PLOS GENETICS

RESEARCH ARTICLE

Estimating variance components in population scale family trees

Tal Shor[1,2]*, **Iris Kalka**[3,4], **Dan Geiger**[1], **Yaniv Erlich**[2,5,6], **Omer Weissbrod**[1,7]*

1 Computer Science Department, Technion—Israel Institute of Technology, Haifa, Israel, **2** MyHeritage Ltd., Or Yehuda, Israel, **3** Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, **4** Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel, **5** The New York Genome Center, New York, NY, United States of America, **6** Department of Computer Science, Fu School of Engineering, Columbia University, NY, United States of America, **7** Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America

* tal.shor@myheritage.com (TS); oweissbrod@hsph.harvard.edu (OW)

Citation: Shor T, Kalka I, Geiger D, Erlich Y, Weissbrod O (2019) Estimating variance components in population scale family trees. PLoS Genet 15(5): e1008124. https://doi.org/10.1371/journal.pgen.1008124

Editor: Peter M. Visscher, The University of Queensland, AUSTRALIA

Received: September 24, 2018

Accepted: April 3, 2019

Published: May 9, 2019

Data Availability Statement: The data studied in this paper is freely available to download from the following URL: http://familinx.org/. The Sci-LMM software code is available from the following URL: https://github.com/TalShor/SciLMM.

Funding: This study was supported by a generous gift from Andria and Paul Heafy (YE) and the Burroughs Wellcome Fund Career Awards at the Scientific Interface. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: TS and YE are employees of MyHeritage Ltd.

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 [10–22]. LMMs are routinely used nowadays to analyze pedigrees of millions of animals and plants [13,23], hundreds of thousands of which are often genotyped (e.g. [22,24,25]).

LMMs and their extensions have recently gained considerable popularity in human genetics studies for the purposes of estimating heritability [26–32] and genetic correlation [33–37], predicting phenotypes [38–41] and modeling sample relatedness [42–46]. 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 have been using 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 [21,47,48].

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 [46], or by approximating the restricted likelihood gradient via Monte-Carlo techniques [28]. 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 [49–53]. Such approaches have become increasingly popular recently [30–35,54–60] owing to their computational tractability and their compatibility with privacy-preserving summary statistics [61]. 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 [56]; they are more robust to modeling

violation because they make fewer distributional assumptions; and they are more flexible, which enables applying techniques to limit confounding factors such as assortative mating (Methods). Moment estimators have also recently been explored in animal breeding studies [62–64] 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 present 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 [65,66] (Methods). Sci-LMM scales HE regression to population-sized pedigrees via sparse matrix tools [67]. The REML estimator combines a direct sparse REML solver [47] with Monte-Carlo gradient approximation [68]. Importantly, existing packages for pedigree-based REML [69–73] cannot handle the analyses performed in this paper because they require the inverse of the epistatic interactions matrix [47,74,75], which is extremely difficult to compute in large pedigrees [76]. 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.

Material and methods

Linear mixed models

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

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{C\beta}; \boldsymbol{\Sigma}) \tag{1}$$

$$\boldsymbol{\Sigma} = \mathbf{G} + \sigma_e^2 \boldsymbol{I} \tag{2}$$

$$\boldsymbol{G} = \sum_{k=1}^{d} \sigma_k^2 \boldsymbol{M}^k. \tag{3}$$

Here, $\boldsymbol{C} = [\boldsymbol{C}_1,\ldots,\boldsymbol{C}_n]^T$ is an $n\times c$ matrix of covariates (including an intercept), $\boldsymbol{\beta}$ is a $c\times 1$ vector of fixed effects, $\boldsymbol{\Sigma}$ is the covariance matrix of the vector $\boldsymbol{y}$, σ_k^2 is the k^{th} variance component, and $\boldsymbol{I}$ is the identity matrix. The parameters to estimate are the fixed effects $\boldsymbol{\beta}$ and the variance components $\sigma_1^2,\ldots,\sigma_d^2,\sigma_e^2$. The Sci-LMM software can currently compute an identity by descent (IBD) matrix, an epistatic covariance matrix and a dominance matrix, as described below.

The restricted log-likelihood $\ell_R(\boldsymbol{\beta},\sigma_1^2,\ldots,\sigma_d^2,\sigma_e^2)$ is given by [77]:

$$\ell_R\left(\boldsymbol{\beta},\sigma_1^2,\ldots,\sigma_d^2,\sigma_e^2\right) = \ell\left(\boldsymbol{\beta},\sigma_1^2,\ldots,\sigma_d^2,\sigma_e^2\right) + \frac{c}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{C}^T\mathbf{C}| - \frac{1}{2}\log|\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}| \tag{4}$$

$$\ell\left(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2\right) = -\frac{1}{2}(\mathbf{y} - \mathbf{C}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{C}\boldsymbol{\beta}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi), \tag{5}$$

where $\ell(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2)$ is the non-restricted likelihood and c is the number of covariates.

An alternative form of Eq 4 often used in animal breeding literature is:

$$\ell_R\left(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2\right) = -\frac{1}{2}(\log|\boldsymbol{B}| + \log|\boldsymbol{\Sigma}| + \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}), \tag{6}$$

where $\boldsymbol{B} = \begin{bmatrix} \sigma_e^{-2}\boldsymbol{C}^T\boldsymbol{C} & \sigma_e^{-2}\boldsymbol{C}^T \\ \sigma_e^{-2}\boldsymbol{C} & \sigma_e^{-2}\boldsymbol{I} + \boldsymbol{G}^{-1} \end{bmatrix}$, $\boldsymbol{P} = \sigma_e^{-2}\boldsymbol{I} - \sigma_e^{-4}\boldsymbol{W}\boldsymbol{B}^{-1}\boldsymbol{W}^T$, $\boldsymbol{W} = [\boldsymbol{C}\,\boldsymbol{I}]$ and we ignored additive constants. This form is particularly convenient when the inverse of each of the matrices $\boldsymbol{M}^1,\ldots,\boldsymbol{M}^d$ is known, as it can be solved efficiently using mixed model equations via Gaussian elimination, without having to directly invert or factorize the matrix $\boldsymbol{\Sigma}$ [47,72]. This makes it particularly convenient to use this form in the presence of only an additive IBD matrix, because the inverse of this matrix is sparse and can be computed analytically [78,79].

Restricted maximum likelihood (REML) estimation

REML estimation consists of finding the parameters $\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2$ that maximize Eq 4. When the inverse of each of the matrices $\boldsymbol{M}^1,\ldots,\boldsymbol{M}^d$ is known, the REML can be found efficiently by using Eq 6, using the so-called mixed model equations method [47,72]. Here we describe a direct solution that can be applied when the inverse of $\boldsymbol{M}^1,\ldots,\boldsymbol{M}^d$ is unknown.

Our solution combines several ideas: (1) we maximize Eq 4 directly, rather than the equivalent form of Eq 6; (2) instead of directly inverting $\boldsymbol{\Sigma}$, we compute its Cholesky factorization $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$ via sparse matrix routines; (3) any product of the form $\boldsymbol{\Sigma}^{-1}\boldsymbol{v}$ for some vector $\boldsymbol{v}$ is computed using L and two triangular solvers (forward and backward substitution); and (4) the gradient of Eq 4 is approximated using Monte Carlo techniques. We now describe our REML approach in detail.

We first describe a solution to the unrestricted log-likelihood (Eq 5) and then extend the solution to the restricted log-likelihood (Eq 4). To compute Eq 5 we need to compute the terms $\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{C}\boldsymbol{\beta})$ and $\log|\boldsymbol{\Sigma}|$. The first term can be computed exactly via either conjugate gradient iterations or by explicitly computing the Cholesky factorization of $\boldsymbol{\Sigma}$ and then applying forward and back substitution. The second term can be computed via the Cholesky factorization of $\boldsymbol{\Sigma}$. The Cholesky factorization can be computed efficiently via the CHOLMOD routines [80]. It remains to find the maximum likelihood estimates of the model parameters.

To find the MLE of $\hat{\boldsymbol{\beta}}$ we note that given $\boldsymbol{\Sigma}$, $\hat{\boldsymbol{\beta}}$ can be computed analytically by deriving Eq 5 with respect to $\boldsymbol{\beta}$ as follows:

$$\frac{\partial\ell(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2)}{\partial\boldsymbol{\beta}} = -\frac{1}{2}\left(\boldsymbol{y} - \mathbf{C}\boldsymbol{\beta}\right)^T\boldsymbol{\Sigma}^{-1}\mathbf{C} \tag{7}$$

By setting the transpose of the gradient to 0, we obtain the MLE:

$$\hat{\boldsymbol{\beta}} = (\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{8}$$

The MLEs of the variance components $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2, \hat{\sigma}_e^2$ are estimated via an optimization procedure, which requires computing the gradient of Eq 5. The partial derivative with respect to

each variance component σ_k^2 is given by:

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2)}{\partial \sigma_k^2} = -\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\right]. \tag{9}$$

The first term on the right-hand side of Eq 9 can be computed efficiently given the Cholesky factorization of $\boldsymbol{\Sigma}$. Unfortunately, the second term cannot be solved efficiently via the above technique because it requires solving n different linear equations, where n can be in the millions. Instead, we use the approximation technique used in [28,68,81]. We first rewrite this term as an expectation (ignoring the scaling factor) as follows:

$$\begin{aligned}
\mathrm{Tr}[\boldsymbol{\Sigma}^{-1}\mathbf{M}^k] &= \mathrm{Tr}[\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}] \\
&= \mathrm{Tr}[\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\boldsymbol{\Sigma}^{-1}\mathrm{E}[\mathbf{y}'\mathbf{y}'^T]] \\
&= \mathrm{E}[\mathrm{Tr}[\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\boldsymbol{\Sigma}^{-1}\mathbf{y}'\mathbf{y}'^T]] \\
&= \mathrm{E}[\mathrm{Tr}[\mathbf{y}'^T\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\boldsymbol{\Sigma}^{-1}\mathbf{y}']] \\
&= \mathrm{E}[\mathbf{y}'^T\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\boldsymbol{\Sigma}^{-1}\mathbf{y}'],
\end{aligned} \tag{10}$$

where $\mathbf{y}' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and we used the fact that the trace of a scalar is equal to the scalar. We therefore approximate Eq 10 by sampling a small number of $\mathbf{y}'$ vectors to approximate the expectation. These vectors can be sampled efficiently given the Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ by sampling a vector $\mathbf{y}_{\boldsymbol{\Sigma}} \sim \mathcal{N}(0, \boldsymbol{I})$ and then using the fact that $\mathbf{L}\mathbf{y}_{\boldsymbol{\Sigma}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The Cholesky factorization can be computed efficiently via the CHOLMOD routines. We found that 100 vectors often yields a very good approximation at a modest computational cost.

We note that [28] proposes an alternative estimation method by completely foregoing the likelihood computation, and instead only trying to minimize the squared gradient elements. However, we found that in sparse settings, this solution often converges into local maxima at the edge of the parameter space (where many variance components are equal to zero) rather than the true maximum likelihood estimate.

We now extend the solution to handle restricted maximum likelihood (Eq 4). Clearly, the restricted maximum likelihood estimate of $\boldsymbol{\beta}$ is the same as the MLE. The derivative of the restricted log likelihood with respect to each variance component σ_k^2 is given by:

$$\frac{\partial \ell_R(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2)}{\sigma_k^2} = \frac{\partial \ell(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2)}{\sigma_k^2} + \frac{1}{2}\mathrm{Tr}\left[(\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{M}^k\boldsymbol{\Sigma}^{-1}\mathbf{C}\right]. \tag{11}$$

The term $\boldsymbol{\Sigma}^{-1}\mathbf{C}$ can be computed by solving c different linear equations, which can be performed efficiently given the Cholesky factorization of $\boldsymbol{\Sigma}$. All the other terms can be computed efficiently, assuming that c is small compared to n.

The standard errors of $\sigma_1^2, \ldots, \sigma_d^2$ can be approximated via the average information REML (AI-REML) procedure [82], which consists of approximating each entry of the Hessian of the restricted log likelihood as follows:

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2)}{\sigma_k^2\sigma_l^2} \approx -\frac{1}{2}\mathbf{y}^T\boldsymbol{P}\mathbf{M}^k\mathbf{P}\mathbf{M}^l\mathbf{P}\mathbf{y}, \tag{12}$$

where $\mathbf{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{C}(\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}$. Afterwards we approximate the standard errors via the square roots of the diagonal entries of the inverse of the negative Hessian. Following [28], we multiply these entries by $\left(1 + \frac{1}{100}\right)$ to account for sampling variance introduced by the 100 $\boldsymbol{y}'$ vectors sampled in the Monte-Carlo approximation.

Implementation details

We implemented our REML solver in Python, using an L-BFGS-B algorithm [83] as implemented in the SciPy package [67]. To prevent the parameters from inducing a non positive-definite matrix, We enforced non-negative parameters by using a log-transformation, which transforms the problem into an unconstrained optimization problem.

Haseman-Elston regression

HE regression estimates variance components via the method of moments, by finding the set of variance components $\sigma_1^2, \ldots, \sigma_d^2, \sigma_e^2$ that minimize the expression:

$$\sum_{i,j} (\mathrm{cov}(y_i - \boldsymbol{C}_i\boldsymbol{\beta}, y_j - \boldsymbol{C}_j\boldsymbol{\beta}) - \Sigma_{ij})^2. \tag{13}$$

Typically, the fixed effects $\boldsymbol{\beta}$ are first estimated without considering the covariance matrices, by solving the multivariate linear regression problem $\boldsymbol{y} = \boldsymbol{C}\boldsymbol{\beta} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_e^2\boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. This solution yields a consistent estimator under mild regularity conditions [84]. Afterwards we plug the fixed effect estimate $\hat{\boldsymbol{\beta}}$ into Eq 13 and estimate the variance component estimates $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2$ as follows:

$$[\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2]^T = ([\mathbf{V}^1, \ldots, \mathbf{V}^d]^T[\mathbf{V}^1, \ldots, \mathbf{V}^d])^{-1}[\mathbf{V}^1, \ldots, \mathbf{V}^d]^T\mathbf{Y}, \tag{14}$$

where $\mathbf{V}^k$ is a vector representation enumerating the elements M_{ij}^k for all pairs of distinct individuals i,j, and $\mathbf{Y}$ is a vector representation of the corresponding elements $(y_i - \mathbf{C}_i\hat{\boldsymbol{\beta}})(y_j - \mathbf{C}_j\hat{\boldsymbol{\beta}})$. Each element q,r of the $d \times d$ matrix $([\mathbf{V}^1,\ldots,\mathbf{V}^d]^T[\mathbf{V}^1\ldots,\mathbf{V}^d])$ can be computed via an element-wise multiplication of the upper-diagonal elements of the matrices $\mathbf{M}^q$, $\mathbf{M}^r$, which can be performed efficiently via sparse matrix routines. The vector $[\mathbf{V}^1,\ldots,\mathbf{V}^d]^T\mathbf{Y}$ can also be computed efficiently in a similar manner.

By following the notation of [56] and denoting $\mathbf{q} \triangleq [\mathbf{V}^1,\ldots,\mathbf{V}^d]^T\mathbf{Y}$, $\mathbf{S} = [\mathbf{V}^1,\ldots,\mathbf{V}^d]^T[\mathbf{V}^1,\ldots,\mathbf{V}^d]$, we have:

$$[\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2]^T = \mathbf{S}^{-1}\mathbf{q}. \tag{15}$$

By applying a few matrix manipulations, we can compute $\mathbf{q}$ and $\mathbf{S}$ efficiently as follows:

$$q_k = \mathbf{y}^T\mathbf{M}^k\mathbf{y} - \sum_i M_{ii}^k y_i^2 = \mathbf{y}^T(\mathbf{M}^k - \mathbf{I})\mathbf{y} \tag{16}$$

$$\mathbf{S}_{kl} = \sum_{i,j} M_{ij}^k M_{ij}^l - \sum_i \mathbf{M}_{ii}^k \mathbf{M}_{ii}^l, \tag{17}$$

where we used the assumption $M_{ii}^k = 1$. Both these quantities can be computed explicitly via sparse matrix routines.

The sampling variance of the estimators is given by $\mathbf{S}^{-1}\mathrm{var}(\mathbf{q})\mathbf{S}^{-1}$, where $\mathrm{var}(\mathbf{q})$ is given by:

$$\mathrm{var}(\boldsymbol{q})_{kl} = 2\mathrm{tr}(\hat{\boldsymbol{\Sigma}}(\mathbf{M}^k - \mathbf{I})\hat{\boldsymbol{\Sigma}}(\mathbf{M}^l - \mathbf{I})). \tag{18}$$

This quantity can be computed in two ways:

1. Exactly, via: $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}(\mathbf{M}^k - \mathbf{I})\hat{\boldsymbol{\Sigma}}(\mathbf{M}^l - \mathbf{I})) = \sum_{ij} [\hat{\boldsymbol{\Sigma}}(\mathbf{M}^k - \mathbf{I})]_{ij}[\hat{\boldsymbol{\Sigma}}(\mathbf{M}^l - \mathbf{I})]_{ij}$
2. Approximately, via: $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}(\mathbf{M}^k - \mathbf{I})\hat{\boldsymbol{\Sigma}}(\mathbf{M}^l - \mathbf{I})) = \mathrm{E}_{\mathbf{y}'}[\mathbf{y}'^T(\mathbf{M}^k - \mathbf{I})\hat{\boldsymbol{\Sigma}}(\mathbf{M}^l - \mathbf{I})\mathbf{y}']$,

where $\mathbf{y}'$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, using a derivation similar to the one in Eq 10.

The approximate approach uses Monte-Carlo approximations, by randomly sampling $\mathbf{y}'$ vectors and approximating the right hand side. It can be substantially faster than the exact approach (because it circumvents expensive matrix-matrix multiplications) and obtain excellent accuracy. The Sci-LMM software uses the approximate approach by sampling ~100 random $\mathbf{y}'$ vectors.

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 by zeroing the covariance matrix entries of corresponding pairs. Importantly, this technique cannot be used in REML, because the resulting covariance matrices may not be positive definite. We note that another potential approach to capture environmental risk factors is including shared effects with a suitable incidence matrix [10], but this approach requires additional assumptions and has not been used here.

Factors affecting estimation accuracy

HE regression is a convenient theoretical framework to analyze the factors affecting estimation accuracy. HE regression can be considered as a special form of linear regression, where off-diagonal entries of covariance matrices serve as explanatory variables. Hence, good accuracy is obtained when measured and unmeasured explanatory variables are uncorrelated with other explanatory variables (Eq 18).

Specifically, obtaining accurate estimates requires (1) that the off-diagonal entries of the LMM covariance matrices are uncorrelated with each other; and (2) that they are uncorrelated with covariance due to unmeasured environmental factors. While the first requirement can be easily tested, the second one requires making strong assumptions about the structure of environmental covariance. For example, if latent environmental factors are shared between spouses but not between parents and children, we may wish to exclude spouses from the analysis. Unfortunately, we not know the structure of environmental covariance for the traits studied in this work, and we leave its investigation for future work.

Identity 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 was originated from the same chromosome of a shared ancestor between individuals i and j [85,86], 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 [85], 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 |path| is the path length.

To efficiently compute the IBD matrix we first construct the matrices $\boldsymbol{L}$ and $\boldsymbol{H}$ of its decomposition $\boldsymbol{A} = \boldsymbol{LHL}^{\boldsymbol{T}}$, 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 (Fig 1A–1C). The matrices $\boldsymbol{L}$ and $\boldsymbol{H}$ can be computed efficiently via iterative techniques [78,79] using sparse matrix routines [80] (S1 Text).

Dominance kinship 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 [10,87]. A necessary condition for nonzero dominancy entry is a nonzero IBD relationship, which enables rapid computation of the dominance matrix.

Epistatic kinship 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., $(A_{k,l})^2$ for two-loci epistasis, $(A_{k,l})^3$ for three-loci epistasis and so on [75]. Therefore, an epistatic covariance matrix is simple to compute 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 [70–73]. 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 (S1 Text; S1 Fig). This algorithm reduces the matrix construction time by several hours.

Computing IBD principal components

In addition to covariance matrices, Sci-LMM can include the top principal components (PCs) of the IBD matrix as fixed effects, using sparse matrix routines [88]. The inclusion of PCs can capture major linear sources of variation in a dataset, and is motivated by large scale human genetic studies, where such PCs often capture population structure [89]. However, we caution that PCs computed from an IBD matrix are not guaranteed to capture population structure [90]. An alternative approach often employed in animal studies is the assignment of unobserved parents to genetic groups [91], but this approach requires knowledge about the location of birth of all individuals without known parents.

Data simulations

We generated pedigrees mimicking real family patterns in the United States, partially based on publications by the United States Census Bureau [92,93]. We iteratively generated generations of individuals, where the first generation included two individuals, and the number of

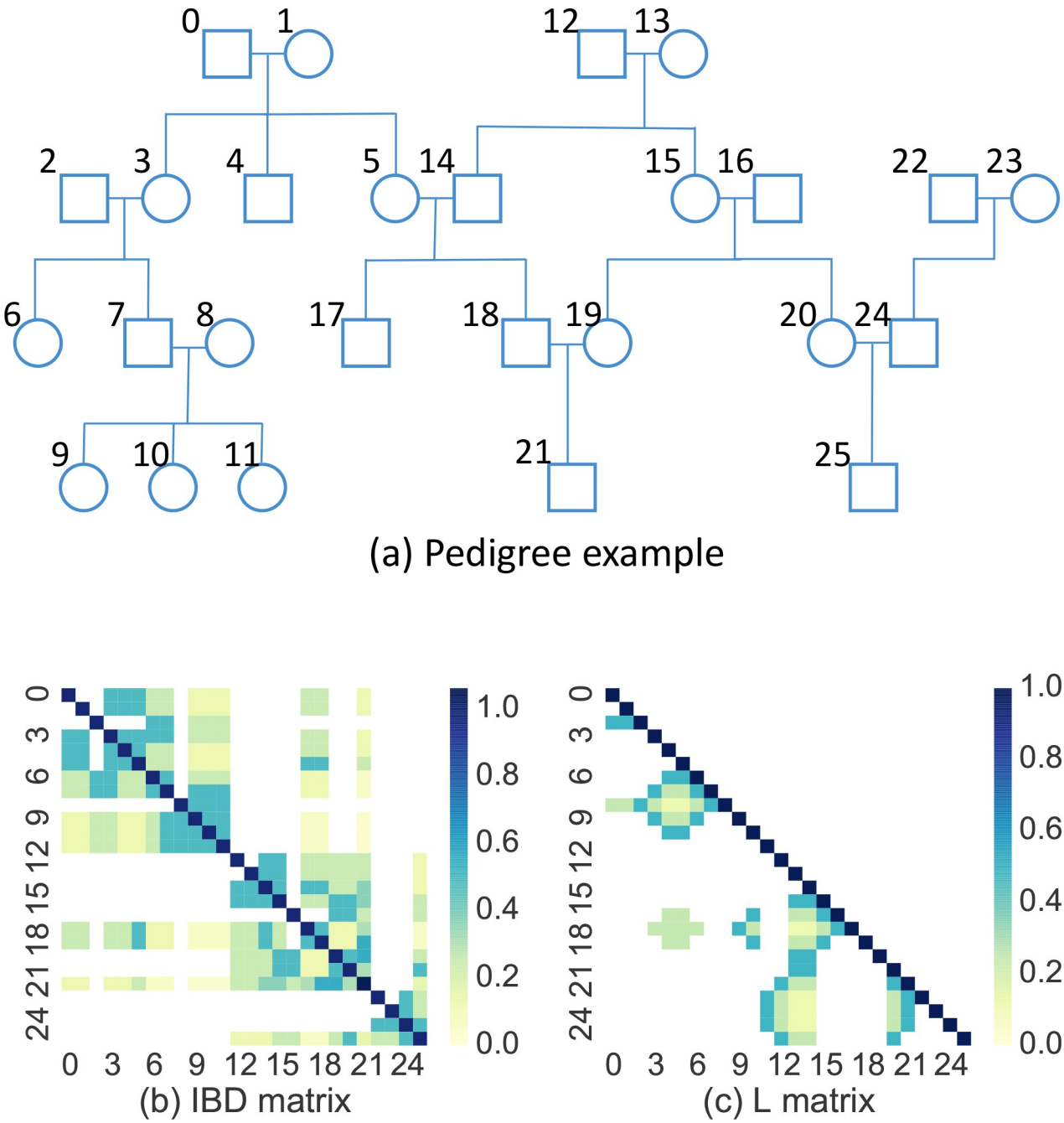

Fig 1. A demonstration of the Sci-LMM IBD matrix construction algorithm. (**a**) An example pedigree with 26 individuals. (**b**) A heat-map representing the IBD matrix, where zero elements are white to emphasize sparsity. (**c**) A heat-map representing the lower Cholesky factorization of the IBD matrix (i.e. the matrix L in the factorization $\boldsymbol{A} = \boldsymbol{LHL}^T$, 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.

https://doi.org/10.1371/journal.pgen.1008124.g001

individuals in each successive generation increases by 40% (approximately the same ratio as in the GENI dataset), until obtaining the desired sample size. Each generation included 50% females and 50% males.

In each generation we generated households, where every household includes either one individual or two individuals with different genders, and every individual can belong to zero,

one or multiple households. The number of households in each generation was 62.5% of the number of individuals in that generation. 68% of the households included pairs of individuals, and the rest included a single individual. Every individual in every generation (except for the top one) was born to parents from a randomly selected household from the previous generation (for 80% of individuals) or from two generations in the past (for the remaining 20% of individuals).

After generating all individuals, we omitted randomly selected edges until obtaining the desired sparsity factor, up to 10% error. We then created corresponding IBD, dominance and epistasis matrices.

Finally, we generated phenotypes using Eq 1 by (1) generating variance components σ_k^2 for each covariance matrix M^k from $U(0,1)$ and scaling them such that they sum to 1.0; (2) Generating 5 binary and 5 normally distributed covariates; and (3) generating fixed effects from $\mathcal{N}(0, 1000/n)$, where n is the sample size.

The parameters differentiating the various experiments are: (1) cohort size (50K, 100K, 250K, 500K, 1M or 2M); (2) sparsity factor (0.0005, 0.001, or 0.005); and (3) the subset of matrices used. We generated 10 different datasets for every unique combination of settings, except for matrices with 2M individuals, for which we generated a single pedigree with ten different phenotype vectors due to runtime considerations.

Computing environment

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

Results

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 (Methods).

In each simulation we generated a normally distributed phenotype, using a covariance matrix with additive, epistatic and dominance effects and ten binary covariates. Unless otherwise stated, the sparsity factor (the fraction of non-zero entries in each matrix) was 0.001. Ten different datasets were generated for each combination of sample size and sparsity factor.

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, (Fig 2A–2C). 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 (Fig 2A–2C), 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 (Fig 2D and 2E).

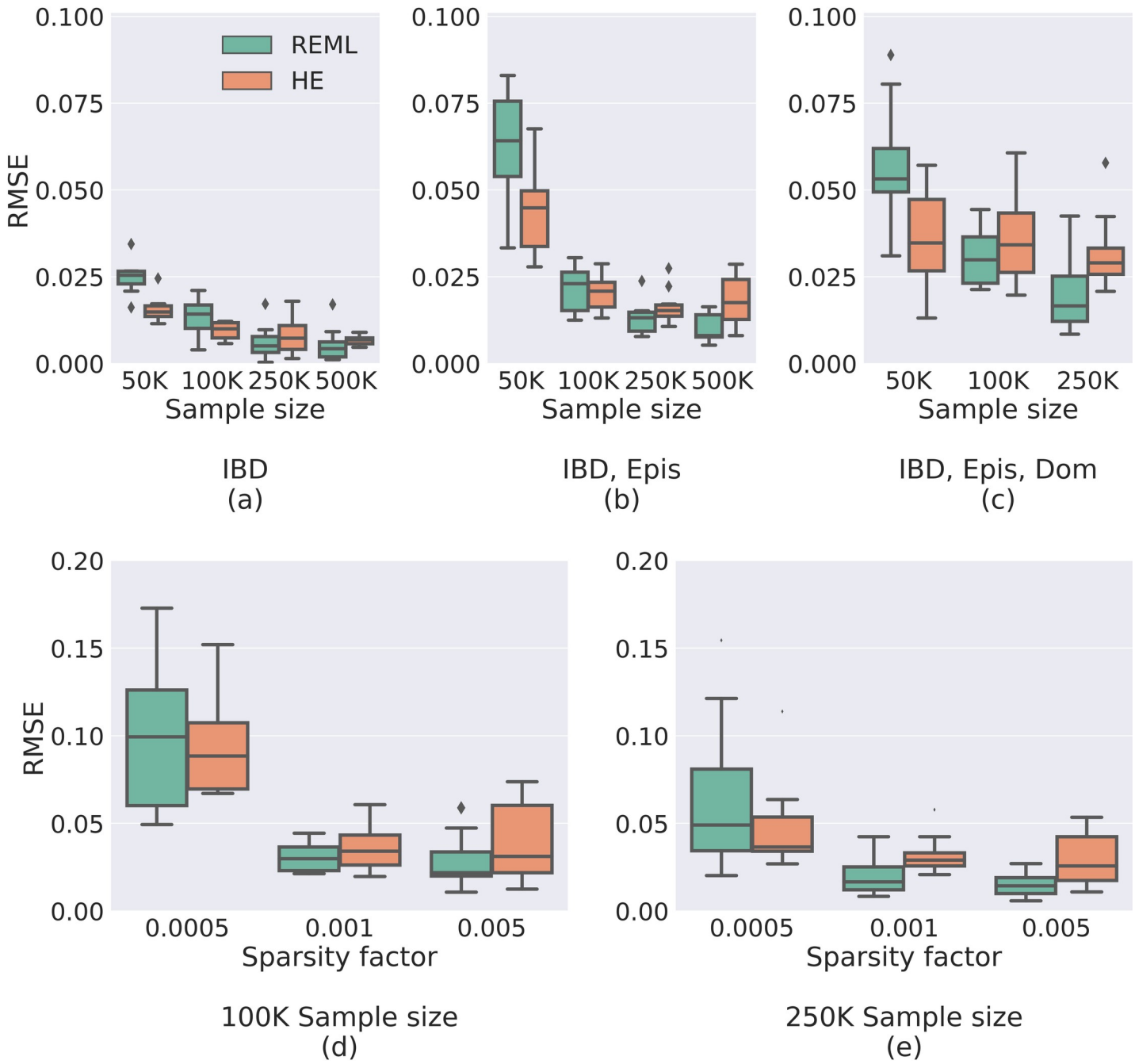

Fig 2. Evaluating the estimation accuracy of Sci-LMM. (**a-c**) Box plots comparing REML and HE estimation accuracy (RMSE) across simulated datasets (each box represents 10 experiments), under varying sample sizes, using **(a)** only IBD, **(b)** IBD and epistasis, or **(c)** IBD, epistasis and dominance 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.

https://doi.org/10.1371/journal.pgen.1008124.g002

Runtime. We evaluated the time Sci-LMM requires to construct covariance matrices. The covariance matrices computation is dominated by the IBD matrix construction, because the other matrices can be computed trivially given this matrix (Methods). The IBD matrix

construction scaled linearly with the number of non-zero entries in the matrix (Fig 3A and 3B). For example, Sci-LMM required less than 4 hours to construct an IBD matrix with 5×10^{11} pairs of possible relatives and a sparsity factor of 0.001.

We next investigated the runtime for variance component estimation using REML and HE. REML estimation for samples with 500,000 individuals (representing 250 million covariance entries) required less than 24 hours (Fig 3C), whereas HE estimation required 16 seconds (Fig 3D). Overall, our results demonstrate that the Sci-LMM framework is scalable to extremely large pedigrees.

We also compared Sci-LMM with several REML software packages [69–72]. We could not invoke any of these packages with an epistatic covariance matrix because they require its inverse, whose computation is more computational demanding than that of the IBD matrix [76]. We verified this by trying to invert the matrix analytically via the software package nadiv [94] and numerically via sparse matrix libraries [80] and via the matrix inversion facilities of the software package WOMBAT [71], 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 [95] and the *lmekin* function of the coxme package [96], both of which can work with sparse covariance matrices. However, both packages could not complete the analysis in 4 days, presumably because they do not use the approximate gradient approximation techniques of Sci-LMM.

Finally, we compared Sci-LMM to WOMBAT in the presence of only an additive covariance matrix. WOMBAT is more computationally efficient in this setting because it uses a mixed model equations (MME) solver [47]. MME solvers scale roughly quadratically with the sample size, compared with the cubic complexity of Sci-LMM, but they require pre-computing the inverse of the LMM covariance matrix. Thus, using an MME solver is advantageous in the presence of only an IBD covariance matrix, whose inverse has an analytical form that can be computed efficiently [78,79]. Indeed, WOMBAT was much faster than the REML solver of Sci-LMM in this setting, completing an analysis of 250,000 individuals in 13 minutes, compared with 164 minutes for Sci-LMM (S1 Table). However, both WOMBAT and the REML solver of Sci-LMM crashed in the presence of $\geq$500,000 individuals, whereas the HE solver of Sci-LMM could complete the analysis in less than 20 minutes. Hence, the HE solver of Sci-LMM is the only tool that we are aware of that can readily scale to population scale pedigrees.

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 (Fig 4A). 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 raised to second and third power, and the top 10 principal

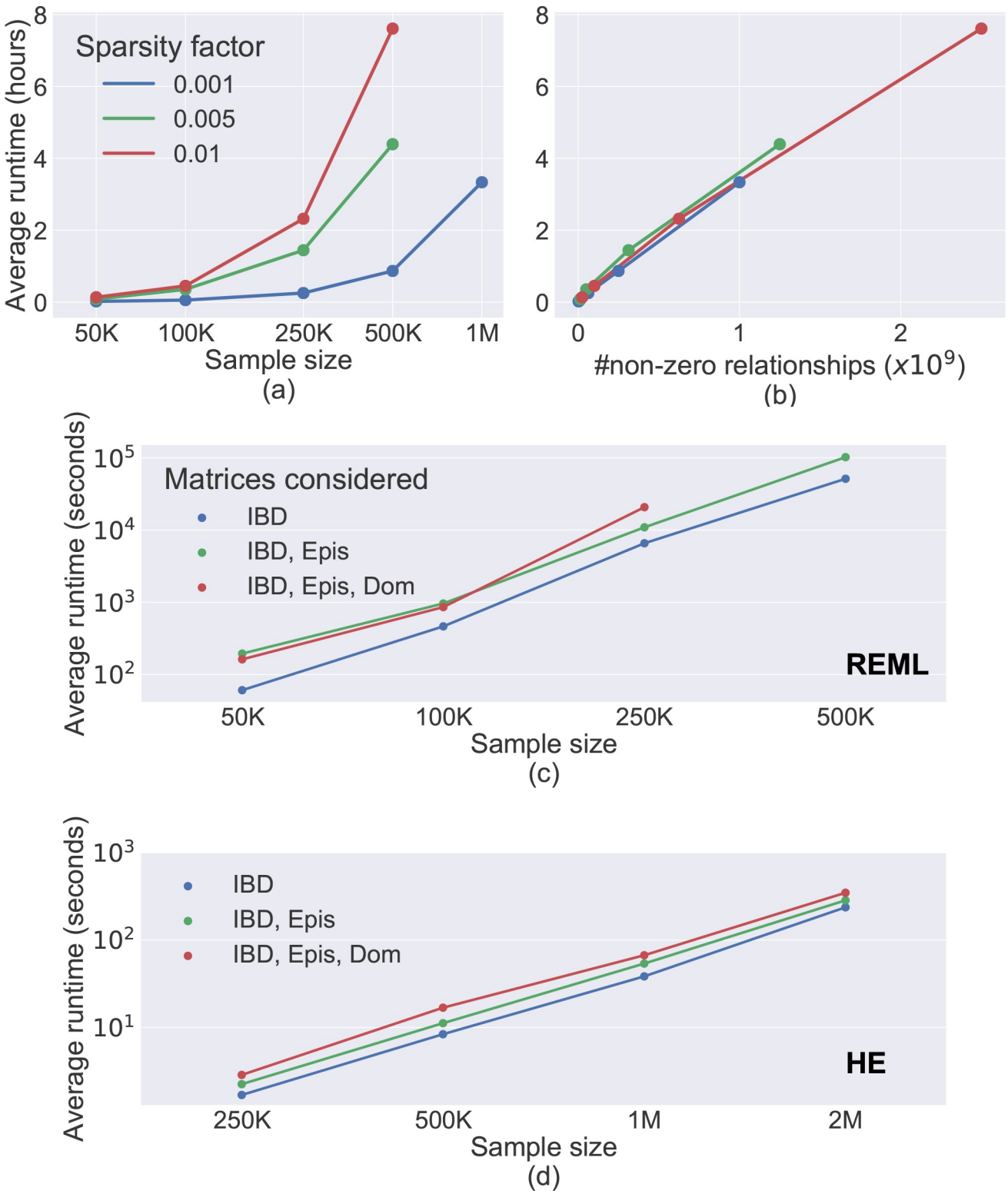

Fig 3. Analysis of Sci-LMM computation time. (a) Computation time required to compute an IBD matrix from pedigree data under different sparsity factors as a function of sample size. **(b)** Computation time required to compute an IBD matrix from pedigree data as a function of the number of nonzero relationships, 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.

https://doi.org/10.1371/journal.pgen.1008124.g003

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 full-sibs or double first cousins, rendering this matrix almost equivalent to the identity matrix (S2 Fig).

The Sci-LMM 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%) (Fig 4B). The HE estimate for IBD was 24.3% (s.e. 0.5%) when omitting the epistatic interactions matrix (HE results with epistatic interactions were inconclusive due to large standard errors). A potential challenge of our framework is that genetic covariance may be correlated with shared environmental factors (Methods) [2,97]. We tried mitigating this problem 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 [2,97] (see Discussion).

We next 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, because the family records of such individuals are more likely to be incomplete. We additionally removed individuals with a shared household (spouses and children) and individuals who are first- or second-degree

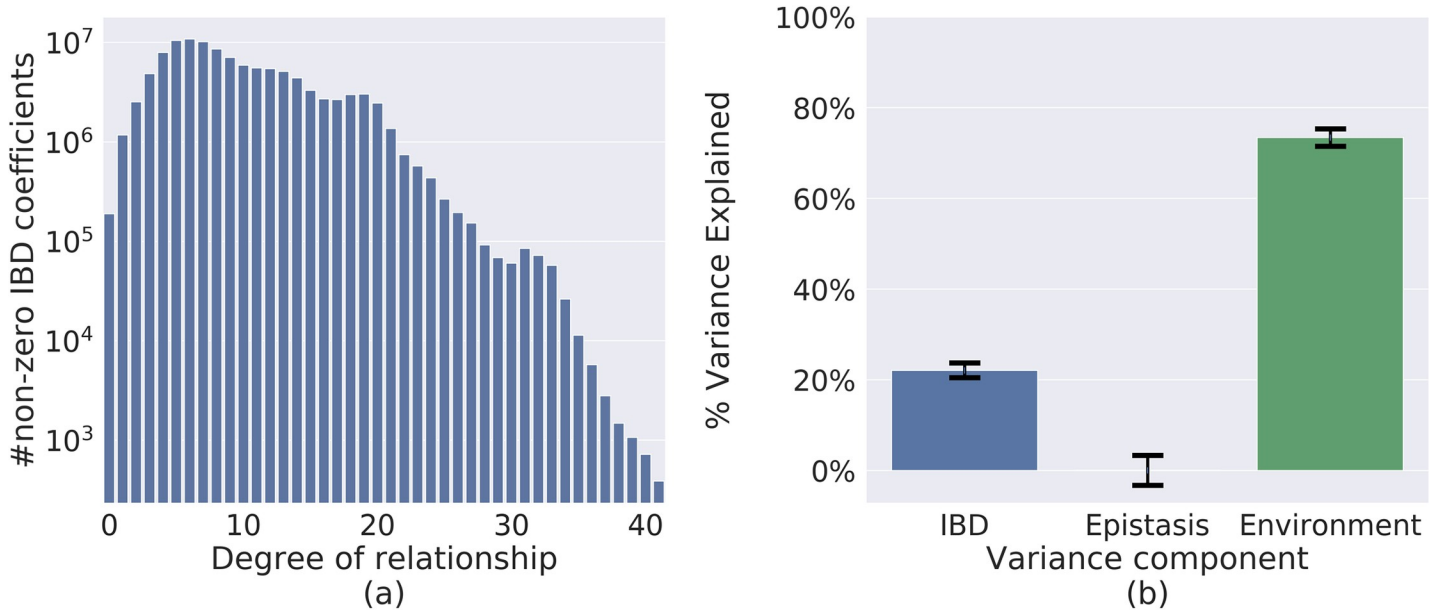

Fig 4. Results of analysis of a real pedigree with 441,000 individuals. (a) A histogram of genetic similarity across 441,000 individuals, using only the closest relationship between every pair of individuals. The degree of relationship between a pair of individuals is given by $-\log_2(K_{ij})$-1, where K_{ij} is their IBD coefficient (Methods). 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).

https://doi.org/10.1371/journal.pgen.1008124.g004

relatives of 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 [98]. However, we note several potential caveats. First, our analysis estimated the heritability of reproductive fitness conditional on having $\geq$2 children; we excluded individuals with fewer children to minimize bias towards individuals with more complete family records. Second, our estimate may be upper-biased due to confounding by non-genetic factors [2,97] (see Discussion). Third, our analysis may be improved by including random effect for shared households or other shared environmental factors. However, such analyses require the introduction of additional modeling assumptions, which we leave for future work.

Finally, we performed a series of experiments to examine the robustness of our longevity analysis to potential confounding factors (S2 Table). First, we restricted the analysis to 283,073 individuals born after 1800, which yielded heritability estimates of 0.23 (0.006) under HE and 0.23 (0.004) under REML. Second, we restricted the analysis to 276,011 individuals who are unlikely to have shared a household during their lifetime (i.e., no spouses or first-degree relatives), which yielded estimates of 0.29 (0.006) under HE and 0.29 (0.005) under REML. Third, we restricted the analysis to 110,237 individuals who were not first or second-degree relatives, which led to estimates of 0.53 (0.040) under HE and 0.51 (0.040) under REML. Finally, we combined the last two restrictions by only including 106,049 individuals who were neither spouses or first or second-degree relatives of each other, leading to estimates of 0.54 (0.045) under HE and 0.51 (0.044) under REML. The inflated estimates when excluding first or second-degree relatives may indicate that weaker levels of IBD covariance are correlated with shared environmental covariance, leading to inflated heritability estimates. Nevertheless, these results indicate that our original analysis is not upper biased due to inclusion of the above potential confounding factors.

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 [99–104], 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 [47], which is implemented in several software packages [69–72].

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 [105,106], environmental bias [97,107], 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 recently proposed techniques to address these issues (e.g. [2,106,107]) 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 [19,21,108–112]. However, privacy and logistical concerns limit public access to human genetic data, necessitating methods based on summary statistics [61]. Thus, approaches for analysis of such combined datasets will combine state of the art techniques from the animal breeding and the human genetics literature, and remain a potential avenue for future work.

Supporting information

S1 Text. Detailed algorithms for covariance matrix construction and pedigree pruning.
(PDF)

S1 Table. Comparison of Sci-LMM and WOMBAT runtime and memory requirements, when using simulations with only an additive IBD matrix. Both tools used only 1 CPU thread, and WOMBAT was executed with the—meuwissen flag. The estimated values were essentially the same for both tools in all cases. WOMBAT and Sci-LMM REML both crashed in the presence of pedigrees with $\geq$500,000 individuals.
(PDF)

S2 Table. Sci-LMM heritability of longevity estimates under different analysis approaches. The epistasis estimates were very close to zero in all cases and are omitted for clarity.
(PDF)

S1 Fig. Stages of removal of uninformative individuals. Nodes represent individuals, and edges represent parent-child relations. Only red individuals have full information records (e.g. year of birth, year of death, etc.)
(TIF)

S2 Fig. The number of nonzero IBD and Dominance entries in the GENI dataset, as a function of degree of relationship.
(TIF)

Acknowledgments

We thank Regev Schweiger, Elior Rahmani, Eran Halperin and Saharon Rosset for fruitful discussion.

Author Contributions

Conceptualization: Tal Shor, Yaniv Erlich, Omer Weissbrod.

Data curation: Tal Shor, Yaniv Erlich.

Formal analysis: Tal Shor, Omer Weissbrod.

Investigation: Tal Shor, Yaniv Erlich, Omer Weissbrod.

Methodology: Tal Shor, Dan Geiger, Yaniv Erlich, Omer Weissbrod.

Project administration: Dan Geiger.

Resources: Yaniv Erlich, Omer Weissbrod.

Software: Tal Shor, Iris Kalka, Omer Weissbrod.

Supervision: Dan Geiger, Omer Weissbrod.

Validation: Iris Kalka, Yaniv Erlich, Omer Weissbrod.

Writing – original draft: Tal Shor, Omer Weissbrod.

Writing – review & editing: Tal Shor, Iris Kalka, Yaniv Erlich, Omer Weissbrod.

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. https://doi.org/10.1126/science.aam9309 PMID: 29496957
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. https://doi.org/10.1534/genetics.118.301613 PMID: 30401766
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. https://doi.org/10.1038/sdata.2015.11 PMID: 25977816
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. https://doi.org/10.1093/bioinformatics/btx569 PMID: 28968884
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. https://doi.org/10.1016/j.cell.2018.04.032 PMID: 29779949
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. https://doi.org/10.1002/ajmg.b.32548 PMID: 28557243
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. https://doi.org/10.1016/j.ajhg.2018.10.017 PMID: 30526866
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. https://doi.org/10.1111/j.1420-9101.2007.01377.x PMID: 17714306

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. https://doi.org/10.1098/rstb.2009.0203 PMID: 20008387
18. Gianola D, Rosa GJM. One Hundred Years of Statistical Developments in Animal Breeding. Annu Rev Anim Biosci. 2015; 3: 19–56. https://doi.org/10.1146/annurev-animal-022114-110733 PMID: 25387231
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. https://doi.org/10.1007/s00122-016-2750-y PMID: 27435734
20. Manfredi E, Tusell L, Vitezica ZG. Prediction of complex traits: Conciliating genetics and statistics. J Anim Breed Genet. 2017; 134: 178–183. https://doi.org/10.1111/jbg.12269 PMID: 28508479
21. Misztal I, Legarra A. Invited review: efficient computation strategies in genomic selection. animal. 2017; 11: 731–736. https://doi.org/10.1017/S1751731116002366 PMID: 27869042
22. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev Genet. 2019; 20: 135–156. https://doi.org/10.1038/s41576-018-0082-2 PMID: 30514919
23. 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.
24. 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. https://doi.org/10.1186/s12711-016-0227-8
25. Mäntysaari EA, Evans RD, Strandén I. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. J Anim Sci. 2017; 95: 4728–4737. https://doi.org/10.2527/jas2017.1912 PMID: 29293736
26. 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. https://doi.org/10.1038/ng.608 PMID: 20562875
27. 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. https://doi.org/10.1016/j.ajhg.2010.11.011 PMID: 21167468
28. 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. https://doi.org/10.1038/ng.3431 PMID: 26523775
29. Lee SH, van der Werf JHJ. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics. 2016; 32: 1420–1422. https://doi.org/10.1093/bioinformatics/btw012 PMID: 26755623
30. Golan D, Lander ES, Rosset S. Measuring missing heritability: Inferring the contribution of common variants. Proc Natl Acad Sci. 2014; 111: E5272–81. https://doi.org/10.1073/pnas.1419064111 PMID: 25422463
31. 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. https://doi.org/10.1038/ng.3211 PMID: 25642630
32. 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. https://doi.org/10.1371/journal.pgen.1006711 PMID: 28388634

33. 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. https://doi.org/10.1038/ng.3406 PMID: 26414676

34. 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. https://doi.org/10.1016/j.ajhg.2017.11.001 PMID: 29220677

35. 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. https://doi.org/10.1016/j.ajhg.2018.06.002 PMID: 29979983

36. 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. https://doi.org/10.1093/bioinformatics/bts474 PMID: 22843982

37. 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. https://doi.org/10.1038/ng.2711 PMID: 23933821

38. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013; 9: e1003264. https://doi.org/10.1371/journal.pgen.1003264 PMID: 23408905

39. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014; 24: 1550–7. https://doi.org/10.1101/gr.169375.113 PMID: 24963154

40. Golan D, Rosset S. Effective genetic-risk prediction using mixed models. Am J Hum Genet. 2014; 95: 383–93. https://doi.org/10.1016/j.ajhg.2014.09.007 PMID: 25279982

41. 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. https://doi.org/10.1016/j.ajhg.2015.09.001 PMID: 26430803

42. 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. https://doi.org/10.1038/nrg2813 PMID: 20548291

43. 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. https://doi.org/10.1038/ng.3190 PMID: 25642633

44. Loh P-R, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nat Genet. 2018; 50: 906–908. https://doi.org/10.1038/s41588-018-0144-6 PMID: 29892013

45. 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. https://doi.org/10.1038/s41588-018-0184-y PMID: 30104761

46. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018; 50: 1593–1599. https://doi.org/10.1038/s41588-018-0248-z PMID: 30349118

47. 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.

48. 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 refinements. J Anim Sci. 2015; 93: 4670–4674. https://doi.org/10.2527/jas.2015-9395 PMID: 26523559

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

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

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

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

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

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

55. 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. https://doi.org/10.1038/ng.3404 PMID: 26414678

56. Zhou X. A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann Appl Stat. 2017; 11: 2027–2051. https://doi.org/10.1214/17-AOAS1052 PMID: 29515717

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

58. Wu Y, Sankararaman S. A scalable estimator of SNP heritability for biobank-scale data. Bioinformatics. 2018; 34: i187–i194. https://doi.org/10.1093/bioinformatics/bty253 PMID: 29950019

59. Wu Y, Yaschenko A, Heydary MH, Sankararaman S. Fast estimation of genetic correlation for Biobank-scale data. bioRxiv. 2019; 525055.

60. Pazokitoroudi A, Wu Y, Burch KS, Hou K, Pasaniuc B, Sankararaman S. Scalable multi-component linear mixed models with application to SNP heritability estimation. bioRxiv. 2019; 522003.

61. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017; 18: 117–127. https://doi.org/10.1038/nrg.2016.142 PMID: 27840428

62. Hu Z, Yang R-C. Marker-Based Estimation of Genetic Parameters in Genomics. Cai X, editor. PLoS ONE. 2014; 9: e102715. https://doi.org/10.1371/journal.pone.0102715 PMID: 25025305

63. Liu H, Chen G-B. A fast genomic selection approach for large genomic data. Theor Appl Genet. 2017; 130: 1277–1284. https://doi.org/10.1007/s00122-017-2887-3 PMID: 28389770

64. Liu H, Chen G-B. A new genomic prediction method with additive-dominance effects in the least-squares framework. Heredity. 2018; 121: 196–204. https://doi.org/10.1038/s41437-018-0099-5 PMID: 29925888

65. Haseman J, Elston R. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet. 1972; 2: 3–19. PMID: 4157472

66. Chen G-B. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman–Elston regression. Front Genet. 2014; 5: 107. https://doi.org/10.3389/fgene.2014.00107 PMID: 24817879

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

68. 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. https://doi.org/10.1371/journal.pone.0080821 PMID: 24339886

69. 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.

70. 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.

71. 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. https://doi.org/10.1631/jzus.2007.B0815 PMID: 17973343

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

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

74. 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. PMID: 17247525

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

76. 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. https://doi.org/10.3168/jds.S0022-0302(91)78204-0 PMID: 2045563

77. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika. 1971; 58: 545–554.

78. 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.

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

80. 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.

81. Lidauer M, Matilainen K, Mäntysaari E, Pitkänen T, Taskinen M, Strandén I, et al. MiX99: Technical reference guide for MiX99 solver. 2017;
82. Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics. 1995; 51: 1440–1450.
83. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput. 1995; 16: 1190–1208.
84. McCulloch CE, Searle SR, Neuhaus JM. Generalized, Linear, and Mixed Models. 2nd ed. Wiley Series in Probability and Statistics; 2008.
85. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015; 16: 33–44. https://doi.org/10.1038/nrg3821 PMID: 25404112
86. Wright S. Coefficients of inbreeding and relationship. Am Nat. 1922; 56: 330–338.
87. Henderson C. Best linear unbiased prediction of nonadditive genetic merits. J Anim Sci. 1985; 60: 111–117.
88. Sorensen DC. Implicitly Restarted Arnoldi/Lanczos Methods for Large Scale Eigenvalue Calculations. Parallel Numerical Algorithms. Dordrecht: Springer Netherlands; 1997. pp. 119–165.
89. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38: 904–909. https://doi.org/10.1038/ng1847 PMID: 16862161
90. 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. https://doi.org/10.1016/j.ajhg.2007.10.009 PMID: 18252216
91. Wolak ME, Reid JM. Accounting for genetic differences among unknown parents in microevolutionary studies: how to include genetic groups in quantitative genetic animal models. J Anim Ecol. 2017; 86: 7–20. https://doi.org/10.1111/1365-2656.12597 PMID: 27731502
92. Kreider RM, Ellis R. Number, timing, and duration of marriages and divorces, 2009. US Department of Commerce, Economics and Statistics Administration, US Census Bureau; 2011.
93. Vespa J, Lewis J, Kreider R. America's families and living arrangements: 2012. Am Community Surv. 2011;
94. 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.
95. 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/
96. Therneau TM. coxme: Mixed Effects Cox Models [Internet]. 2018. Available: https://CRAN.R-project.org/package=coxme
97. Feldman MW, Ramachandran S. Missing compared to what? Revisiting heritability, genes and culture. Philos Trans R Soc B Biol Sci. 2018; 373: 20170064.
98. Kosova G, Abney M, Ober C. Heritability of reproductive fitness traits in a human population. Proc Natl Acad Sci. 2010; 107: 1772–1778. https://doi.org/10.1073/pnas.0906196106
99. Buhmann MD. A new class of radial basis functions with compact support. Math Comput. 2000; 70: 307–319.
100. Gneiting T. Correlation functions for atmospheric data analysis. Q J R Meteorol Soc. 1999; 125: 2449–2464.
101. Gaspari G, Cohn SE. Construction of correlation functions in two and three dimensions. Q J R Meteorol Soc. 1999; 125: 723–757.
102. Gneiting T. Compactly supported correlation functions. J Multivar Anal. 2002; 83: 493–508.
103. Sansò F, Schuh W-D. Finite covariance functions. Bull Géod. 1987; 61: 331–347.
104. Wendland H. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Adv Comput Math. 1995; 4: 389–396.
105. Bijma P. Estimating maternal genetic effects in livestock. J Anim Sci. 2006; 84: 800–806. PMID: 16543556
106. 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. https://doi.org/10.1126/science.aan6877 PMID: 29371463

107. 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. https://doi.org/10.1038/s41588-018-0178-9 PMID: 30104764

108. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009; 92: 4656–4663. https://doi.org/10.3168/jds.2009-2061 PMID: 19700729

109. 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. https://doi.org/10.3168/jds.2009-2730 PMID: 20105546

110. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010; 42: 2. https://doi.org/10.1186/1297-9686-42-2 PMID: 20105297

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

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