Bioinformatics, 36(24), 2020, 5632–5639
doi: 10.1093/bioinformatics/btaa1079
Advance Access Publication Date: 26 December 2020
Original Paper

Genetics and population analysis

A non-linear regression method for estimation of gene–environment heritability

Matthew Kerin[1] **and Jonathan Marchini**[2,*]

[1]Wellcome Trust Center for Human Genetics, Oxford, OX3 7BN, UK and [2]Regeneron Genetics Center, Tarrytown, NY 10591, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on April 30, 2020; revised on November 27, 2020; editorial decision on December 11, 2020; accepted on December 16, 2020

Abstract

Motivation: Gene–environment (GxE) interactions are one of the least studied aspects of the genetic architecture of human traits and diseases. The environment of an individual is inherently high dimensional, evolves through time and can be expensive and time consuming to measure. The UK Biobank study, with all 500 000 participants having undergone an extensive baseline questionnaire, represents a unique opportunity to assess GxE heritability for many traits and diseases in a well powered setting.

Results: We have developed a randomized Haseman–Elston non-linear regression method applicable when many environmental variables have been measured on each individual. The method (GPLEMMA) simultaneously estimates a linear environmental score (ES) and its GxE heritability. We compare the method via simulation to a whole-genome regression approach (LEMMA) for estimating GxE heritability. We show that GPLEMMA is more computationally efficient than LEMMA on large datasets, and produces results highly correlated with those from LEMMA when applied to simulated data and real data from the UK Biobank.

Availability and implementation: Software implementing the GPLEMMA method is available from https://jmarchini.org/gplemma/.

Contact: jonathan.marchini@regeneron.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advent of genome-wide association studies (The Wellcome Trust Case Control Consortium, 2007) has catalyzed a huge number of discoveries linking genetic markers to many human complex diseases and traits. For the most part, these discoveries have involved common variants that confer relatively small amounts of risk and only account for a small proportion of the phenotypic variance of a trait (Manolio *et al.*, 2009). This has led to a surge of interest in methods and applications that measure the joint contribution to phenotypic variance of all measured variants throughout the genome (SNP heritability), and in testing individual variants within this framework. Most notably the seminal paper of Yang et al. (2010), who used a linear mixed model (LMM) to show that the majority of missing heritability for height could be explained by genetic variation by common SNPs (Yang *et al.*, 2010). When testing variants for association these LMMs can reduce false positive associations due to population structure, and improve power by implicitly conditioning on other loci across the genome (Listgarten *et al.*, 2012; Loh *et al.*, 2015; Yang *et al.*, 2014). These methods model the unobserved polygenic contribution as a multivariate Gaussian with covariance structure proportional to a genetic relationship matrix (GRM) (Eskin *et al.*, 2008; Lippert *et al.*, 2011; Zhou and Stephens, 2012). This approach is mathematically equivalent to a whole genome regression (WGR) model with a Gaussian prior over SNP effects (Listgarten *et al.*, 2012).

Subsequent research has shown that the simplest LMMs make assumptions about the relationship between minor allele frequency (MAF), linkage disequilibrium (LD) and trait architecture that may not hold up in practice (Evans *et al.*, 2018; Speed *et al.*, 2012) and generalizations have been proposed that stratify variance into different components by MAF and LD (Speed *et al.*, 2012, 2017; Yang *et al.*, 2015). Other flexible approaches have been proposed in both the animal breeding (de los Campos *et al.*, 2013; Hayes *et al.*, 2001) and human literature (Carbonetto and Stephens, 2012; Logsdon *et al.*, 2010; Zhou *et al.*, 2013) to allow different prior distributions that better capture SNPs of small and large effects. For example, a mixture of Gaussians (MoG) prior can increase power to detect associated loci in some (but not all) complex traits (Carbonetto and Stephens, 2012; Loh *et al.*, 2015). Other methods have been proposed that estimate heritability only from summary statistics and LD reference panels (Bulik-Sullivan *et al.*, 2015; Speed and Balding, 2019). Heritability can also be estimated using Haseman–Elston regression (Haseman and Elston, 1972) and has recently been extended using a randomized approach (Wu and Sankararaman, 2018) that has $\mathcal{O}(NM)$ computational complexity and works for multiple variance components (Pazokitoroudi *et al.*, 2020). Other recent work has shown that LMM approaches such as these are not able to disentangle direct and indirect genetic effects, the balance of which will vary depending on the trait being studied (Young *et al.*, 2018).

There has been less exploration of methods for estimating heritability that account for gene–environment interactions. One interesting approach has proposed using spatial location as a surrogate for environment (Heckerman *et al.*, 2016) using a three component LMM—one based on genomic variants, one based on measured spatial location as a proxy for environmental effects, and a gene–environment component, modelled as the Hadamard product of the genomic and spatial covariance matrices. Other authors have used this method to account for gene-gene interactions (Crawford *et al.*, 2017; Ober *et al.*, 2015).

Modelling gene–environment interactions when many different environmental variables are measured is a more challenging problem. If several environmental variables drive interactions at individual loci, or if an unobserved environment that drives interactions is better reflected by a combination of observed environments, it can make sense to include all variables in a joint model. StructLMM (Moore *et al.*, 2019) focuses on detecting GxE interactions at individual markers. Environmental similarity between individuals is modelled (over multiple environments) as a random effect, and each SNP is tested independently for GxE interactions. However, this approach does not model the genome wide contribution of all the markers, which is often a major component of phenotypic variance.

We recently proposed a WGR approach called LEMMA applicable to large human datasets such as UK Biobank, where many potential environmental variables are available (Kerin and Marchini, 2020). The LEMMA regression model includes main effects of each genotyped SNP across the genome, and also interactions of each SNP with a environmental score (ES), that is a linear combination of the environmental variables. The ES is estimated as part of the method using a Variational Bayes algorithm to fit the WGR model. The model uses mixture of Gaussian (MoG) priors on main and GxE SNP effects, that allow for a range of different genetic architectures from polygenic to sparse genetic effects (Carbonetto and Stephens, 2012; Logsdon *et al.*, 2010; Zhou *et al.*, 2013). The ES can be readily interpreted and its main use is to test for GxE interactions one variant at a time, typically at a larger set of imputed SNPs in the dataset. However, the ES can also be used to estimate the proportion of phenotypic variability that is explained by GxE interactions (SNP GxE heritability), using a two component randomized Haseman–Elston (RHE) regression (Pazokitoroudi *et al.*, 2020).

The main contribution of this article is to combine the estimation of the LEMMA ES into a stand-alone RHE framework. This results in a non-linear optimization problem that we solve using the Levenburg-Marquardt (LM) algorithm. The method implicitly assumes a Gaussian prior on main effect and GxE effect sizes. We also propose a separate RHE method that estimates the independent GxE contribution of each measured environmental variable. We set out the differences between these two models and present a simulation study to compare them to LEMMA. We show that GPLEMMA is more computationally efficient than LEMMA on large datasets. We also apply the method to UK Biobank data and show that GPLEMMA produces estimates very close to LEMMA. Software implementing the GPLEMMA algorithm in C++ is available at https://jmarchini.org/gplemma/.

2 Materials and methods

2.1 Modelling SNP heritability

The simplest model for estimating SNP heritability has the form

$$y = X\beta + e, \quad \beta_l \sim \mathcal{N}\left(0, \frac{\sigma_g^2}{M}\right), \quad e \sim \mathcal{N}\left(0, \sigma_e^2\right)$$

where y is a continuous phenotype, X is an $N \times M$ matrix of genotypes that has been normalized to have column mean zero and column variance one, and β is an M-vector of SNP effect sizes. Integrating out β leads to the variance component model

$$y \sim \mathcal{N}\left(0, \sigma_g^2 K + \sigma_e^2 I\right),$$

where $K = \frac{XX^T}{M}$ is known as the genomic relationship matrix (GRM) (Yang *et al.*, 2010). Estimating the two parameters in this model σ_g and σ_e leads to an estimate of SNP heritability of $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$. This is commonly referred to in the literature as the single component model. Subsequent research has shown that the single component model makes assumptions about the relationship between minor allele frequency (MAF), linkage disequilibrium (LD) and trait architecture that may not hold up in practice (Evans *et al.*, 2018; Speed *et al.*, 2012). There have been many follow up methods, including; generalizations that stratify variance into different components by MAF and LD (Yang *et al.*, 2015), approaches that assign different weights for the GRM (Speed *et al.*, 2012, 2017), methods that replace the Gaussian prior on β with a spike and slab on SNP effect sizes (Powell *et al.*, 2018) and methods that estimate heritability only from summary statistics and LD reference panels (Bulik-Sullivan *et al.*, 2015; Finucane *et al.*, 2015).

2.2 Randomized Haseman–Elston regression

An alternative method used to compute heritability is known as Haseman–Elston (HE) regression (Haseman and Elston, 1972). HE-regression is a method of moments (MoM) estimator that optimizes variance components (σ_g^2, σ_e^2) in order to minimize the squared difference between the observed and expected trait covariances. The MoM estimator $(\hat{\sigma}_g^2, \hat{\sigma}_e^2)$ can be obtained by solving the minimization

$$\arg\min_{\sigma_g^2, \sigma_e^2} ||yy^T - (\sigma_g^2 K + \sigma_e^2 I)||_F^2$$

or equivalently by solving the linear regression problem

$$\text{vec}(yy^T) = \sigma_g^2 \text{vec}(K) + \sigma_e^2 \text{vec}(I) + \epsilon'$$

where vec(A) is the vectorization operator that transforms an $N \times M$ matrix into an NM-vector. In matrix format, both of these forms correspond to solving the following linear system

$$\begin{pmatrix} \text{tr}(K^2) & \text{tr}(K) \\ \text{tr}(K) & N \end{pmatrix} \begin{pmatrix} \sigma_\beta^2 \\ \sigma_e^2 \end{pmatrix} = \begin{pmatrix} y^T K y \\ y^T y \end{pmatrix} \quad (1)$$

HE-regression methods are widely acknowledged to be more computationally efficient (Golan *et al.*, 2014; Wu and Sankararaman, 2018; Yang *et al.*, 2017) and do not require any assumptions on the phenotype distribution beyond the covariance structure (Golan *et al.*, 2014) (in contrast to maximum-likelihood estimators). However, HE-regression based estimates typically have higher variance (Yang *et al.*, 2017), thus implying that they have less power.

Recent method developments (Pazokitoroudi *et al.*, 2020; Wu and Sankararaman, 2018) have shown that a randomized HE-regression (RHE) approach can be used to compute efficiently on genetic datasets with hundreds of thousands of samples. Wu and Sankararaman (2018) observed that Equation (1) can be solved efficiently without ever having to explicitly compute the kinship matrix K by using Hutchinson's estimator (Hutchinson, 1990), which states that $\text{tr}(A) = \mathbb{E}[z^T A z]$ for any matrix where z is a random vector with mean zero and covariance given by the identity matrix. The proposed method involves approximating tr(K) and tr(K^2) using only matrix vector multiplications with the genotype matrix X, to compute the following expressions

$$\text{tr}(K) \approx \frac{1}{B}\frac{1}{M^2}\sum_b ||X^T z_b||_2^2,$$
$$\text{tr}(K^2) \approx \frac{1}{B}\frac{1}{M^2}\sum_b ||XX^T z_b||_2^2.$$

Thus an approximate solution can be obtained in $\mathcal{O}(NMB)$ time, where B denotes a relatively small number of random samples. Subsequent work by Pazokitoroudi *et al.* extended this approach to a multiple component model (Pazokitoroudi *et al.*, 2020)

$$y \tilde{} \mathcal{N}\left(0, \sum_k \sigma_k^2 K_k + I\sigma_e^2\right).$$

With parameter estimates obtained as solution to the linear system given by

$$\begin{pmatrix} T & b \\ b^T & N \end{pmatrix} \begin{pmatrix} \sigma_\beta^2 \\ \sigma_e^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{c} \\ N \end{pmatrix}, \qquad (2)$$

where $T_{kl} = \mathrm{tr}(K_k K_l)$, $b_k = \mathrm{tr}(K_k)$ and $c_k = y^T K_k y$. Finally, both papers show how to efficiently control for covariates by projecting them out of all terms in the system of equations. Thus with covariates included the multiple component model becomes

$$y \tilde{} \mathcal{N}\left(C\alpha, \sum_k \sigma_k^2 K_k + I\sigma_e^2\right),$$

and terms in the subsequent linear system are given by $T_{kl} = \mathrm{tr}(WK_k WK_l W)$, $b_k = \mathrm{tr}(WK_k W)$ and $c_k = y^T WK_k Wy$, where $W = I_N - C^T(C^T C)^{-1}C$. The GPLEMMA and MEMMA approaches developed in this article use this method of handling covariates.

2.3 Modelling GxE heritability

We introduce two extensions of the RHE framework for modelling GxE interactions with multiple environmental variables. In both models we let E be an $N \times L$ matrix of environmental variables and C be an $N \times D$ matrix of covariates, where both matrixes are normalized such that columns have mean zero and variance one. We use notation E_l to denote the lth column of E, and $\mathrm{diag}(x)$ denotes a null matrix with elements of vector x on its diagonal. We note that columns of E are always included in C, so that $D > L$.

2.3.1 MEMMA

The first model assumes that each environmental variable interacts independently with the genome

$$y = C\alpha + X\beta + \sum_l (E_l \odot X)\lambda_l + \epsilon, \qquad (3)$$

where $\beta \tilde{} \mathcal{N}\left(0, \frac{\sigma_\beta^2}{M} I_M\right), \lambda_l \tilde{} \mathcal{N}\left(0, \frac{\sigma_{w_l}^2}{M} I_M\right), \epsilon \tilde{} \mathcal{N}(0, \sigma_e^2 I_N)$ and $E_l \odot X$ denotes the element-wise product of E_l with each column of X. Integrating out β and λ leads to the variance component model

$$y \tilde{} \mathcal{N}\left(C\alpha, \sum_{k=1}^{L+2} \theta_k K_k\right),$$

where $\theta = \left\{\sigma_\beta^2, (\sigma_{w_l}^2)_{l=1}^L, \sigma_e^2\right\}$, $F_k = E_k \odot X$ and

$$K_k = \begin{cases} \dfrac{XX^T}{M} & \text{if } k = 1, \\ \dfrac{F_{k-1}F_{k-1}^T}{M} & \text{if } 1 < k \leq L+1, \\ I & \text{if } k = L+2. \end{cases}$$

Fitting the variance components is done analytically by solving the system of equations $T\theta = c$ where $T_{kl} = \mathrm{tr}(WK_k WK_l W)$, $c_k = y^T WK_k Wy$ and $W = I_N - C(C^T C)^{-1}C^T$. As shown in the original RHE method (Pazokitoroudi *et al.*, 2020; Wu and Sankararaman, 2018), Hutchinson's estimator can be used to efficiently estimate T_{kl}. To do this our software streams SNP markers from a file and computes $y^T WXX^T Wy$ and the following N-vectors

$$u_b = XX^T Wz_b, \qquad (4)$$

$$v_{b,l} = XX^T E_l Wz_b, \qquad (5)$$

where $z_b \tilde{} \mathcal{N}(0, I_N)$ for $1 \leq b \leq B$ are random N-vectors. Then

$$T_{kl} = \frac{1}{M^2 B} \sum_b (\xi_b^k)^T \xi_b^k,$$

where ξ_b^k is defined as

$$\xi_b^k = \begin{cases} u_b & \text{if } k = 1, \\ v_{b,l} & \text{if } 1 < k \leq L+1, \\ z_b & \text{if } k = L+2. \end{cases}$$

Finally, the variance components are converted to heritability estimates using the following formula

$$\hat{h}_k^2 = \frac{\hat{\theta}_k \mathrm{tr}(K_k)}{\sum_k \hat{\theta}_k \mathrm{tr}(K_k)}.$$

We call this approach MEMMA (Multiple Environment Mixed Model Analysis). MEMMA costs $\mathcal{O}(NMLB)$ in compute and $\mathcal{O}(NLB)$ in RAM.

2.3.2 GPLEMMA

The second model involves the estimating a linear combination of environments, or environmetal score (ES), that interacts with the genome. The model is given by

$$y = C\alpha + X\beta + (\eta \odot X)\gamma + \epsilon, \qquad (6)$$

$$\beta \tilde{} \mathcal{N}\left(0, \frac{\sigma_\beta^2}{M} I_M\right), \qquad (7)$$

$$\gamma \tilde{} \mathcal{N}\left(0, \frac{\sigma_\gamma^2}{M} I_M\right), \qquad (8)$$

where $\eta = Ew$ is a column vector that we refer to as the linear environmental score (ES). This is the same model used by LEMMA (Kerin and Marchini, 2020), which we include below for completeness, except the mixture of Gaussians priors on SNP effects (β and γ) have been replaced with Gaussian priors. For this reason, we call this approach GPLEMMA (Gaussian Prior Linear Environment Mixed Model Analysis). Integrating out the SNP effects yields the model

$$y \tilde{} \mathcal{N}\left(C\alpha, \sigma_\beta^2 K + \sigma_\gamma^2 K_2(w) + \sigma_e^2 I\right),$$

where $K_2(w) = \mathrm{diag}(Ew)K\mathrm{diag}(Ew) = \frac{1}{M}\sum_{l,m} w_l w_m F_l F_m^T$ and $F_l = E_l \odot X$. Minimizing the squared loss between the expected and observed covariance is equivalent to the following regression problem

$$\mathrm{vec}(yy^T) = \sigma_\beta^2 \mathrm{vec}(K) + \sum_{l,m} \sigma_\gamma^2 w_l w_m \mathrm{vec}(F_l F_m^T) + \sigma_e^2 \mathrm{vec}(I) + \epsilon'. \quad (9)$$

In this format it is clear that optimizing $\sigma_\beta^2, \sigma_\gamma^2, w, \sigma_e^2$ is a non-linear regression problem. Further, including a parameter for σ_γ^2 is no longer necessary. From here on we set $\tilde{w}_l = \sqrt{\sigma_\gamma^2} w_l$ and drop the parameterization without loss of generality.

2.3.3 Levenburg–Marquardt algorithm

We use the Levenburg-Marquardt (LM) algorithm (Zolfaghari *et al.*, 2005), which is commonly used for non-linear least squares problems. The algorithm effectively interpolates between the Gauss-Newton algorithm and the method of steepest gradient descent, by use of an adaptive damping parameter. In this manner, it is more robust than the straight forward Gauss-Newton algorithm but should have faster convergence than a gradient descent approach.

Without loss of generality, consider the model

$$Y = f(\theta) + \epsilon, \qquad (10)$$

where $f(\theta)$ is a function that is non-linear in the parameters θ. Given a starting point θ_0, LM proposes a new point $\theta_{\mathrm{new}} = \theta_0 + \delta$ by solving the normal equations

$$(J(\theta_0)^T J(\theta_0) + \mu I)\delta = J(\theta_0)^T \epsilon(\theta_0), \tag{11}$$

where $J(\theta_0) = \frac{\delta f(\theta_0)}{\delta \theta_0}$ and $\epsilon(\theta_0) = Y - f(\theta_0)$ are respectively the Jacobian and the residual vector evaluated at θ_0.

If θ_{new} has lower squared error than θ_0, then the step is accepted and the adaptive damping parameter μ is reduced. Otherwise, μ is increased and a new step δ is proposed. For small values of μ Equation (11) approximates the quadratic step appropriate for a fully linear problem, whereas for large values of μ Equation (11) behaves more like steepest gradient descent. This allows the algorithm to defensively navigate regions of the parameter space where the model is highly non-linear. If $\theta + \delta$ reduces the squared error, then the step is accepted and μ is reduced, otherwise μ is increased and a new step δ is proposed.

In summary the LM algorithm requires computation of the matrices $J(\theta)^T J(\theta)$, $J(\theta)^T \epsilon(\theta)$ at each step, as well as the squared error (which we define as $S(\theta)$). We now give statements of the equations used to compute each of these values, and show that each iteration can be performed in $\mathcal{O}(NL^2B)$ time.

We apply the LM algorithm with $\theta = \left\{\sigma^2_\beta, w, \sigma^2_e\right\}$, $Y = \text{vec}(yy^T)$ and

$$f(\theta) = \sigma^2_\beta \text{vec}(K) + \sum_{l,m} w_l w_m \text{vec}(F_l F_m^T) + \sigma^2_e \text{vec}(I).$$

Several quantities can be pre-calculated and re-used in the LM algorithm. The N-vectors u_b, $v_{b,l}$ and $y^T WXX^T Wy$ are needed and have been defined above. In addition, GPLEMMA also benefits from the pre-calculation of

$$H_{l,m} = E_l^T \text{diag}(Wy) XX^T \text{diag}(Wy) E_m, \quad 1 \leq l, m \leq L$$

which can also be computed as genotypes are streamed from file.

Let $(J^T J)_{\theta_i,\theta_j}$ denote the entry of the J^{TJ} that corresponds to $\frac{f(\theta)^T}{\partial\theta_i}\frac{f(\theta)}{\partial\theta_i}$ for $\theta_i, \theta_j \in \left\{w, \sigma^2_\beta, \sigma_e\right\}$ and define the N-vector $v_b(w) = \sum_l w_l v_{b,l}$. Then the $(L+2) \times (L+2)$ matrix $J(\theta)^T J(\theta)$ is given by

$$\begin{aligned}
\left(J^T J\right)_{w_l,w_m} &= \text{tr}\left(\text{diag}(\eta) K \text{diag}(E_l) \text{diag}(E_m) K \text{diag}(\eta)\right),\\
&= \frac{1}{M^2B} \sum_b \left(v_b(w)^T \text{diag}(E_l) \text{diag}(E_m) v_b(w)\right),\\
\left(J^T J\right)_{w_l,\sigma^2_\beta} &= \text{tr}\left(\text{diag}(\eta) K \text{diag}(E_l) K\right) = \frac{1}{M^2B} \sum_b \left(v_b(w)^T \text{diag}(E_l) u_b\right),\\
\left(J^T J\right)_{\sigma^2_\beta,\sigma^2_\beta} &= \text{tr}(KK) = \frac{1}{M^2B} \sum_b ||u_b||_2^2,\\
\left(J^T J\right)_{\sigma^2_\beta,\sigma^2_e} &= \text{tr}(K) = \frac{1}{M^2B} \sum_b z_b^T W u_b,\\
\left(J^T J\right)_{w_l,\sigma^2_e} &= \text{tr}\left(\text{diag}(\eta) K \text{diag}(E_l)\right) = \frac{1}{M^2B} \sum_b z_b^T W v_b(w),\\
\left(J^T J\right)_{\sigma^2_e,\sigma^2_e} &= \text{tr}(W).
\end{aligned}$$

$J(\theta)^T \epsilon(\theta)$ is given by

$$\begin{aligned}
(J(\theta)^T \epsilon(\theta))_{\sigma^2_\beta} &= \text{tr}(y^T WKWy) - J(\theta)^T J(\theta) \sigma^2_\beta,\\
(J(\theta)^T \epsilon(\theta))_{w_l} &= \text{tr}(y^T W \text{diag}(E_l) K \text{diag}(Ew) Wy) - J(\theta)^T J(\theta) w_l,\\
(J(\theta)^T \epsilon(\theta))_{\sigma^2_e} &= \text{tr}(y^T Wy) - J(\theta)^T J(\theta) \sigma^2_e.
\end{aligned}$$

where

$$\text{tr}(y^T W \text{diag}(E_l) K \text{diag}(Ew) Wy) = \sum_m H_{l,m}$$

Finally the squared error, which we define as $S(\theta)$, is given by

$$\begin{aligned}
S\left(\sigma^2_\beta, w\right) &= ||\left(yy^T - \text{Cov}(y)\right)||_F^2,\\
&= \text{tr}\left(\left(yy^T - \text{Cov}(y)\right)\left(yy^T - \text{Cov}(y)\right)\right),\\
&= \text{tr}(yy^T yy^T) - 2 \begin{pmatrix} \sigma^2_\beta \\ 1 \\ \sigma^2_e \end{pmatrix}^T \begin{pmatrix} \text{tr}(y^T Ky) \\ \text{tr}\left(y^T K_2(w) y\right) \\ \text{tr}(y^T y) \end{pmatrix}\\
&+ \begin{pmatrix} \sigma^2_\beta \\ 1 \\ \sigma^2_e \end{pmatrix}^T \begin{pmatrix} \text{tr}(KK) & \text{tr}\left(KK_2(w)\right) & \text{tr}(K) \\ \text{tr}\left(KK_2(w)\right) & \text{tr}\left(K_2(w) K_2(w)\right) & \text{tr}\left(K_2(w)\right) \\ \text{tr}(K) & \text{tr}\left(K_2(w)\right) & N \end{pmatrix} \begin{pmatrix} \sigma^2_\beta \\ 1 \\ \sigma^2_e \end{pmatrix}
\end{aligned}$$

where

$$\text{tr}(K_2(w) K_2(w)) \approx \frac{1}{M^2B} \sum_b ||v_b(w)||_2^2$$

The initial preprocessing step has costs $\mathcal{O}(NMLB + NML^2)$ in compute and $\mathcal{O}(NLB)$ in RAM. The remaining algorithm does not require much RAM in addition to that required in the preprocessing step, so also costs $\mathcal{O}(NLB)$ in RAM. Construction of the summary variable $v_b(w) = \sum_l w_l v_{b,l}$ costs $\mathcal{O}(NLB)$ in compute. Each iteration of the LM algorithm costs $\mathcal{O}(NL^2B)$.

It is possible to parallelize GPLEMMA using OpenMPI by partitioning samples across cores, in a similar manner to that used by LEMMA (Kerin and Marchini, 2020). Given that evaluating the objective function $S(\sigma^2_\beta, w)$ is characterized by BLAS level 1 array operations, a distributed algorithm using OpenMPI should have superior runtime versus an the same algorithm using OpenMP as well as providing RAM limited only by the size of a researchers compute cluster.

We perform 10 repeats of the LM algorithm with different initializations, and keep results from the solution with lowest squared error $S(\hat{\theta})$. Each run is initialized with a vector of interaction weights $\tilde{w}$, where each entry set to $\frac{1}{L}$ and a small amount of Gaussian noise is added.

$$\tilde{w} = \frac{1}{L}\vec{1} + \mathcal{N}\left(0, \frac{2}{L^2} I_L\right).$$

To transform the initial weights vector $\tilde{w}$ to the initial parameters θ_0 we let $\left(\hat{\sigma}^2_\beta, \hat{\sigma}^2_\gamma, \hat{\sigma}^2_e\right)$ be solutions to

$$\left(\hat{\sigma}^2_\beta, \hat{\sigma}^2_\gamma, \hat{\sigma}^2_e\right) = \min_{\sigma^2_\beta, \sigma^2_\gamma, \sigma^2_e} ||yy^T - \left(\sigma^2_\beta K + \sigma^2_\gamma K_2(\tilde{w}) + \sigma^2_e I\right)||_F^2.$$

The GPLEMMA algorithm is then initialized with $\theta_0 = \left(\hat{\sigma}^2_\beta, w, \hat{\sigma}^2_e\right)$ where $w = \sigma_\gamma \tilde{w}$.

2.3.4 LEMMA

The LEMMA model is given by

$$y = C\alpha + X\beta + (\eta \odot X)\gamma + \epsilon, \tag{12}$$

$$\beta \tilde{} \psi \mathcal{N}\left(0, \frac{\sigma^2_{\beta 1}}{M} I_M\right) + (1 - \psi) \mathcal{N}\left(0, \frac{\sigma^2_{\beta 2}}{M} I_M\right), \tag{13}$$

$$\gamma \tilde{} \pi \mathcal{N}\left(0, \frac{\sigma^2_{\gamma 1}}{M} I_M\right) + (1 - \pi) \mathcal{N}\left(0, \frac{\sigma^2_{\gamma 2}}{M} I_M\right), \tag{14}$$

where $\eta = Ew$ is the linear environmental score (ES). The use of the MoG priors makes it harder to analytically integrate out the parameters. The LEMMA algorithm uses a Variational Bayes approach to first estimate the ES and fit the whole genome regression pararmeters β and γ. The primary use of the LEMMA model is for testing individual SNPs for GxE effects. However, LEMMA can also estimate GxE heritability. It uses the VB algorithm to estimate the ES, and then plugs this estimate into Eq 6 and uses a randomized Haseman–Elston linear regression to estimate the variance components.

2.3.5 Relationship between MEMMA and GPLEMMA

Comparing Equation (3) with Section 6, suggests that the GPLEMMA model can be expressed at the MEMMA model with the added constraint that

$$\Lambda = w\gamma^T$$

where $\Lambda = [\lambda_1, \dots, \lambda_L]^T$ is the $L \times M$ matrix of GxE effect sizes in MEMMA for the L environments and M SNPs.

We can expect the two models to give similar heritability estimates, under the simplifying assumptions that GxE interactions do occur with a single linear combination of the environments and that the set of random variables $\left\{g, (E_l \odot g)_{l=1}^{L}\right\}$ is mutually independent. Let $g \sim \mathcal{N}(0, K)$ and $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I)$. Then connection between the two models is revealed by observing

$$\begin{aligned} y &= \mathcal{N}\left(C\alpha, \sigma_\beta^2 K + \sigma_\gamma^2 K_2(w) + \sigma_e^2 I\right), \\ &= \sigma_\beta g + \left(\sigma_\gamma \textstyle\sum_l w_l E_l\right) \odot g + \epsilon, \\ &= \sigma_\beta g + \sum_l \sigma_{w_l} E_l \odot g + \epsilon, \\ &= \mathcal{N}\left(0, \sigma_\beta^2 K + \textstyle\sum_l \sigma_{w_l}^2 E_l \odot K \odot E_l^T + \sigma_e^2 I\right), \end{aligned}$$

where $\sigma_{w_l}^2 = \sigma_\gamma^2 w_l^2$. Thus we should expect both models to have the same estimate for the proportion of variance explained by GxE interaction effects.

Even in the case that MEMMA and GPLEMMA have the same expected heritability estimate, there are still some differences between the two. GPLEMMA is a constrained model, so the variance of its heritabiity estimates may be smaller. Further, although $\hat{\sigma}_{w_l}^2$ is proportional to the square of the weights used to construct the ES the sign of the interaction weight w_l has been lost. Thus it is not possible to reconstruct an ES for use in single SNP testing using MEMMA.

2.4 Simulated data

We carried out simulation studies to compare the performance of MEMMA, LEMMA and GPLEMMA in a variety of different scenarios, using three different phenotype simulation strategies; a baseline phenotype that was simulated according to the LEMMA model (Kerin and Marchini, 2020), a phenotype which generalizes the LEMMA model to three orthogonal ESs and a misspecified phenotype that had squared dependancy on a heritable environmental variable S.

The simulations use real data subsampled from genotyped SNPs in the UK Biobank (Bycroft *et al.*, 2018), drawing SNPs from all 22 chromosomes in proportion to chromosome length and using unrelated samples of mixed ancestry ($N = 25$k; 12 500 white British, 7500 Irish and 5000 white European, $N = 50$k; 29 567 white British, 7500 Irish and 12 568 white European, $N = 100$k; 79 567 white British, 7500 Irish and 12 568 white European; using self-reported ancestry in field *f.21000.0.0*). All samples were genotyped using the UKBB genotype chip and were included in the internal principal component analysis performed by the UK Biobank. Environmental variables were simulated from a standard Gaussian distribution.

Let N be the number of individuals, M the total number of SNPs, L the total number of environmental variables and h_{G}^2 and h_{GxE}^2 the herirtability of main effects and GxE effects. The baseline phenotype was simulated using the model

$$\begin{aligned} y &= C\alpha + X\beta + (\eta \odot X)\gamma + \epsilon, \\ \eta &= Ew, \\ \epsilon &\sim \mathcal{N}(0, I), \end{aligned}$$

where X represents the $N \times M$ genotype matrix after columns have been standardized to have mean zero and variance one, C is the first principle component of the genotype matrix and E is the $N \times L$ matrix of environmental variables. In all simulations α was set such that $C\alpha$ explained one percent of trait variance. The interaction weight vector w contained L^{active} non-zero elements, which were drawn from a decreasing sequence

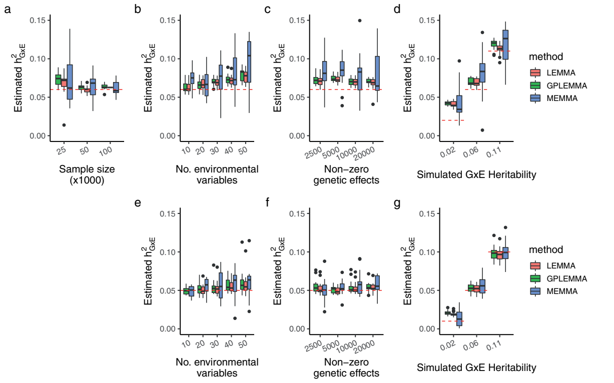

Fig. 1. PVE estimation. Estimates of the proportion of variance explained by GxE effects by LEMMA, MEMMA and GPLEMMA whilst varying the number of environments, the number of active environments, the number of non-zero SNP effects and GxE heritability. All simulations constructed with the baseline phenotype. **a–d** Simulations results using $N = 25K$ samples and $M = 100K$ variants by default, **e–g** show simulation results using $N = 100K$ samples and $M = 300K$ variants. Results from 15 repeats shown

$$w_i = \begin{cases} (-1)^i \left(1 - \dfrac{i}{2L^{\text{active}}}\right) & i \le L^{\text{active}}, \\ 0 & o/w. \end{cases}$$

The effect size parameters β and γ were simulated from a spike and slab prior such that the number of non-zero elements was governed by M_{G} and M_{GxE} for main and interaction effects respectively. Non-zero elements were drawn from a standard Gaussian, and then subsequently rescaled to ensure that the heritability given by main and interaction effects was h_g^2 and h_{GxE}^2 respectively. We chose a set of default parameters: $N = 25K; M = 100K; L = 30; L^{\text{active}} = 6$, $M_{\mathrm{G}} = 2500; M_{\mathrm{GxE}} = 1250; h_g^2 = 20\%$; $h_{\mathrm{GxE}}^2 = 5\%$, and then varied one parameter at a time to examine the effects of sample size, number of environments, number of non-zero SNP effects and GxE heritability. In addition, we investigated performance using a larger baseline simulation with $N = 100K$ samples and $M = 300K$ variants. The first genetic principal component was provided as a covariate to all methods.

Figure 1 compares estimates of the proportion of variance explained (PVE) by GxE effects from all three methods. In general, all methods had upwards bias that decreased with sample size and increased with the number of environments. While heritability estimates from LEMMA and GPLEMMA appeared quite similar, estimates from MEMMA had much higher variance and also appeared to have higher upwards bias as the total number of environments increase. All the methods exhibited less bias in the larger simulations with $N = 100K$ samples and $M = 300K$ variants *(*Fig. 1e–g). Supplementary Figure S1 shows the estimated PVE by main effects from the same set of simulations. In general the estimated PVE by main effects from the three methods was extremely similar.

Figure 2 compares the absolute correlation between the simulated ES and the ES inferred by LEMMA and GPLEMMA (note that MEMMA does not provide an estimate of the ES). In general, the estimated ES from GPLEMMA had slightly lower absolute correlation with the true ES than the estimated ES from LEMMA. LEMMA models SNP effects directly, so can estimate the SNPs involved in the interaction, and this is likely the reason for the small improvement in the accuracy of the ES. In large sample sizes ($N = 100k$ with $M = 300k$ SNPs), both methods achieve a correlation of over 0.98 with the simulated ES.

Next, we tested the methods against a generalized phenotype with three orthogonal ESs that interacted with disjoint sets of SNPs. More explicitly, the generalized phenotype was simulated from the following model

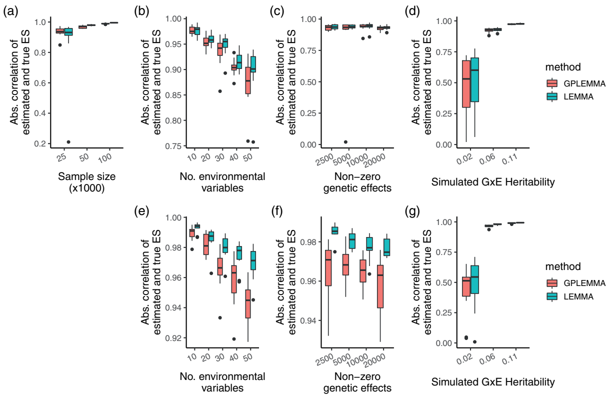

Fig. 2. Correlation between the true ES and the inferred ES. Absolute correlation between the true ES and the ES inferred by LEMMA and GPLEMMA whilst varying the number of environments, the number of active environments, the number of non-zero SNP effects and GxE heritability. All simulations constructed with the baseline phenotype. **a–d** Simulations results using $N = 25K$ samples and $M = 100K$ variants by default, **e–g** simulation results using $N = 100K$ samples and $M = 300K$ variants. Results from 15 repeats shown

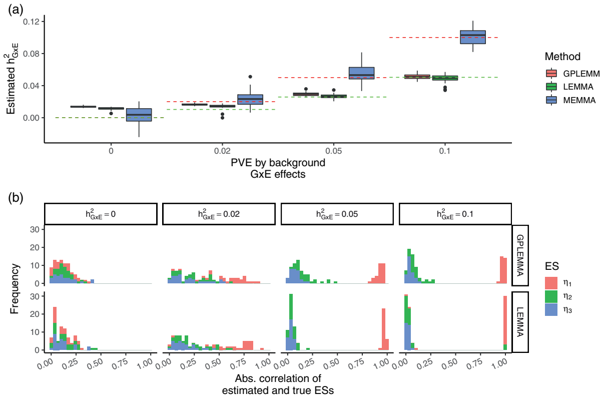

Fig. 3. Comparison with 3 simulated ESs. **a** Estimates of the proportion of variance explained by GxE effects by LEMMA, MEMMA and GPLEMMA whilst varying the simulated GxE heritability. The red dashed line indicates the total GxE heritability. The dashed green line indicates the heritability of the first GxE component η_1. **b** The (absolute) correlation between the estimated ES and the three simulated ESs. Simulations constructed using $N = 100K$ samples and $M = 100K$ variants. Results from 15 repeats shown

$$y \tilde{} \mathcal{N}\left(C\alpha + X\beta + \sum_{k=1}^{3} \eta_k \odot X\gamma_k, I_N\right),$$

where $\eta_k = EW_k$ is an N-vector and W is an $L \times K$ matrix of environmental weights with $K = 3$. W contained $L^{\text{active}} = 6$ rows with non-zero elements drawn from a standard Gaussian distribution, and columns were pairwise orthogonal. Similarly, γ_k denotes an M-vector where the non-zero elements of γ_k and γ_j are disjoint. SNP coefficients were drawn from a spike and slab prior in a similar manner to the baseline phenotype with $M_G = 2500$ and $M_{GxE} = 1250$.

Figure 3 displays results from a set of simulations with $N = 100k$ samples and $M = 100k$ SNPs. The three ESs were scaled such that the singular values of Λ where decreasing (with values 80, 60, 40). Figure 3b shows that the ES estimated by both LEMMA and GPLEMMA was, in general, highly correlated with the ES with highest singular value when the simulated GxE heritability was over 0.05. In this scenario, estimates of GxE heritability from both methods was centered around the true simulated heritability of that ES (see Fig. 3a). However at lower levels of simulated GxE heritability, both methods generally failed to identify one of the true ESs (consistent with Fig. 2). In contrast, the estimates of GxE heritability from MEMMA were centered on the sum of the GxE heritabilities from all three ESs. This suggests that running both GPLEMMA and MEMMA may help to elucidate the architecture of GxE interactions for a given trait.

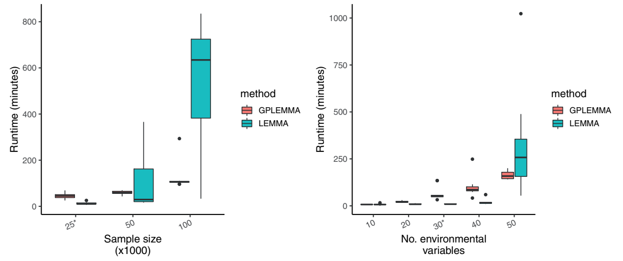

Fig. 4. Computational cost of GPLEMMA. Comparison of the computational cost of LEMMA and GPLEMMA as (**a**) sample size and (**b**) the number of environments increases. Runtime shown on a $\log_{10}$ scale. By default each run used four cores, $N = 25k$ samples, $L = 30$ environments and 10 random starts of the Levenburg-Marquardt algorithm. Results from 10 repeats shown. Runtime for LEMMA excludes time spent on single SNP hypothesis testing

For a more challenging scenario, we reran the high GxE heritability ($h^2_{\text{GxE}} = 0.1$) simulation using a larger number of SNPs ($M = 300K$) so that $N > M$ and scaled the ESs such that the singular values corresponding to the first two ESs were roughly equal (the singular values were 80, 78, 60). In this scenario the ES estimated by GPLEMMA did not correlate well with any one of the three simulated ESs, but was highly correlated with a linear combination of them (Supplementary Fig. S2). We hypothesize that with a large number of ‘null’ variants GPLEMMA becomes less able to identify a single ES and instead infers a mixture. In contrast, LEMMA consistently identified one of the simulated ESs. The GxE heritability estimates from all three methods were roughly as before, but with some inflation (Supplementary Fig. S3).

In the third batch of simulation results we used a misspecified phenotype, with squared dependancy on a heritable environmmental variable S. The misspecified phenotype was simulated using the model

$$y_{\text{misspec}} = \alpha_s S^2 + y_{\text{baseline}},$$

where S was simulated to have heritability of 30% with 2500 causal SNPs drawn from a spike and slab prior, and a range of values for α_s was used to vary the strength of the non-linear relationship between y and S.

Supplementary Figure S4 compares MEMMA, GPLEMMA and LEMMA in a simulation where the functional form of a heritable environmental variable was misspecified (or more specifically; the phenotype depended on the squared effect of a heritable environment). All methods were first tested without any attempt to control for model misspecification, and second using a preprocessing strategy where each environment was tested independently for squared effects on the phenotype and any squared effects with p-value $< 0.01/L$ were included as covariates. These are referred to as *(−SQE)* and *(+SQE)* respectively in the figures. Using the *(−SQE)* strategy, all methods showed upwards bias in estimates of GxE heritability that increased with the strength of the squared effect on the phenotype (Supplementary Fig. S4b). Model misspecification also caused bias in the ES of both GPLEMMA and LEMMA, however bias in the ES from GPLEMMA appeared to be much worse (Supplementary Fig. S4a). Using the (+*SQE*) strategy, all GxE heritability estimates were unbiased, consistent with earlier simulation results.

Figure 4 shows a comparison of the runtime of LEMMA and GPLEMMA as the sample size and number of environments varied. The figure shows that on relatively small datasets ($N = 25k$ samples or $L = 10$ environmental variables) LEMMA was slightly faster, however on large datasets GPLEMMA is faster due to superior algorithmic complexity. To give a direct comparison, on a simulated data with $N = 100k$ samples, $M = 100k$ SNPs and $L = 30$

environmental variables using 4 cores for each run, LEMMA took an average of 648 minutes to run whereas GPLEMMA took an average of 233 minutes.

Supplementary Figure S5 displays simulation results on the computational complexity of GPLEMMA. Supplementary Figure S5a and b shows that GPLEMMA achieved perfect strong scaling (A parallel algorithm has perfect strong scaling if the runtime on T processors is linear in $\frac{1}{T}$, including communication costs.) on the range of cores tested. This suggests that GPLEMMA has superior scalability to LEMMA, as for LEMMA the speedup due to increased cores began to decay after the number of samples per core dropped below 3000 (Kerin and Marchini, 2020).

Time to compute the preprocessing step and solve the non-linear least squared problem are shown in Supplementary Figure S5c–f, while the number of environments and sample size were varied. As expected, the preprocessing step appeared to be linear in both the number of environments and sample size. Time to solve the non-linear least squares problem appeared to be quadratic in the number of environments and approximately linear in sample size N. As a single LM iteration should have complexity $\mathcal{O}(NL^2B)$, this suggests that the number of iterations required for convergence of the LM algorithm was independent of sample size and the number of environments (at least for the range of values tested).

Finally, we tested GPLEMMA in simulation where we simulated ordinal environmental variables, created using a binomial distribution $Bin(n, p)$ where n with $n \in \{3, 4, 5\}$ levels and $p \sim U(0, 0.5)$. Comparing GPLEMMAs performance in these simulations with previous results using a continuous environmental variable (Supplementary Fig. S6) suggests that GPLEMMA is not sensitive to the choice of ordinal or continuous environmental variables.

2.5 Analysis of UK Biobank data

To compare GPLEMMA and LEMMA on real data we ran both methods on body mass index (log BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP) and pulse pressure (PP) measured on individuals from the UK Biobank. We filtered the SNP genotype data based on minor allele frequency (≥ 0.01) and IMPUTE info score (≥ 0.3), leaving approximately 642 000 variants per trait. We used 42 environmental variables from the UK Biobank, similar to those used in previous GxE analyses of BMI in the UK Biobank (Moore *et al.*, 2019; Young *et al.*, 2016). After filtering on ancestry and relatedness, sub-setting down to individuals who had complete data across the phenotype, covariates and environmental factors we were left with approximately $280,000$ samples per trait. The sample, SNP and covariate processing and filtering applied is the same as that reported in the LEMMA paper (Kerin and Marchini, 2020).

Table 1 shows the estimates and standard errors for SNP main effects (h^2_{G}) and GxE effects (h^2_{GxE}) for GPLEMMA and LEMMA applied to the 4 traits. In all cases there is good agreements between the estimates from both methods.

Finally we ran RHE-regression on the four UK Biobank traits whilst controlling for the same set of covariates. Heritability estimates from RHE-regression were marginally higher than those obtained by LEMMA and GPLEMMA (see Supplementary Table S1).

Table 1. Comparison of GPLEMMA and LEMMA on 4 UK Biobank traits

Trait	h^2_{G} (s.e)		h^2_{GxE} (s.e)	
	GPLEMMA	LEMMA	GPLEMMA	LEMMA
log BMI	0.256 (0.078)	0.259 (0.069)	0.074 (0.008)	0.071 (0.009)
PP	0.230 (0.042)	0.233 (0.039)	0.063 (0.007)	0.075 (0.018)
SBP	0.237 (0.057)	0.240 (0.053)	0.036 (0.003)	0.033 (0.003)
DBP	0.273 (0.037)	0.277 (0.034)	0.021 (0.003)	0.014 (0.001)

Note: Heritability estimates obtained using genotyped SNPs.

3 Discussion

Primarily this article develops a novel randomized Haseman–Elston non-linear regression approach for modelling GxE interactions of quantitative traits in large genetic studies with multiple environmental variables. This approach estimates GxE heritability at the same time as estimating the linear combination of environmental variables (called an ES) that underly that heritability. This general idea was pioneered in our previous approach LEMMA (Kerin and Marchini, 2020) which used a whole-genome regression approach to learn the ES, and this was then used in a randomized Haseman–Elston approach to estimate GxE heritability. The GPLEMMA approach introduced in this article does not need that first whole-genome regression step, and this leads to substantial computational savings. The model underlying GPLEMMA is very similar to that in LEMMA, but implicity assumes a Gaussian distribution for main SNP effects and GxE effects at each SNP.

We compared GPLEMMA to a simpler approach, which we called MEMMA, that estimates GxE heritability of each environmental variable in a joint model, but does not attempt to find the best linear combination of them. We found that estimates of GxE heritability from MEMMA had higher variance than estimates from LEMMA and GPLEMMA, suggesting that the usefulness of MEMMA might be limited. Results from LEMMA and GPLEMMA were very similar, both in terms of estimating the ES and GxE heritability. The primary advantage of GPLEMMA over LEMMA is in computational complexity, as the empirical complexity of GPLEMMA appeared to be linear in sample size whereas LEMMA was shown to be super-linear (Kerin and Marchini, 2020).

The methods LEMMA, GPLEMMA and MEMMA have all been developed for quantitative traits, and we have not explored their use when applied directly to binary traits, as if they were continuous, as was carried out in Pazokitoroudi *et al.* (2020). Developing GPLEMMA to directly model binary traits is a direction for future work. It maybe that transformations that exist for single component LMMs to convert heritability estimates to the liability scale may be also work here.

In the future it may also be interesting to extend the GPLEMMA model to partition variance using multiple orthogonal linear combinations of environmental variables. This could be expressed using the model

$$y = C\alpha + X\beta + \sum_{j=1}^{J}(\eta_j \odot X)\gamma_j + \epsilon, \tag{15}$$

where $\eta_j = Ew_j$ is an N-vector, w_j is an L-vector and $w_j \perp w_k \forall j, k \in \{1, \ldots, J\}$.

LEMMA is also able to perform single SNP hypothesis testing whereas GPLEMMA (currently) does not. The linear weighting parameter w from GPLEMMA could be used to initialize LEMMA, or the estimated ES could be used as a single environmental variable in LEMMA. Exploring these, and other, approaches is future work.

While we are enthusiastic about the potential of GPLEMMA (and LEMMA) to elucidate the contribution of GxE interactions to disease traits, we suggest that more care is needed than a standard genetic heritability analysis for a number of reasons. As our simulations show, including a variable as an 'environment' that is itself under genetic control can lead to bias if the relationship between that variable and trait of interest is not well modelled, and this should be carefully considered. Also, GPLEMMA is only able to assess the contribution of those variables included in the model. It may well be the case that a relevant environment is not available and so its GxE contribution cannot be assessed. Finally, the scale the measured trait can impact results (Kerin and Marchini, 2020) so it can be useful to assess results on a range of scales.

Acknowledgements

Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the

NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The authors are grateful to Sriram Sankararaman for discussions about the RHE approach during the 2019 CGSI at UCLA.

Funding

M.K. acknowledges funding for this work from the Wellcome Trust [109106/Z/15/Z]. J.M. acknowledges funding for this work from the European Research Council [617306].

Conflict of Interest: J.M. owns stocks and stock options in Regeneron Pharmaceuticals.

Data availability

The genetic and phenotype datasets generated by UK Biobank analyzed during the current study are available via the UK Biobank data access process. The resource is available to all bona fide researchers from academic, charity, public, and commercial sectors for all types of health-related research that is in the public interest: there is no preferential or exclusive access for any person. More details are available at http://www.ukbiobank.ac.uk/register-apply/.

References

Bulik-Sullivan,B.K. *et al.*; Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Bycroft,C. *et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.

Carbonetto,P. and Stephens,M. (2012) Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.*, **7**, 73–108.

Crawford,L. *et al.* (2017) Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet.*, **13**, e1006869.

de los Campos,G. *et al.* (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, **193**, 327–345.

Eskin,E. *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Evans,L.M. *et al.*; Haplotype Reference Consortium. (2018) Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.*, **50**, 737–745.

Finucane,H.K. *et al.*; ReproGen Consortium. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.

Golan,D. *et al.* (2014) Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA*, **111**, E5272–E5281.

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

Hayes,B. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Heckerman,D. *et al.* (2016) Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl. Acad. Sci. USA*, **113**, 7377–7382.

Hutchinson,M.F. (1990) A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Commun. Stat. Simulation Comput.*, **19**, 433–450.

Kerin,M. and Marchini,J. (2020) Inferring Gene-by-Environment Interactions with a Bayesian Whole-Genome Regression Model. *Am. J. Hum. Genet.*, **107**, 698–713.

Lippert,C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

Listgarten,J. *et al.* (2012) Improved linear mixed models for genome-wide association studies. *Nat. Methods*, **9**, 525–526.

Logsdon,B.A. *et al.* (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, **11**, 58.

Loh,P.R. *et al.* (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genetics*, **47**, 284–290.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Moore,R., BIOS Consortium. *et al.* (2019) A linear mixed model approach to study multivariate gene–environment interactions. *Nat. Genetics*, **51**, 180–186.

Ober,U. *et al.* (2015) Accounting for genetic architecture improves sequence based genomic prediction for a drosophila fitness trait. *PLoS One*, **10**, e0126880.

Pazokitoroudi,A. *et al.* (2020) Scalable multi-component linear mixed models with application to SNP heritability estimation. *Nat. Commun.*, **11**, 4020.

Powell,J.E. *et al.* (2018) Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genetics*, **50**, 746–753.

Speed,D. and Balding,D.J. (2019) SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.*, **51**, 277–284.

Speed,D. *et al.* (2012) Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.*, **91**, 1011–1021.

Speed,D. *et al.*; The UCLEB Consortium. (2017) Reevaluation of SNP heritability in complex human traits. *Nat. Genet.*, **49**, 986–992.

The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Wu,Y. and Sankararaman,S. (2018) A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics*, **34**, i187–i194.

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

Yang,J. *et al.* (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.

Yang,J. *et al.*; The LifeLines Cohort Study. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.*, **47**, 1114–1120.

Yang,J. *et al.* (2017) Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.*, **49**, 1304–1310.

Young,A.I. *et al.* (2016) Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nat. Commun.*, **7**, 12724.

Young,A.I. *et al.* (2018) Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat. Genet.*, **50**, 1304–1310.

Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.

Zhou,X. *et al.* (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.*, **9**, e1003264.

Zolfaghari,A. *et al.* (2005) An algorithm for the least-squares estimation of nonlinear parameters. *Int. J. Soil Sci.*, **3**, 270–277.