Non-linear randomized Haseman-Elston regression for estimation of gene-environment heritability

Matthew Kerin[1] & Jonathan Marchini[2]

[1] *Wellcome Trust Center for Human Genetics, Oxford, UK.*

[2] *Regeneron Genetics Center, Tarrytown, USA.*

Corresponding author *Jonathan Marchini (jonathan.marchini@regeneron.com)*

Abstract

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. We have developed a non-linear randomized Haseman-Elston (RHE) 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 computationally efficient and produces results highly correlated with those from LEMMA when applied to simulated data and real data from the UK Biobank.

Introduction

The advent of genome-wide association studies[1] 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[2]. 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[3]. 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[4–6]. These methods model the unobserved polygenic contribution as a multivariate Gaussian with covariance structure proportional to a genetic relationship matrix (GRM) [7–9]. This approach is mathematically equivalent to a whole genome regression (WGR) model with a Gaussian prior over SNP effects [4].

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[10,11] and generalisations have been proposed that stratify variance into different components by MAF and LD[10,12,13]. Other flexible approaches have been proposed in both the animal breeding[14,15] and human literature[16–18] 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[6,17]. Other methods have been proposed that estimate heritability only from summary statistics and LD reference panels [19,20]. Heritability can also be estimated using Haseman-Elston regression [21] and has recently been extended using a randomised approach [22] that has $\mathcal{O}(NM)$ computational complexity and works for multiple variance components [23]. 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. [24].

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[25] 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, modeled as the Hadamard product of the genomic and spatial covariance matrices. Other authors have used this method to account for gene-gene interactions [26,27].

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[28] focuses on detecting GxE interactions at individual markers, and models the environmental similarity between individuals (over multiple environments) as a random effect, and

then tests each SNP 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 whole genome regression approach called LEMMA applicable to large human datasets such as UK Biobank, where many potential environmental variables are available[29]. 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. 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[16–18]. 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 randomised Haseman-Elston (RHE) regression [23].

The main contribution of this paper 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 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/`.

Methods

Modeling SNP Heritability The simplest model for estimating SNP heritability has the form

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

where y is a continuous phenotype, X is an $N \times M$ matrix of genotypes that has been normalised 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

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

where $K = \frac{XX^T}{M}$ is known as the genomic relationship matrix (GRM)[3]. 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[10,11]. There have been many follow up methods, including; generalizations that stratify variance into different components by MAF and LD [13], approaches that assign different weights for the GRM[10,12], methods that replace the Gaussian prior on β with a spike and slab on SNP effect sizes [30] and methods that estimate heritability only from summary statistics and LD

reference panels [20,31].

Haseman-Elston (HE) regression An alternative method used to compute heritability is known as HE-regression[21]. HE-regression is a method of moments (MoM) estimator that optimizes variance components (σ_g^2, σ_e^2) in order to minimise 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

$$\underset{\sigma_g^2,\sigma_e^2}{\operatorname{argmin}} ||yy^T - (\sigma_g^2 K + \sigma_e^2 I)||_F^2$$

or equivalently by solving the linear regression problem

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

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

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

HE-regression methods are widely acknowledged to be more computationally efficient[22,32,33] and do not require any assumptions on the phenotype distribution beyond the covariance structure [32] (in contrast to maximum-likelihood estimators). However, HE-regression based estimates typically have higher variance [33], thus implying that they have less power.

Recent method developments[22,23] 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 et al. (2018) observed that Equation (1) can be solved efficiently without ever having to explicitly compute the kinship matrix K by using Hutchinson's estimator [34], which states that $\text{tr}\,(A) = \mathbb{E}\left[z^T A z\right]$ for any matrix where z is a random vector with mean zero and covariance given by the identity matrix. The proposed method involves approximating $\text{tr}\,(K)$ and $\text{tr}\,(K^2)$ using only matrix vector multiplications with the genotype matrix X, to compute the following expressions

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

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 extended this approach to a multiple component model [23]

$$y \sim \mathcal{N}(0, \sum_k \sigma_k^2 K_k + I\sigma_e^2)$$

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

$$\begin{pmatrix} T & b \\ b^T & N \end{pmatrix} \begin{pmatrix} \sigma_\beta^2 \\ \sigma_e^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{c} \\ N \end{pmatrix} \tag{2}$$

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

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

and terms in the subsequent linear system are given by $T_{kl} = \text{tr}\,(WK_kWK_lW)$, $b_k = \text{tr}\,(WK_kW)$ and $c_k = y^TWK_kWy$, where $W = I_N - C^T(C^TC)^{-1}C$.

Modeling 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, C is an $N \times D$ matrix of covariates, each with columns normalised to have mean zero and variance one.

MEMMA

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

$$y = C\alpha + X\beta + \sum_l (E_l \odot X)\lambda_l + \epsilon \tag{3}$$

where $\beta \sim \mathcal{N}(0, \frac{\sigma^2_\beta}{M} I_M)$, $\lambda_l \sim \mathcal{N}(0, \frac{\sigma^2_{w_l}}{M} I_M)$, $\epsilon \sim \mathcal{N}(0, \sigma^2_e I_N)$ 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

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

where $\theta = \{\sigma^2_\beta, (\sigma^2_{w_l})_{l=1}^L, \sigma^2_e\}$, $F_k = E_k \odot X$ and

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

Fitting the variance components is done analytically by solving the system of equations $T\theta = c$ where $T_{kl} = \text{tr}\,(WK_kWK_lW)$, $c_k = y^TWK_kWy$ and $W = I_N - C(C^TC)^{-1}C^T$. As shown in the original RHE method[22,23], 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^TWXX^TWy and the following N-vectors

$$u_b = XX^TWz_b, \tag{4}$$

$$v_{b,l} = XX^TE_lWz_b, \tag{5}$$

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

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

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

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

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

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

$$y = C\alpha + X\beta + (\eta \odot X)\gamma + \epsilon \tag{6}$$

where $\eta = Ew$ is the linear environmental score (ES), $\beta \sim \mathcal{N}(0, \frac{\sigma^2_\beta}{M} I_M)$ and $\gamma \sim \mathcal{N}(0, \frac{\sigma^2_\gamma}{M} I_M)$. This is the same model used by LEMMA[29] 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

$$y \sim \mathcal{N}(C\alpha, \sigma^2_\beta K + \sigma^2_\gamma K_2(w) + \sigma^2_e I),$$

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

$$\text{vec}\left(yy^T\right) = \sigma^2_\beta \text{vec}\,(K) + \sum_{l,m} \sigma^2_\gamma w_l w_m \text{vec}\left(F_l F_m^T\right) + \sigma^2_e \text{vec}\,(I) + \epsilon'. \tag{7}$$

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

Levenburg-Marquardt algorithm

We use the Levenburg-Marquardt (LM) algorithm[35], 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

$$Y = f(\theta) + \epsilon, \tag{8}$$

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 $\theta_{\text{new}} = \theta_0 + \delta$ by solving the normal equations

$$\left(J(\theta_0)^T J(\theta_0) + \mu I\right) \delta = J(\theta_0)^T \epsilon(\theta_0), \tag{9}$$

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 (9) approximates the quadratic step appropriate for a fully linear problem, whereas for large values of μ Equation (9) 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 $\theta = \{\sigma_\beta^2, w, \sigma_e^2\}$, $Y = \text{vec}\left(yy^T\right)$ and

$$f(\theta) = \sigma_\beta^2 \text{vec}\left(K\right) + \sum_{l,m} w_l w_m \text{vec}\left(F_l F_m^T\right) + \sigma_e^2 \text{vec}\left(I\right).$$

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

$$H_{l,m} = E_l^T \text{diag}\left(Wy\right) X X^T \text{diag}\left(Wy\right) E_m, \qquad 1 \leq l, m \leq L$$

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

Let $(J^T J)_{\theta_i,\theta_j}$ denote the entry of the $J^T J$ that corresponds to $\frac{f(\theta)}{\partial\theta_i}^T \frac{f(\theta)}{\partial\theta_i}$ for $\theta_i, \theta_j \in \{w, \sigma_\beta^2, \sigma_e\}$ 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

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

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

$$\begin{aligned}
\left(J(\theta)^T \epsilon(\theta)\right)_{\sigma_\beta^2} &= \text{tr}\left(y^T W K W y\right) - J(\theta)^T J(\theta) \sigma_\beta^2, \\
\left(J(\theta)^T \epsilon(\theta)\right)_{w_l} &= \text{tr}\left(y^T W \text{diag}\left(E_l\right) K \text{diag}\left(Ew\right) W y\right) - J(\theta)^T J(\theta) w_l, \\
\left(J(\theta)^T \epsilon(\theta)\right)_{\sigma_e^2} &= \text{tr}\left(y^T W y\right) - J(\theta)^T J(\theta) \sigma_e^2.
\end{aligned}$$

where

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

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

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

where

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

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 parallelise GPLEMMA using OpenMPI by partitioning samples across cores, in a similar manner to that used by LEMMA [29]. Given that evaluating the objective function $S(\sigma^2_\beta, w)$ is characterised 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 initialisations, and keep results from the solution with lowest squared error $S(\hat{\theta})$. Each run is initialised 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.

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

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

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

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

Relationship between MEMMA and GPLEMMA

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

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

where $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ 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 $\{g, (E_l \odot g)_{l=1}^L\}$ 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

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

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 that 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.

Results

Simulated data We carried out a simulation study to assess the relative properties of MEMMA and GPLEMMA. In addition, we compared to running the whole genome regression model in LEMMA, which estimates an ES and then uses it to estimate the GxE heritability.

The simulations use real data subsampled from genotyped SNPs in the UK Biobank[36], drawing

SNPs from all 22 chromosomes in proportion to chromosome length and using unrelated samples of mixed ancestry ($N = 25k$; 12500 white British, 7500 Irish and 5000 white European, $N = 50k$; 29567 white British, 7500 Irish and 12568white European, $N = 100k$; 79567 white British, 7500 Irish and 12568 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.

Phenotypes for the baseline simulations were all simulated according to the LEMMA model [29]. Let N be the number of individuals, M the total number of SNPs, M_g the number of causal main effect SNPs, M_{GxE} the number of SNPs with GxE effects, L the total number of environmental variables, L^{active} the number of 'active' environments with non-zero contribution to the ES vector w, and h^2_g and h^2_{GxE} the herirtability of main effects and GxE effects. The model used to simulate data is

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

$$\eta = Ew,$$

$$\epsilon \sim \mathcal{N}(0, I),$$

where X represents the $N \times M$ genotype matrix after columns have been standardised 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.

Non-zero elements of the interaction weight vector w were drawn from a decreasing sequence

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

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^2_g and h^2_{GxE} respectively. We chose a set of baseline parameter choices : $N = 25K; M = 100K; L = 30; L^{\text{active}} = 6, M_g = 2500; M_{\text{GxE}} = 1250; h^2_g = 20\%; h^2_{\text{GxE}} = 5\%$, 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 percentage 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 (Figure 1 (e-g)).

Figure 2 compares the absolute correlation between the simulated ES and the ES inferred by LEMMA and GPLEMMA. Models like MEMMA do 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, likely due to the data having been simulated from the LEMMA model, with sparse main and GxE SNP effects, whereas the GPLEMMA model assumes a polygenic or infinitesimal model. In large sample sizes $(N = 100k)$, both methods achieve a correlation of over 0.98 with the simulated ES.

Figure 3 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 (Figure 3b). 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 (Figure 3a). Using the $(+SQE)$ strategy, all GxE heritability estimates were unbiased, consistent with earlier simulation results.

Figure 4 displays simulation results on the computational complexity of GPLEMMA. Figures 4a

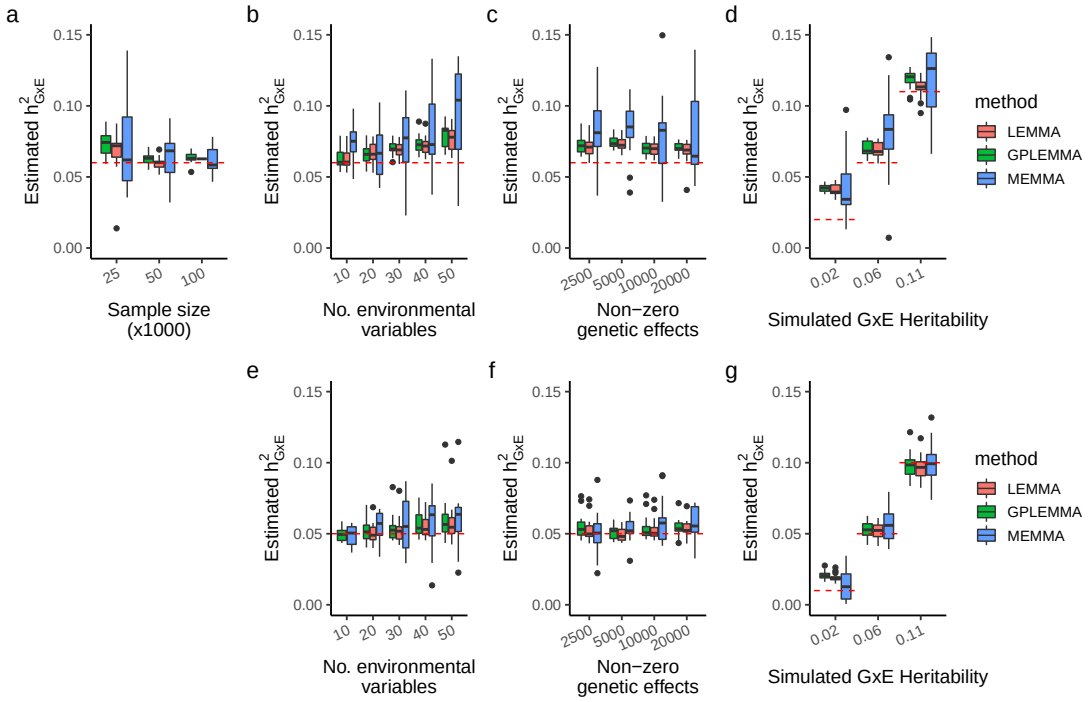

Figure 1: **PVE estimation**. Estimates of the proportion of variance explained by GxE effects by LEMMA, MEMMA and GPLEMMA on baseline simulations with using $N = 25K$ samples and $M = 100K$ variants, whilst varying sample size (a), the number of environments (b), the number of non-zero SNP effects (c) and GxE heritability (d). Panels (e-g) shows results of simulations with $N = 100K$ samples and $M = 300K$ variants, whilst varying the number of environments (e), the number of non-zero SNP effects (f) and GxE heritability (g).

and 4b show that GPLEMMA achieved perfect strong scaling[1] 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 [29].

[1] A parallel algorithm has perfect strong scaling if the runtime on T processors is linear in $\frac{1}{T}$, including communication costs.

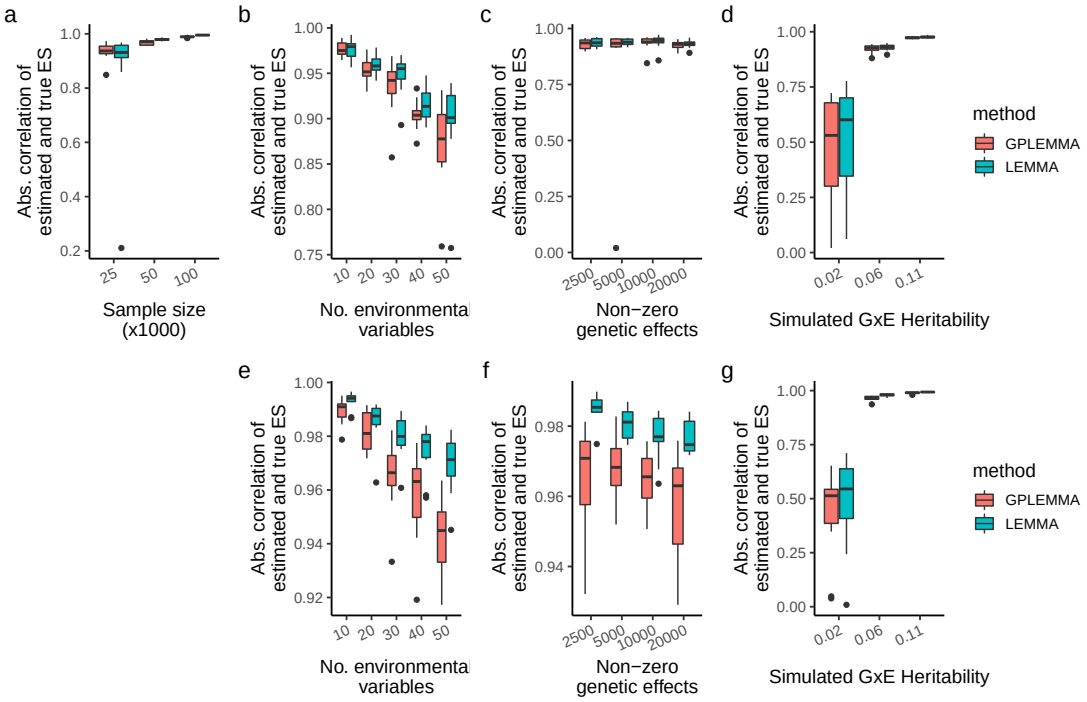

Figure 2: **Comparison on baseline simulations**. Absolute correlation between the true ES and the ES inferred by LEMMA and GPLEMMA whilst varying the number of environments, the number of active environments, the number of non-zero SNP effects and GxE heritability. The top row contains simulations using $N = 25K$ samples and $M = 100K$ variants, the bottom row containhs simulations using $N = 100K$ samples and $M = 300K$ variants. Results from 15 repeats shown.

Time to compute the preprocessing step and solve the non-linear least squared problem are shown in Figures 4c to 4f, 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

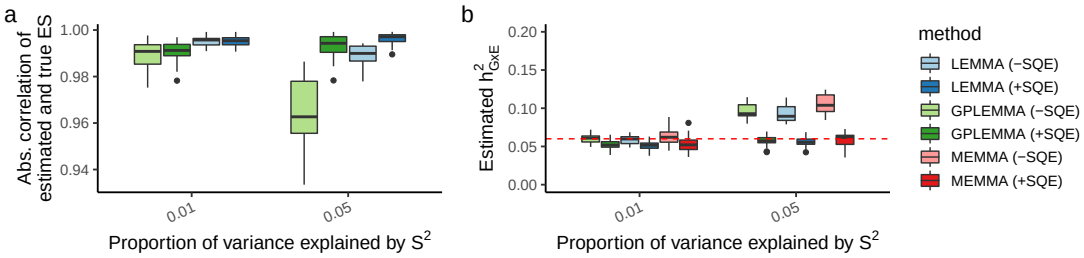

Figure 3: **Comparison on simulations with a misspecified heritable environment** Estimated proportion of trait variance explained by GxE effects is shown on the left, absolute correlation between the inferred ES and the true ES shown in the right. Results shown using LEMMA, MEMMA and GPLEMMA. Phenotypes simulated with a squared effect from a heritable confounder (see **??**). Results from 20 repeats shown. Abbreviations; (-SQE), no attempt to control for squared effects; (+SQE), squared effects with $p < 0.01$ (Bonferroni correction for multiple envs) included as covariates

LM algorithm was independent of sample size and the number of environments (at least for the range of values tested).

Finally, to give a direct comparison between LEMMA and GPLEMMA, we ran each method on simulated data with $N = 100k$ samples, $M = 100k$ SNPs and $L = 30$ environmental variables using 4 cores for each run. Over 20 repeats, LEMMA took an average of 648 minutes to run whereas GPLEMMA took an average of 233 minutes.

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

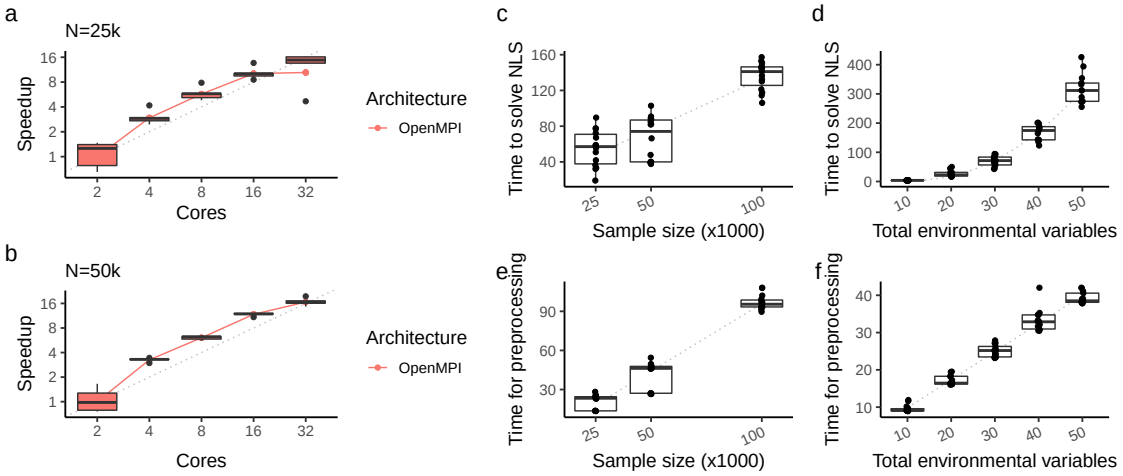

Figure 4: **Computational complexity of GPLEMMA in simulation**. Strong scaling of GPLEMMA using OpenMPI to parrallelise across cores with (a) $N = 25k$ samples and (b) $N = 50k$ samples. Comparison of the runtime of the Levenburg-Marquardt algorithm (c, d) and runtime of the preprocessing step (e, f). By default each run used; four cores, $N = 25k$ samples, $L = 30$ environments and 10 random starts of the Levenburg-Marquardt algorithm. Results from 15 repeats shown.

(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 [28,37]. 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[29].

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.

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)

Table 1: **Comparison of GPLEMMA and LEMMA on 4 UK Biobank traits**. Heritability estimates obtained using genotyped SNPs.

Discussion

Primarily this paper develops a novel randomized Haseman-Elston non-linear regression approach for modelling GxE interactions 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 [29] 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 paper 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 [29].

In the future it may also be interesting to explore the idea of further partitioning variance using multiple orthogonal linear combinations of environmental variables. This could be expressed using the model

$$y = C\alpha + X\beta + \sum_{j=1}^{J}(\eta_j \odot X)\gamma_j + \epsilon \tag{10}$$

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

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.

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. We are grateful to Sriram Sankararaman for discussions about the RHE approach during the 2019 CGSI at UCLA.

Author contributions

J.M. and M.K. conceived the ideas for the model and methods development. M.K. conducted all analyses and developed the software that implemented the methods with guidance from J.M. M.K. and J.M. wrote the manuscript.

References

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

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

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

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

5. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods (2014).

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

7. Eskin, E. *et al.* Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709–1723 (2008).

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

9. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012).

10. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* (2012).

11. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics* **50**, 737–745 (2018).

12. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nature Genetics* **49**, 986–992 (2017).

13. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47**, 1114–1120 (2015).

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

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

16. Logsdon, B. A., Hoffman, G. E. & Mezey, J. G. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11** (2010).

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

18. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics* **9**, e1003264 (2013).

19. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics* **51**, 277–284 (2019).

20. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295 (2015).

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

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

23. Pazokitoroudi, A. *et al.* Scalable multi-component linear mixed models with application to SNP heritability estimation. *bioRxiv* 522003 (2019).

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

25. Heckerman, D. *et al.* Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences* **113**, 7377–7382 (2016).

26. Ober, U. *et al.* Accounting for Genetic Architecture Improves Sequence Based Genomic Prediction for a Drosophila Fitness Trait. *PLOS ONE* **10**, e0126880 (2015).

27. Crawford, L., Zeng, P., Mukherjee, S. & Zhou, X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genetics* **13**, e1006869 (2017).

28. Moore, R. *et al.* A linear mixed model approach to study multivariate gene-environment interactions. *Nature Genetics* **51**, 180–186 (2018).

29. Kerin, M. & Marchini, J. Gene-environment interactions using a bayesian whole genome regression model. *bioRxiv* (2019). URL `https://doi.org/10.1101/797829`.

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

31. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

32. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences* **111**, E5272–E5281 (2014).

33. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics* **49**, 1304–1310 (2017).

34. Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation* **19**, 433–450 (1990).

35. Zolfaghari, A. *et al.* An algorithm for the least-squares estimation of nonlinear parameters. *International Journal of Soil Science* **3**, 270–277 (2005).

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

37. Young, A. I., Wauthier, F. & Donnelly, P. Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nature Communications* **7**, 12724 (2016).