldsymbol{\Sigma}^\dagger$ is the Moore-Penrose pseudo-inverse of the covariance matrix $\sum_{k=1}^{d} \sigma_k^2 M^k$ after projecting out the covariates (Supplementary Material), and $|\boldsymbol{\Sigma}^\dagger|_+$ is the product of its non-zero eigenvalues[104]. This expression, as well as its partial derivative according to each variance component, can be computed efficiently via sparse matrix routines. Exact gradient estimation is more computationally challenging, but can be approximated efficiently via Monte Carlo techniques described in[27,65]. Since REML is sensitive to initial values, we obtain initial estimates via HE regression.

Computing environment

All experiments were conducted using a Linux workstation with a 24-cores 2GHz Xeon E5 processor and 256Gb of RAM.

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.).

References

1. Kaplanis J, Gordon A, Shor T, Weissbrod O, Geiger D, Wahl M, et al. Quantitative analysis of population-scale family trees with millions of relatives. Science. 2018;360: 171–175.

2. Ruby JG, Wright KM, Rand KA, Kermany A, Noto K, Curtis D, et al. Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating. Genetics. 2018;210: 1109–1124.

3. Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA, Zink F, et al. Sequence variants from whole genome sequencing a large group of Icelanders. Sci Data. 2015;2: 150011.

4. Huang X, Elston RC, Rosa GJ, Mayer J, Ye Z, Kitchner T, et al. Applying family analyses to electronic health records to facilitate genetic research. Bioinformatics. 2018;34: 635–642.

5. Polubriaginof FCG, Vanguri R, Quinnies K, Belbin GM, Yahi A, Salmasian H, et al. Disease Heritability Inferred from Familial Relationships Reported in Medical Records. Cell. 2018;173: 1692-1704.e11.

6. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. Am J Med Genet B Neuropsychiatr Genet. 2018;177: 601–612.

7. Nelson D, Moreau C, de Vriendt M, Zeng Y, Preuss C, Vézina H, et al. Inferring Transmission Histories of Rare Alleles in Population-Scale Genealogies. Am J Hum Genet. 2018;103: 893–906.

8. Kruuk LEB, Hadfield JD. How to separate genetic and environmental causes of similarity between relatives: Separating genetic and environmental effects. J Evol Biol. 2007;20: 1890–1903.

9. Henderson CR, Kempthorne O, Searle SR, Von Krosigk C. The estimation of environmental and genetic trends from records subject to culling. Biometrics. 1959;15: 192–218.

10. Lynch M, Walsh B. Genetics and Analysis of Quantitative Traits. Sunderland, Massachusetts, USA: Sinauer Associates; 1998.

11. Gianola D. Statistics in Animal Breeding. J Am Stat Assoc. 2000;95: 296–299.

12. Hofer A. Variance component estimation in animal breeding: a review. J Anim Breed Genet. 1998;115: 247–265.

13. Kruuk LEB. Estimating genetic parameters in natural populations using the “animal model.” Philos Trans R Soc B Biol Sci. 2004;359: 873–890.

14. Thompson R, Mäntysaari E. Prospects for statistical methods in animal breeding. J Ind Soc Agric Stat. 2004;57: 15–25.

15. Thompson R, Brotherstone S, White IMS. Estimation of quantitative genetic parameters. Philos Trans R Soc B Biol Sci. 2005;360: 1469–1477.

16. Thompson R. Estimation of quantitative genetic parameters. Proc R Soc B Biol Sci. 2008;275: 679–686.

17. Hill WG. Understanding and using quantitative genetic variation. Phil Trans R Soc B. 2010;365: 73–85.

18. Gianola D, Rosa GJM. One Hundred Years of Statistical Developments in Animal Breeding. Annu Rev Anim Biosci. 2015;3: 19–56.

19. Xavier A, Muir WM, Craig B, Rainey KM. Walking through the statistical black boxes of plant breeding. Theor Appl Genet. 2016;129: 1933–1949.

20. Manfredi E, Tusell L, Vitezica ZG. Prediction of complex traits: Conciliating genetics and statistics. J Anim Breed Genet. 2017;134: 178–183.

21. Misztal I, Legarra A. Invited review: efficient computation strategies in genomic selection. animal. 2017;11: 731–736.

22. Silva MVB, dos Santos DJA, Boison SA, Utsunomiya ATH, Carmo AS, Sonstegard TS, et al. The development of genomics applied to dairy breeding. Livest Sci. 2014;166: 66–75.

23. Fernando RL, Cheng H, Golden BL, Garrick DJ. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genet Sel Evol. 2016;48.

24. Mäntysaari EA, Evans RD, Strandén I. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals1. J Anim Sci. 2017;95: 4728–4737.

25. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42: 565–569.

26. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88: 76–82.

27. Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet. 2015;47: 1385–1392.

28. Lee SH, van der Werf JHJ. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information: Table 1. Bioinformatics. 2016;32: 1420–1422.

29. Golan D, Lander ES, Rosset S. Measuring missing heritability: Inferring the contribution of common variants. Proc Natl Acad Sci. 2014;111: E5272-81.

30. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47: 291–295.

31. Ge T, Chen C-Y, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. Domingue BW, editor. PLOS Genet. 2017;13: e1006711.

32. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47: 1236–1241.

33. Lu Q, Li B, Ou D, Erlendsdottir M, Powles RL, Jiang T, et al. A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. Am J Hum Genet. 2017;101: 939–964.

34. Weissbrod O, Flint J, Rosset S. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. Am J Hum Genet. 2018;103: 89–99.

35. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28: 2540–2542.

36. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet. 2013;45: 984–994.

37. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9: e1003264.

38. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24: 1550–7.

39. Golan D, Rosset S. Effective genetic-risk prediction using mixed models. Am J Hum Genet. 2014;95: 383–93.

40. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97: 576–592.

41. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11: 459–463.

42. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47: 284–290.

43. Loh P-R, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nat Genet. 2018;50: 906–908.

44. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50: 1335–1341.

45. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018;50: 1593–1599.

46. Lee S, van der Werf J. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genet Sel Evol. 2006;38: 1–19.

47. Masuda Y, Aguilar I, Tsuruta S, Misztal I. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements1,2. J Anim Sci. 2015;93: 4670–4674.

48. Rao CR. Estimation of Heteroscedastic Variances in Linear Models. J Am Stat Assoc. 1970;65: 161.

49. Rao CR. Estimation of variance and covariance components—MINQUE theory. J Multivar Anal. 1971;1: 257–275.

50. Rao CR. Minimum variance quadratic unbiased estimation of variance components. J Multivar Anal. 1971;1: 445–456.

51. Rao CR. Estimation of Variance and Covariance Components in Linear Models. J Am Stat Assoc. 1972;67: 112–115.

52. LaMotte LR. Quadratic Estimation of Variance Components. Biometrics. 1973;29: 311.

53. Bulik-Sullivan B. Relationship between LD Score and Haseman-Elston Regression. bioRxiv. 2015; 018283.

54. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47: 1228–1235.

55. Zhou X. A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann Appl Stat. 2017;11: 2027–2051.

56. Bonnet A. Heritability estimation in case-control studies. Electron J Stat. 2018;12: 1662–1716.

57. Wu Y, Sankararaman S. A scalable estimator of SNP heritability for biobank-scale data. Bioinformatics. 2018;34: i187–i194.

58. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017;18: 117–127.

59. Hu Z, Yang R-C. Marker-Based Estimation of Genetic Parameters in Genomics. Cai X, editor. PLoS ONE. 2014;9: e102715.

60. Liu H, Chen G-B. A fast genomic selection approach for large genomic data. Theor Appl Genet. 2017;130: 1277–1284.

61. Liu H, Chen G-B. A new genomic prediction method with additive-dominance effects in the least-squares framework. Heredity. 2018;121: 196–204.

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

63. Chen G-B. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman–Elston regression. Front Genet. 2014;5: 107.

64. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for Python [Internet]. 2001. Available: http://www.scipy.org/

65. Matilainen K, Mäntysaari EA, Lidauer MH, Strandén I, Thompson R. Employing a Monte Carlo Algorithm in Newton-Type Methods for Restricted Maximum Likelihood Estimation of Genetic Parameters. Wu R, editor. PLoS ONE. 2013;8: e80821.

66. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee D. BLUPF90 and related programs (BGF90). Proceedings of the 7th world congress on genetics applied to livestock production. 2002. pp. 743–744.

67. Madsen P, Sørensen P, Su G, Damgaard LH, Thomsen H, Labouriau R, et al. DMU-a package for analyzing multivariate mixed models. 8th World Congress on Genetics Applied to Livestock Production. Belo Horizonte; 2006.

68. Meyer K. WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). J Zhejiang Univ Sci B. 2007;8: 815–821.

69. Gilmour A, Gogel B, Cullis B, Welham Sj, Thompson R. ASReml user guide release 4.1 structural specification. Hemel Hempstead VSN Int Ltd. 2015;

70. Hadfield JD. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. J Stat Softw. 2010;33: 1–22.

71. Cockerham C. An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present. Genetics. 1954;39: 859—882.

72. Kempthorne O. The correlation between relatives in a random mating population. Proc R Soc Lond B Biol Sci. 1954;143: 103–113.

73. VanRaden PM, Hoeschele I. Rapid Inversion of Additive by Additive Relationship Matrices by Including Sire-Dam Combination Effects. J Dairy Sci. 1991;74: 570–579.

74. Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans R Soc Edinb. 1919;52: 399–433.

75. Henderson CR. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics. 1976;32: 69–83.

76. Quaas RL. Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix. Biometrics. 1977;32: 949.

77. Meuwissen T, Luo Z. Computing inbreeding coefficients in large populations. Genet Sel Evol GSE. 1991;24: 305–313.

78. Sorensen DC. Implicitly Restarted Arnoldi/Lanczos Methods for Large Scale Eigenvalue Calculations. Parallel Numerical Algorithms. Dordrecht: Springer Netherlands; 1997. pp. 119–165.

79. Zhu X, Li S, Cooper RS, Elston RC. A Unified Association Analysis Approach for Family and Unrelated Samples Correcting for Stratification. Am J Hum Genet. 2008;82: 352–365.

80. Wolak ME. nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. Methods Ecol Evol. 2012;3: 792–796.

81. Chen Y, Davis TA, Hager WW, Rajamanickam S. Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. ACM Trans Math Softw TOMS. 2008;35: 22.

82. Conomos MP, Gogarten SM, Brown L, Chen H, Rice K, Sofer T, et al. GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness [Internet]. 2018. Available: https://rdrr.io/bioc/GENESIS/

83. Therneau TM. coxme: Mixed Effects Cox Models [Internet]. 2018. Available: https://CRAN.R-project.org/package=coxme

84. Feldman MW, Ramachandran S. Missing compared to what? Revisiting heritability, genes and culture. Philos Trans R Soc B Biol Sci. 2018;373: 20170064.

85. Kosova G, Abney M, Ober C. Heritability of reproductive fitness traits in a human population. Proc Natl Acad Sci. 2010;107: 1772–1778.

86. Buhmann MD. A new class of radial basis functions with compact support. Math Comput. 2000;70: 307–319.

87. Gneiting T. Correlation functions for atmospheric data analysis. Q J R Meteorol Soc. 1999;125: 2449–2464.

88. Gaspari G, Cohn SE. Construction of correlation functions in two and three dimensions. Q J R Meteorol Soc. 1999;125: 723–757.

89. Gneiting T. Compactly supported correlation functions. J Multivar Anal. 2002;83: 493–508.

90. Sansò F, Schuh W-D. Finite covariance functions. Bull Géod. 1987;61: 331–347.

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

92. Bijma P. Estimating maternal genetic effects in livestock. J Anim Sci. 2006;84: 800–806.

93. Kong A, Thorleifsson G, Frigge ML, Vilhjalmsson BJ, Young AI, Thorgeirsson TE, et al. The nature of nurture: Effects of parental genotypes. Science. 2018;359: 424–428.

94. Young AI, Frigge ML, Gudbjartsson DF, Thorleifsson G, Bjornsdottir G, Sulem P, et al. Relatedness disequilibrium regression estimates heritability without environmental bias. Nat Genet. 2018;50: 1304–1310.

95. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92: 4656–4663.

96. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93: 743–752.

97. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42: 2.

98. Legarra A, Christensen OF, Aguilar I, Misztal I. Single Step, a general approach for genomic selection. Livest Sci. 2014;166: 54–65.

99. Meuwissen T, Hayes B, Goddard M. Genomic selection: A paradigm shift in animal breeding. Anim Front. 2016;6: 6–14.

100. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015;16: 33–44.

101. Wright S. Coefficients of inbreeding and relationship. Am Nat. 1922;56: 330–338.

102. Henderson C. Best linear unbiased prediction of nonadditive genetic merits. J Anim Sci. 1985;60: 111–117.

103. McCulloch CE, Searle SR, Neuhaus JM. Generalized, Linear, and Mixed Models. 2nd ed. Wiley Series in Probability and Statistics; 2008.

104. Lippert C, Xiang J, Horta D, Widmer C, Kadie C, Heckerman D, et al. Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. Bioinformatics. 2014;30: 3206–14.

Figures

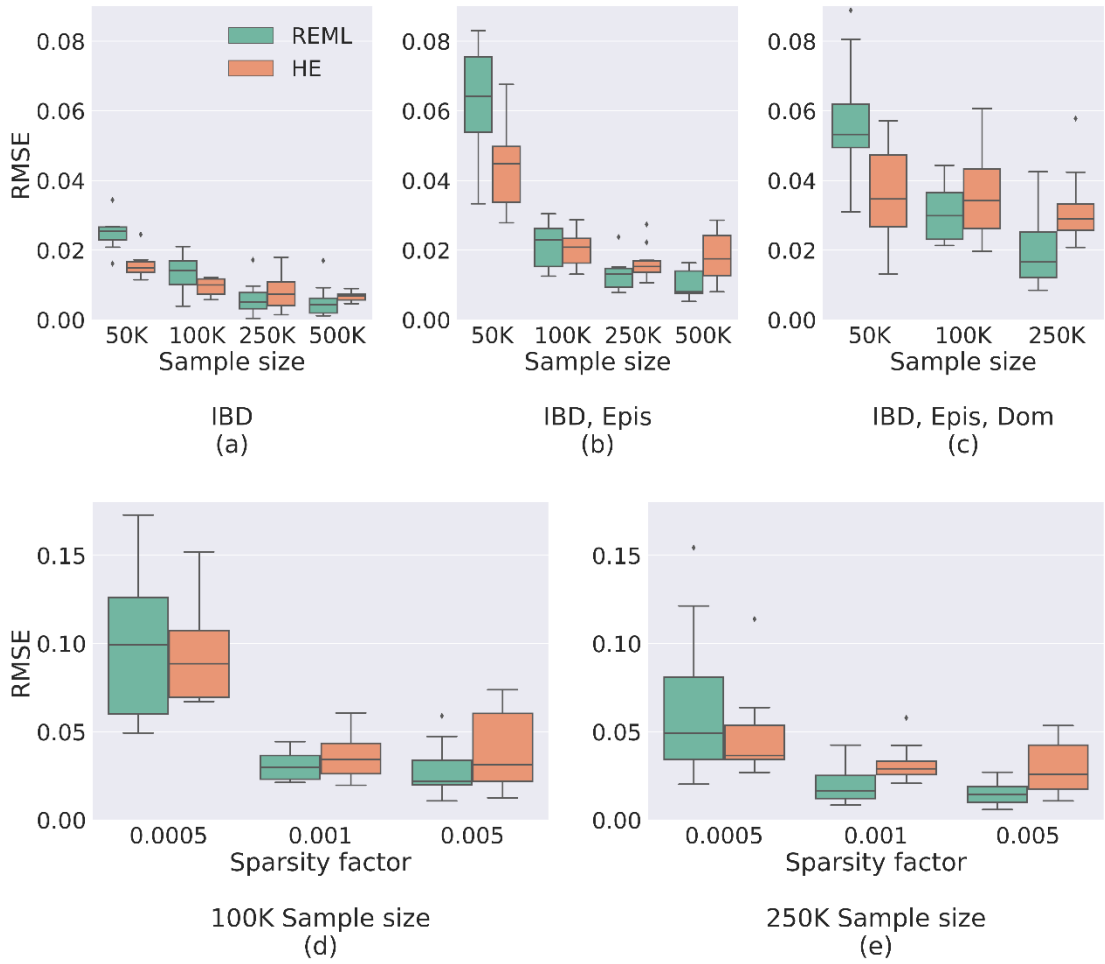

Figure 1: Evaluating the estimation accuracy of Sci-LMM. (**a-c**) Box plots comparing REML and HE estimation accuracy (RMSE) across simulated datasets (each box represents 10 experiments), under varying sample sizes, using **(a)** only IBD, **(b)** IBD and epistasis, or **(c)** IBD, epistasis and dominance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

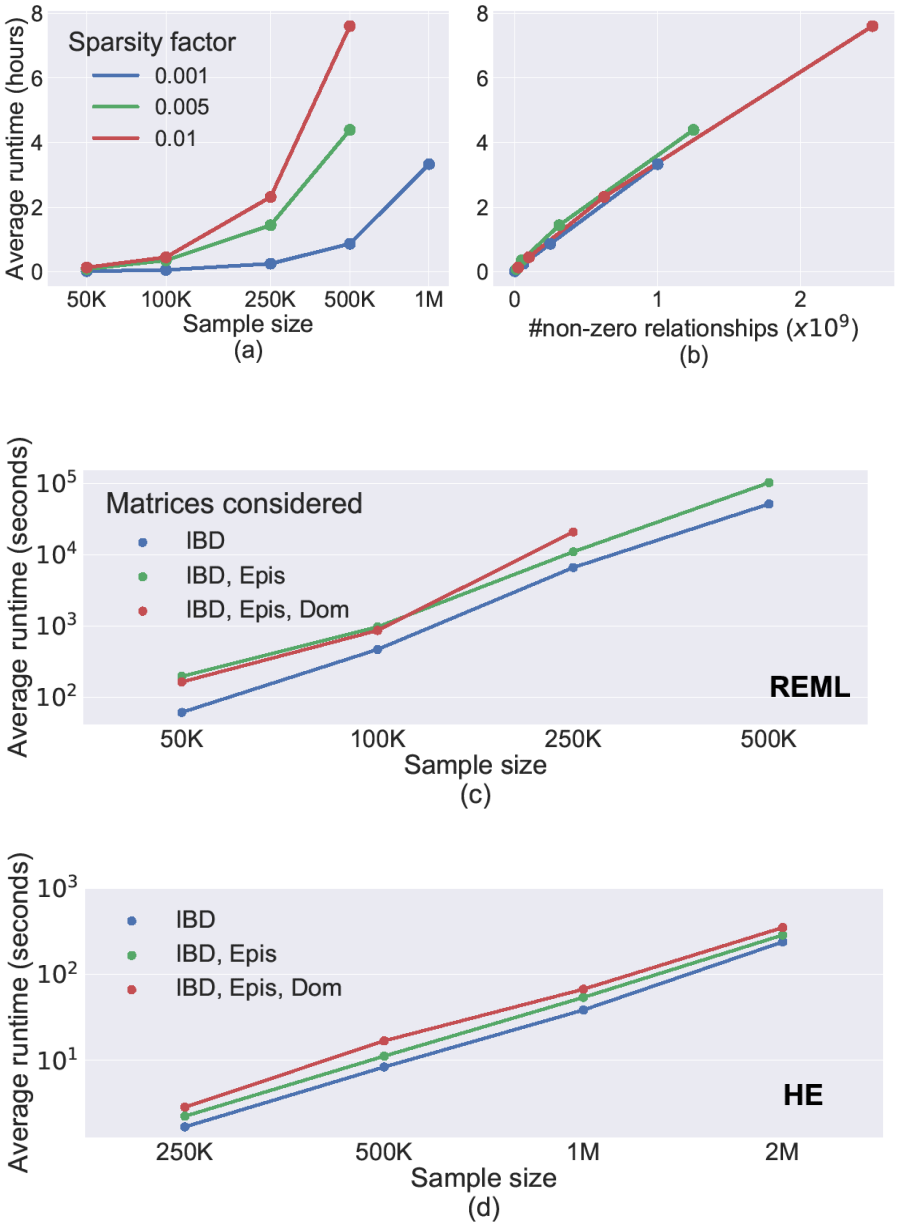

Figure 2: Analysis of Sci-LMM computation time. **(a)** Computation time required to compute an IBD matrix from pedigree data under different sparsity factors as a function of sample size. **(b)** Computation time required to compute an IBD matrix from pedigree data as a function of the number of nonzero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

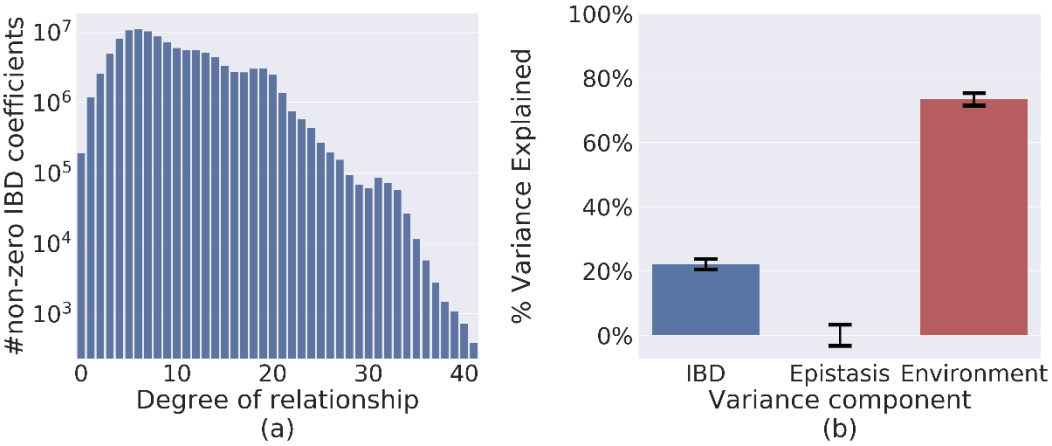

Figure 3: Results of analysis of a real pedigree with 441,000 individuals. **(a)** A histogram of genetic similarity across 441,000 individuals, using only the closest relationship between every pair of individuals. The dataset includes approximately 9.7 million pairs of individuals whose least common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

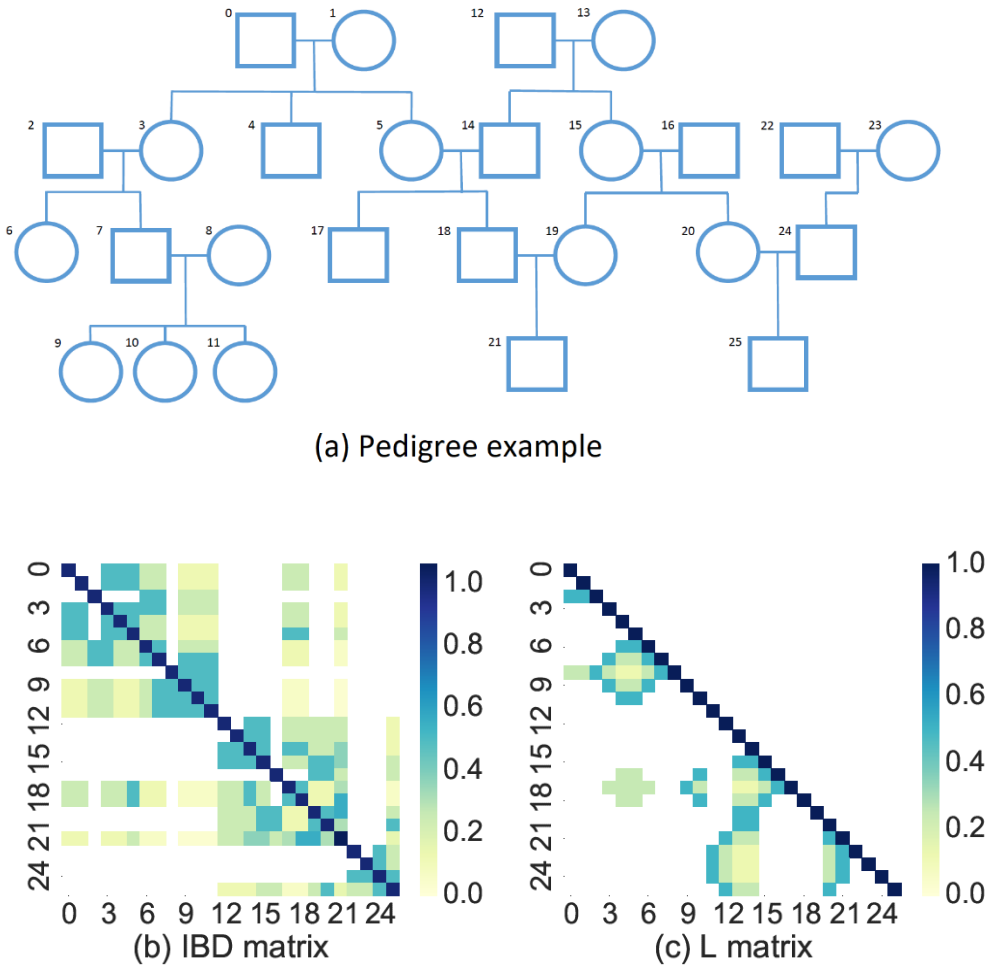

Figure 4: A demonstration of the Sci-LMM IBD matrix construction algorithm. (**a**) An example pedigree with 26 individuals. (**b**) A heat-map representing the IBD matrix, where zero elements are white to emphasize sparsity. (**c**) A heat-map representing the lower Cholesky factorization of the IBD 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.

Supporting Information Legends

Supplementary Material: Detailed derivations and algorithms for the methodologies described in the main text.