GxEMM: Extending linear mixed models to general gene-environment interactions

Andy Dahl[1,*], Na Cai[2,3], Jonathan Flint[4], and Noah Zaitlen[1,*]

[1]Department of Medicine, University of California San Francisco, San Francisco, CA
[2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[4]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA
[*]andywdahl@gmail.com, noah.zaitlen@ucsf.edu

August 21, 2018

Abstract

Gene-environment interaction (GxE) is a well-known source of non-additive inheritance. GxE can be important in applications ranging from basic functional genomics to precision medical treatment. Further, GxE effects elude inherently-linear LMMs and may explain missing heritability. We propose a simple, unifying mixed model for polygenic interactions (GxEMM) to capture the aggregate effect of small GxE effects spread across the genome. GxEMM extends existing LMMs for GxE in two important ways. First, it extends to arbitrary environmental variables, not just categorical groups. Second, GxEMM can estimate and test for environment-specific heritability. In simulations where the assumptions of existing methods do not hold, we show that GxEMM improves estimates of ordinary and GxE heritability and increases power to test for polygenic GxE. We then use GxEMM to prove that the heritability of major depression (MD) is reduced by stress, which we previously conjectured but could not prove with prior methods, and that a tail of polygenic GxE effects remains unexplained by MD GWAS.

1 Introduction

Gene-environment interaction (GxE) has been thoroughly documented at the level of individual genetic variants. In functional genomics, some variants have effects on expression that depend on external context [1–6], age [7], tissue [8], cell type [9, 10], or other genetic variants [11, 12]. In complex traits, GxE has been found for exposures like air pollution [13], child abuse [14], microbe exposure [15] and major lifetime stress [16]. Genetics can also interact with medical interventions, sometimes rendering treatment ineffective [17–19]. G-by-sex, perhaps the most studied form of genetic interaction, has been found for diseases including asthma, autism and diabetes [20–23].

GxE effects are among many explanations given for 'missing heritability', i.e. the gap between the variance explained by GWAS-significant SNPs and twin-based heritability estimates (h^2_{twin}) [24]. Although linear mixed models (LMMs) have shown that much of this missing heritability can be explained by many variants of small effect [25, 26], ordinary LMM heritability (h^2_g) can still be substantially lower than h^2_{twin}. For example, major depression (MD) has h^2_{twin} of roughly 30-40%, but h^2_g is more like 20-30% [27]. Because the linearity assumption inherent in LMMs excludes heritability derived from genetic interactions, GxE may explain some of this gap.

Despite these two parallel streams of work–seeking individual GxE effects and explaining missing heritability–comparatively little has been done in the intersection, i.e. to study GxE effects in aggregate across the genome. However, the success of ordinary LMMs suggests that GxE may be better understood at the polygenic level, by modeling genome-wide random effects, rather than by studying a few large-effect variants. Further, whether known GxE loci represent exceptions or the rule is under active debate for many complex traits, including MD, making genome-wide GxE inference valuable for basic research, experimental design, and precision treatment.

Several methods to estimate GxE genome-wide have already been developed. A useful heuristic, lacking a generative model, was proposed for the special case where environments are categorical and have equal heritabilities [28]. More recently, this has been generalized in several directions: to allow genetic and residual correlations across two contexts measured on all samples [29]; or to use only summary statistics [30]; or, for a single SNP, to allow high-dimensional environments [31].

We present a new, general form for polygenic gene-environment interactions that we call a GxE mixed model (GxEMM). While complex in some settings, GxEMM has a simple visual interpretation in the case of discrete environments (Figure 1). GxEMM generalizes and unifies several existing GxE mixed models, which provides theoretical value. But GxEMM's primary value is pragmatic, as it extends the reach of GxE mixed models in two ways. First, it allows arbitrary environmental covariates, regardless whether they are continuous, discrete, proportions, or some combination. Second, GxEMM allows different environments to explain different amounts of GxE heritability; in particular, in the case of discrete environments, GxEMM can estimate and test for differential environment-specific heritability. Computationally, GxEMM uses existing software and thus has the same cost as generic REML methods.

We introduce ordinary genetic linear mixed models in Section 2.1. We then generalize them to model genetic interactions, without imposing the simplifying assumptions used in previous work, in Section 2.2. We discuss issues of restricting the parameter space, identifiability, scaling, testing, and defining environment-specific heritability in Sections 2.3-5. We then show GxEMM delivers roughly unbiased estimates and calibrated tests in simulations from the GxEMM model, and also that GxEMM can increase power to detect heterogeneity over existing tests. Finally, to demonstrate the practical utility of GxEMM, we re-analyze an MD dataset and validate our previous conjecture that major lifetime stress obscures the genetic impact on MD risk [16]. We conclude with comments on possible future methodological extensions and practical applications.

2 Polygenic GxE Mixed Model

2.1 Modelling polygenic additive effects

We assume a quantitative trait $y \in \mathbb{R}^N$ measured on N samples. We allow Q background covariates in $X \in \mathbb{R}^{N \times Q}$. Our focus is on the L SNPs measured in the genotype matrix $G \in \mathbb{R}^{N \times L}$. We assume columns of G are scaled to mean zero, variance one.

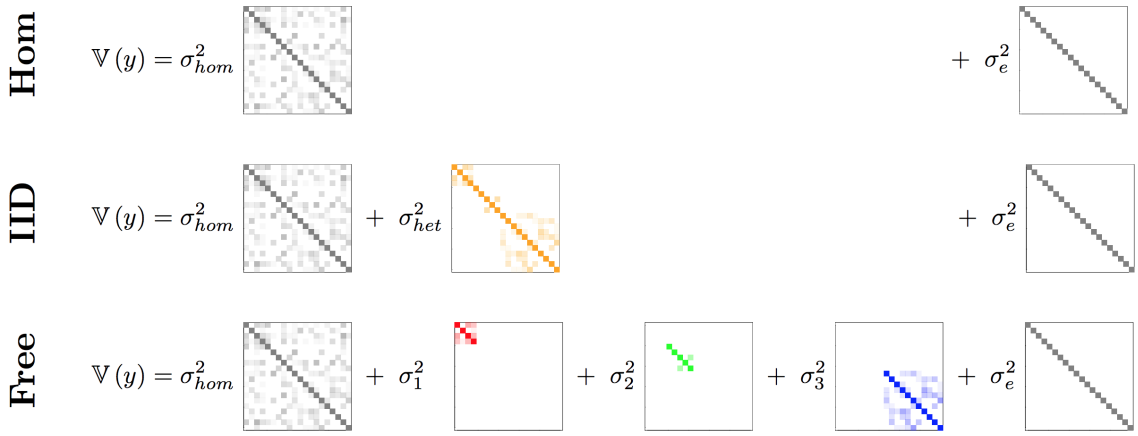

Figure 1: Depiction of the covariance matrices fit by the three LMMs we study. Samples are grouped by membership in one of three environments (red, blue and green). The ordinary GREML model (top) fits the additive heritability (governed by σ^2_{hom}, in grey) and noise (σ^2_ϵ). Darker colors indicate absolute values closer to 1, the value for all diagonal entries. The IID GxEMM (middle) adds a heterogeneous effect that modifies heritability only within environments (σ^2_{het}, orange). The Free GxE mixed model (bottom) allows different per-environment heritabilities (σ^2_k).

With this standard notation, the additive linear mixed model can be written:

$$y = X\alpha + G\beta + \epsilon \tag{1}$$

$$\beta \sim \mathcal{N}\left(0, \frac{\sigma^2_g}{L} I_L\right) \tag{2}$$

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2_\epsilon I_N\right) \tag{3}$$

We assume $Q \ll N$ and spend the full Q degrees of freedom to estimate α as a fixed effect. In contrast, we model β as a random effect. This can be motivated as a genuine prior that all SNPs have small, nonzero effects, or as a convenient model that yields an accurate maximum likelihood estimate for $\|\beta\|^2$ under more general assumptions [32].

The normality assumption on β has the substantial practical advantage that $\|\beta\|$ can be estimated without calculating the high-dimensional β. This is accomplished by marginalizing out β, which gives a simpler and (essentially) equivalent formulation of the above mixed model:

$$y \sim \mathcal{N}\left(X\alpha, \sigma^2_g K + \sigma^2_\epsilon I_N\right) \tag{4}$$

$$K := \frac{1}{L} G G^T \tag{5}$$

This defines K as the kinship/genetic relatedness matrix. Because G is assumed to have centered and scaled columns, K is a natural estimator of genetic similarity. While the model in (1) appears to require expensive computations in β, the marginalized model in (4) depends only on σ^2_g (and α and σ^2_ϵ). (4) can be fit by REML, which fits the variance components after projecting out X.

One main use for LMMs in genetics is to estimate the additive, or homogeneous, or narrow-sense, heritability, which we define as $h_g^2 := \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$ (assuming that G has centered and scaled columns). Conceptually, this quantifies the aggregate genetic contribution without identifying specific effects.

Another main use for LMMs in genetics is GWAS, accomplished by adding one SNP (at a time) as a column of X and testing if its corresponding entry in α is zero. The random genetic effect can be seen as adjusting for either polygenic background signal or non-genetic confounding [25]. A simple, common approximation to LMM-based GWAS is to only fit the variance components once, under the null hypothesis that SNPs have no effect [33], which works well in practice because SNP effects tend to be individually negligible (but see [34]).

2.2 Modelling polygenic interaction effects

We now assume P environmental variables have been measured in $Z \in \mathbb{R}^{N\times P}$. Like X, Z can contain binary and/or continuous values. We define the GxE mixed model (GxEMM) by adding environmental main effects and polygenic SNP-environment interactions to the ordinary LMM:

$$y_i = \left(\sum_q X'_{iq}\alpha_q\right) + \left(\sum_s G_{is}\beta_s\right) + \left(\sum_{s,p} G_{is}Z_{ip}\gamma_{(sp)}\right) + \epsilon_i \implies$$

$$y = X'\alpha + G\beta + (G * Z)\,\gamma + \epsilon \tag{6}$$

$$\beta \sim \mathcal{N}\left(0, \frac{\sigma_{hom}^2}{L} I_L\right); \quad \epsilon \sim \mathcal{N}\left(0, \sigma_\epsilon^2 I_N\right) \tag{7}$$

$$\gamma \sim \mathcal{N}\left(0, \frac{1}{L} I_L \otimes V\right) \tag{8}$$

where $\otimes$ is the Kronecker product, $*$ is the column-wise Khatri-Rao product, and X' are the fixed effects. We assume $P \ll N$ and estimate Z as a fixed effect, i.e. we define $X' := (X : Z)$. We retain the spherical Gaussian priors on β and ϵ from the additive LMM, though we rewrite σ_g^2 as σ_{hom}^2 to emphasize that the interaction model permits heterogeneous sources of inheritance. We write $\gamma_{(sp)}$ to indicate the entry of γ corresponding to the s-th SNP and p-th environment.

To our knowledge, the GxEMM model for the GxE coefficients, γ, is new. As for β, we assume that interaction effects are independent between SNPs, which is encoded in the assumed across-SNP covariance I_L. However, we allow covariance between environments, in the same way for all SNPs, as encoded in the $P \times P$ across-environment covariance V. Unlike the β prior, which depends on one parameter, the γ prior depends on a matrix of parameters, V. Intuitively, σ_{hom}^2 describes the size of β, while V_{pp} describes the size of the genetic interaction effects with environment p. The off-diagonal terms, $V_{pp'}$, account for genetic effects that are shared between environments p and p'.

Just as the random β can be marginalized for additive LMMs, the random β and γ in GxEMM can be marginalized. This gives the random-effect formulation:

$$y \sim \mathcal{N}\left(X'\alpha, \sigma_{hom}^2 K + K \circ \left(ZVZ^T\right) + \sigma_\epsilon^2 I_N\right) \tag{9}$$

The GxE variance component decomposes into the Hadamard product ($\circ$) of genetic- and environmental-similarity matrices. This is actually if and only if: the only covariance matrices for γ that give (9) are of the form $I_L \otimes V$ (assuming either large sample size or random G, Appendix).

The MLE for the model in (9), which involves a complicated covariance function of V, is not obvious. But, because the Hadamard product is linear, the matrix multiplications in ZVZ^T can be

expanded and commuted through the Hadamard product with K, giving the equivalent expression:

$$y \sim \mathcal{N}\left(X'\alpha, \sigma^2_{hom}K + \sum_{p,q} v_{pq}\left[K \circ \left(Z_{,p}Z^T_{,q}\right)\right] + \sigma^2_\epsilon I_N\right) \tag{10}$$

This covariance expression has a simple visual description for discrete environments (Figure 1).

(10) takes the form of a standard variance component estimation problem, which is useful because it can be solved by standard REML approaches. Specifically, we can use the GCTA software package [28] to fit Equation (10) with REML by using X as fixed effects and the similarity matrices K, $\{K \circ (Z_i Z_i^T)\}_{i=1}^K$, $\{K \circ (Z_i Z_j^T + Z_j Z_i^T)\}_{i \neq j}$, and I_N.

Various restrictions reduce GxEMM to existing models. First, StructLMM is obtained if $L = 1$ SNP is used and $V = \sigma^2_{het} I_P$, i.e. if interaction effects are i.i.d. across environments [31]. If, instead, Z contains only zeros and ones and $V = \sigma^2_{het} I_P$, GxEMM becomes equivalent to the original GxE model in GCTA [28]. A related model was developed specifically for sex as the environment, which fits the sexes separately to allow sex-dependent genetic architecture, noise, and covariate effects [23]. GxEMM is most similar to iSet [29], except iSet focuses on low-rank K, assumes two discrete environments, and assumes samples are observed in all environments.

2.3 Restricting V

We never fit GxEMM with a general V in this paper because it has $O(P^2)$ parameters, which is computationally intensive for generic REML algorithms. We instead consider two nested restrictions to the model. First, Free GxEMM restricts V to be diagonal, i.e. $V_{pp} = \sigma^2_p$ and $V_{pq} = 0$ for $p \neq q$. While this imposes independence between environment-specific effects–after accounting for homogeneous effects–it allows the GxE heritability explained to vary freely between factors in Z.

Second, we evaluate an even simpler IID GxEMM that restricts $V = \sigma^2_{het} I_P$, imposing that all columns of Z have similar genome-wide interaction effects. In particular, if Z contains unscaled discrete environment proportions, the IID model enforces constant heritability across environments. If, further, Z contains only 0s and 1s, IID GxEMM reduces to the GxE model in [28].

Enforcing diagonal V also makes σ^2_g and V jointly identified. Otherwise, variance explained can be passed between σ^2_g and V without changing the likelihood. To see this, let $1_P \in \mathbb{R}^P$ be a vector of 1s, let $\equiv$ denote equal GxEMM likelihood, and assume, say, that Z has rows summing to 1. Then

$$(\sigma^2_g, V) \equiv (\sigma^2_g - \lambda, V + \lambda 1_P 1_P^T) \quad \forall \lambda \in \mathbb{R}$$

The constraint that V is diagonal, however, makes the right hand side infeasible for $\lambda \neq 0$ (assuming V is feasible). Constraining $\sigma^2_g \geq 0$ or $V \succeq 0$, instead, would restrict λ to a proper interval of $\mathbb{R}$.

When there are P discrete environments, $\binom{P}{2} - 1$ degrees of freedom are identified for V, while only $P - 1$ are identified for fixed effect interactions (per SNP). The extra degrees of freedom are identified by the prior on γ; e.g. when V is diagonal, the model prefers (β, γ) pairs where the average across-environment effects (per SNP) are explained by β rather than γ.

2.4 Centering and Scaling G and Z

G can be assumed mean zero WLOG because the full data likelihood is invariant under the mapping $G \mapsto G + 1_N \mu^T$ for all $\mu \in \mathbb{R}^L$. Z, however, cannot be centered without changing the likelihood. This asymmetry between Z and G is because we treat α as fixed and β as random. The perturbation

induced by centering G lives in the span of Z, and because Z is a fixed effect this perturbation is annihilated when fitting variance components by REML.

We assume columns of G have been normalized to empirical variance 1, corresponding to a particular assumption about the decay of effect sizes with MAF [35]. Almost nothing changes, however, if another (deterministic) column scaling is used, as long as $\frac{1}{\sqrt{L}}\|G\|_F^2 = \mathrm{tr}\,(K) = N$.

We do not assume Z is scaled to have columns with mean zero and variance one. Unlike with G, which has roughly exchangeable columns, columns of Z may have meaningful relative scales. We do, however, require that $\|Z\|_F^2 = N$, which comes without loss of generality and permits simpler formulas for heritability. For example, if environments are discrete and the data are restricted only to samples from environment k, the heritability in the new dataset would be

$$h_k^2 := \frac{\sigma_{hom}^2 + \sigma_k^2}{\sigma_{hom}^2 + \sigma_k^2 + \sigma_e^2} \tag{11}$$

The full heritability in the observed data, accounting for both homogeneous and environment-specific heritability, replaces σ_k^2 in the above equation with an average of the σ_k^2 that is weighted by the size of each environment, $\frac{1}{N}\|Z_{,k}\|^2$.

The scale of columns of Z is irrelevant for Free GxEMM because the feasible set for V is closed under conjugation by diagonal matrices–multiplying $Z_{,k}$ by λ can be counterbalanced by multiplying row k and column k of V by λ^{-1}. However, the column scaling of Z affects results when V is forced to be spherical, as in IID GxEMM and existing GxE mixed models with $P > 2$.

2.5 Testing for significant GxE heritability

We use the standard LRT for GxE heritability terms and allow negative estimates for σ_{het}^2 (and non-positive definite estimates for V) for three reasons. First, this reduces bias–provably so in the case of ordinary heritability [32]–which is important for aggregating estimates across traits. Second, if GxE heritability is defined as $\tilde{h}^2 := \mathbb{E}\left(\|\text{genetic terms}\|^2\right)$, then the constraint $\tilde{h}^2 \geq 0$ no longer requires either that $\sigma_{hom}^2 \geq 0$ or that $V \succeq 0$. Third, it is conceivable that environmental similarity could attenuate the importance of genetic similarity.

3 Simulations

Because genome-wide genetic data is complex, we validated GxEMM with simulations based on real genotypes from CONVERGE (see Section 4). For each simulation, we randomly choose $S = 1,000$ SNPs to have causal additive and interaction effects and draw traits from the GxEMM model. We then fit three mixed models: the ordinary LMM with only a homogeneous genetic effect; the IID GxEMM with one parameter for GxE heritability; and the Free GxEMM allowing each heritability to vary freely between environments. Inside all mixed models, we use the standard genome-wide relatedness matrix as the causal SNPs are unknown in practice.

Specifically, we generate phenotypes from the following polygenic interaction model:

$$y = X\alpha + G\beta + (G * Z)\,\gamma + \epsilon$$

$$\beta_s \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma_{hom}^2}{S}\right); \quad \gamma_{sk} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \frac{\sigma_k^2}{S}\right); \quad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}\left(0, 1 - \sigma_g^2 - \sum_k \left(\frac{\sigma_k^2}{N}\|Z_{,k}\|^2\right)\right)$$

We define X to include the environments Z and the fixed effects we used in CONVERGE (Section 4) and draw $\alpha_q \overset{\text{iid}}{\sim} \mathcal{N}(0, .1)$; note that X is essentially irrelevant because it is projected out by REML. G contains the randomly chosen S causal SNPs. We scale columns of X and G to mean zero, variance one so that the effect sizes are easily interpretable. We choose $\mathbb{V}(\epsilon)$ so that the σ^2 terms can be interpreted as fractions of (residual) phenotypic variance explained.

We define the environment z_i for each person to be i.i.d. draws from one three discrete categories:

$$z_i \overset{\text{iid}}{\sim} \text{Categorical}(.1, .3, .6) \rightarrow Z_{ik} = \begin{cases} 1 & \text{if} \quad z_i = k \\ 0 & \text{if} \quad z_i \neq k \end{cases} \quad \forall i, k$$

$\rightarrow$ means that we define Z by encoding the K-level factor z as a binary matrix with K columns.

We varied the variance components in turn. First, we simulated under the homogeneous model by setting all $\sigma^2_k = 0$ and varying $\sigma^2_{hom} \in [0, .6]$ (Figure 2a,d). The IID and Free GxEMMs performed well in this null setting, each giving genetic heterogeneity estimates centered around zero and roughly calibrated tests (Figure 2a,d).

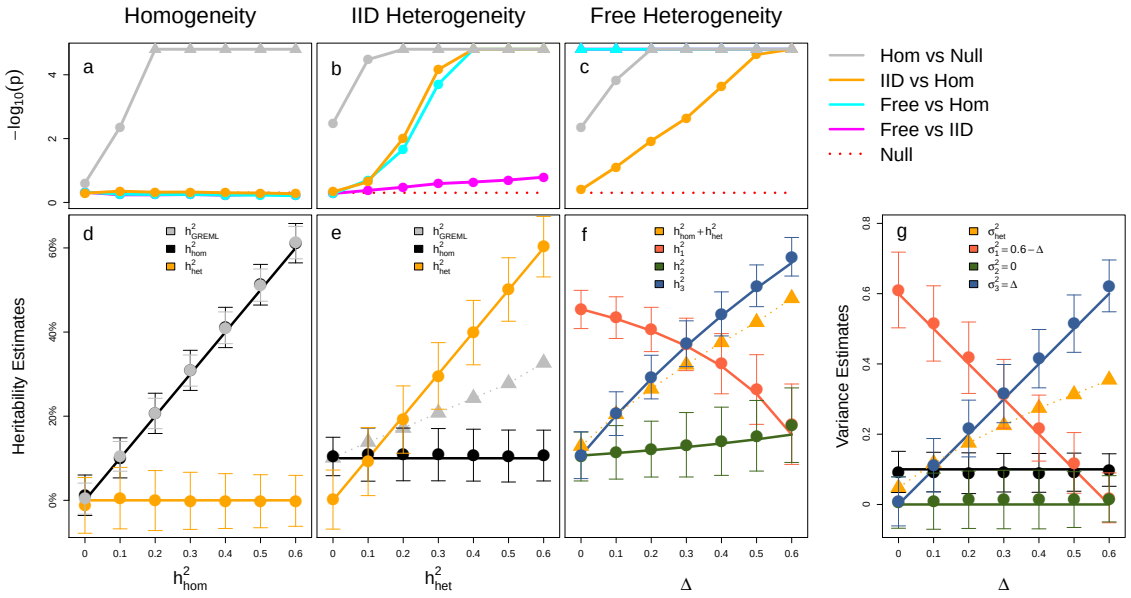

Figure 2: GxEMM tests (a-c) and genetic variance estimates (d-g) in polygenic GxE simulations with real genotypes. For each parameter setting, median $-\log_{10}(p)$ across 100 independent simulations are shown for four LRTs under (a) genetic homogeneity, (b) constant genetic heterogeneity across environments, or (c) variable genetic heterogeneity. Large y-axis values are truncated as triangles (in c, pink and blue overlap). Null gives the median of 10,000 independent, null $-\log_{10}(p)$ values. GxEMM estimates for the variance components are shown as medians ($\pm$ one standard deviation). The ordinary heritabilities that would be obtained by restricting to one environment are shown in (g) by applying equation (11) to estimates from (f). For comparison, estimates from too-simple models are shown in triangles in (b,c).

To add IID genetic heterogeneity, we fix $\sigma^2_{hom} = .2$ and vary $\sigma^2_{het} \in [0, .6]$ while requiring that $\sigma^2_{het} = \sigma^2_1 = \sigma^2_2 = \sigma^2_3$ (Figure 2b,e). GxEMM accurately estimated the heterogeneous heritability,

and the test for differences between the σ_k^2 was roughly calibrated, especially for realistic values of genetic heterogeneity ('Free vs IID'). Note that even though the Free model is needlessly general in this simulation, its test for genetic homogeneity only loses minimal power relative to IID GxEMM.

Finally, we fix $\sigma_{hom}^2 = .1$ and add differences in the σ_k^2 by varying a tradeoff parameter $\Delta \in [0, .6]$ (Figure 2c,f,g). We fix $\sigma_2^2 = 0$ to reflect a moderately-sized subgroup without any specific genetic basis, and we interpolate the heterogeneous heritability between the small and large group by $\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2\right) := (.6 - \Delta, 0, \Delta)$. When Δ is small, most of heritability resides in the smallest environment, and the Free GxEMM test for genetic heterogeneity is substantially more powerful than existing, IID-based tests. The IID heterogeneity test gains power as Δ increases, meaning that we shift heritability from the small group to the large group, though Free GxEMM remains more powerful and explains more heritability. We also display the raw GxEMM estimates for σ_k^2 before rescaling to environment-specific heritabilities with equation (11) (Figure 2g).

4 Polygenic heterogeneity in major depression

Major depression is a moderately heritable trait ($h_{twin}^2 \approx 40\%$ [36]) that is likely highly heterogeneous. The CONVERGE consortium [37] was established to enrich genetic signal for MD by recruiting only women with recurrent episodes of MD, between the ages of 30-60, and of Han Chinese ethnicity. Women were chosen because they have higher MD heritability (42%) than men (29%) [36]. Recurrent, clinically ascertained MD cases were selected because they likely harbor more heritable disease etiologies [38]. CONVERGE collected data from 5,303 cases and 5,337 matched controls. CONVERGE identified the first replicated GWAS results for MD, supporting the hypothesis that heterogeneity can mask causal genetic MD signals [37].

Further, in a direct test of this hypothesis, CONVERGE found three SNPs that are active only in MD cases without major stressful life events (SLE) [16]. We aim to use GxEMM to go beyond SNP-level tests by capturing differential genome-wide genetic architectures between SLE groups. This is motivated by analyses in [16] that found a suggestive and plausible but statistically insignificant difference in heritability between SLE groups.

We studied 9,303 samples with genotype and SLE data. We used an intercept, age, and SLE status as fixed effects. To control for population structure, we also included the top ten genetic PCs (from an LDAK relatedness matrix [39]) and their interactions with SLE [40].

We fit the homogeneous model and found $\hat{h}_g^2 = 31.8\%$ (s.e. 4.4%). This estimate (and s.e.) is on the observed scale due to complexities of transforming to the liability scale in the presence of environmental main effects and differential prevalences. We next fit IID GxEMM and found $\hat{h}_{IID}^2 = 37.7\%$ (s.e. 5.6%), which combines the homogeneous (h_{hom}^2) and IID heterogeneous (h_{het}^2) heritability components; we use the delta method to approximate the standard error. Finally, we fit Free GxEMM and found $h_{stress}^2 = 31.2\%$ (s.e. 6.5%) and $h_{non-stress}^2 = 39.3\%$ (s.e. 5.6%). The Free model fit better than the IID model (LRT p=0.00018), showing that the stressed samples harbor significantly lower genetic signal than the non-stressed.

[16] found three SNPs that interacted with SLE (rs7526682, rs11577545, and rs950893). To estimate the aggregate size of GxE effects that remain unaccounted for, we refit Free GxEMM while including these SNPs as fixed effects. While the stressed heritability estimate is basically unchanged (31.2%), the non-stressed heritability slightly decreases (38.0%). This trend suggests these three SNPs primarily act in non-stressed samples and that many GxE effects remain unexplained.

A significant caveat is that our current GxEMM implementation models binary trait heritability

on the observed scale, treating the trait as quantitative rather than translating to the more natural liability scale. This translation is important even in the homogeneous case, as standard LMM gives biased heritability estimates for binary traits [41]. Further, even after translation, REML estimates are still biased, especially with preferential case recruitment and large covariate effects, and moment-based approaches are required instead [42, 43]. However, these methods are not easily applied under genetic heterogeneity, and REML is a common method in practice. We will extend GxEMM to more carefully handle binary traits and ascertainment in future work.

5 Discussion

Gene-environment interactions and small, genome-wide effects are separately well-documented. We unified approaches to study these two types of genetic effects with a Khatri-Rao linear mixed model for polygenic GxE called GxEMM. Marginalizing the GxE effects gives a simple, visually intuitive variance component model that can be fit with existing software. We showed the importance of the generality offered by GxEMM in simulations where the assumptions of existing methods fail. We showed the practical significance of GxEMM by analyzing a human major depression dataset and showing, for the first time, that major lifetime stress obscures the genetic impact on MD.

GxEMM can be applied to any genome-wide datasets with relevant environmental measurements. It is particularly useful and interpretable when partitioning heritability amongst discrete environmental factors. Though we have used the word environment, our approach can be used for anything that putatively interacts with the genome, including other genotypes [44] or study indicators in a random effect meta-analysis [45, 46]. The significant practical utility of GxEMM is that environmental variables need not be categorical and that different amounts of heritability can be attributed to different environments.

GxEMM could also be used for GWAS. Models similar to IID GxEMM have been shown to improve power and calibration [40], suggesting the Free GxEMM can further improve results when the IID assumption fails. In fact, GxEMM GWAS would cost no more than standard LMM-based GWAS when approximating variance components by their null estimates, which is common practice.

The main limitation of our GxEMM implementation relative to other GxE mixed models is computation time. In special cases, the problem can be rewritten to roughly reduce the complexity to $O(N)$ [29, 31], which can be a major advantage over our $O(N^3)$ implementation. Fundamentally, we do not exploit the structure of the GxE kinship matrices, which almost certainly enables faster computation. Methodologically, it would be interesting to model environment-specific variances, like iSet [29]. Noise heterogeneity can bias our current GxEMM implementation, though we could in principle add $K - 1$ relatedness matrices for environment-specific noise.

References

[1] Luis B Barreiro et al. “Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection.” *Proceedings of the National Academy of Sciences of the United States of America* 109.4 (Jan. 2012), pp. 1204–1209.

[2] Julien Gagneur et al. “Genotype-Environment Interactions Reveal Causal Pathways That Mediate Genetic Effects on Phenotype”. *PLoS Genetics* 9.9 (2013), e1003803.

[3] M N Lee et al. “Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells”. *Science* 343.6175 (Mar. 2014), pp. 1246980–1246980.

[4] B P Fairfax et al. "Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression". *Nature* 343.6175 (2014), pp. 1246949–1246949.

[5] Alfonso Buil et al. "Quantifying the degree of sharing of genetic and non-genetic causes of gene expression variability across four tissues." *BioRxiv* (May 2016), p. 053355.

[6] David A Knowles et al. "Allele-specific expression reveals interactions between genetic variation and environment". *Nature Methods* 14.7 (July 2017), pp. 699–702.

[7] Daniel Glass et al. "Gene expression changes with age in skin, adipose tissue, blood and brain". *Genome Biology* 14.7 (July 2013), R75.

[8] GTEx Consortium. "Genetic effects on gene expression across human tissues". *Nature* 550.7675 (Oct. 2017), pp. 204–213.

[9] Daria V Zhernakova et al. "Identification of context-dependent expression quantitative trait loci in whole blood". *Nature Genetics* 49.1 (Jan. 2017), pp. 139–145.

[10] Hyun Min Kang et al. "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation". *Nature Biotechnology* 36.1 (Dec. 2017), pp. 89–94.

[11] Andrew Anand Brown et al. "Genetic interactions affecting human gene expression identified by variance association mapping." *eLife* 3 (Apr. 2014), e01381.

[12] Simon K G Forsberg et al. "Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast". *Nature Genetics* 49.4 (Apr. 2017), pp. 497–503.

[13] Marie-Julie Favé et al. "Gene-by-environment interactions in urban populations modulate risk phenotypes." *Nature communications* 9.1 (Mar. 2018), p. 827.

[14] Avshalom Caspi et al. "Role of Genotype in the Cycle of Violence in Maltreated Children". *Science* 297.5582 (Aug. 2002), pp. 851–854.

[15] Luke Jostins et al. "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease." *Nature* 491.7422 (Nov. 2012), pp. 119–124.

[16] Roseann E Peterson et al. "Molecular Genetic Analysis Subdivided by Adversity Exposure Suggests Etiologic Heterogeneity in Major Depression." *The American journal of psychiatry* 175.6 (June 2018), pp. 545–554.

[17] D V Exner et al. "Lesser response to angiotensin-converting-enzyme inhibitor therapy in black as compared with white patients with left ventricular dysfunction." *New England Journal of Medicine* 344.18 (May 2001), pp. 1351–1357.

[18] Jessica L Mega et al. "Reduced-Function CYP2C19 Genotype and Risk of Adverse Clinical Outcomes Among Patients Treated With Clopidogrel Predominantly for PCI: A Meta-analysis". *JAMA* 304.16 (Oct. 2010), pp. 1821–1830.

[19] Nadeem Riaz et al. "Recurrent SERPINB3 and SERPINB4 mutations in patients who respond to anti-CTLA4 immunotherapy." *Nature Genetics* 48.11 (Nov. 2016), pp. 1327–1329.

[20] Rachel A Myers et al. "Genome-wide interaction studies reveal sex-specific asthma risk alleles." *Human Molecular Genetics* 23.19 (Oct. 2014), pp. 5251–5259.

[21] Ileena Mitra et al. "Pleiotropic Mechanisms Indicated for Sex Differences in Autism". *PLoS Genetics* 12.11 (Nov. 2016), e1006425.

[22] Kerrin S Small et al. "Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition". *Nature Genetics* 50.4 (Apr. 2018), pp. 572–580.

[23] Eun Yong Kang et al. "An Association Mapping Framework To Account for Potential Sex Difference in Genetic Architectures". *Genetics* 209.3 (May 2018), genetics.300501.2017–698.

[24] Teri A Manolio et al. "Finding the missing heritability of complex diseases". *Nature* 461.7265 (Aug. 2009), pp. 747–753.

[25] Hyun Min Kang et al. "Efficient control of population structure in model organism association mapping." *Genetics* 178.3 (Mar. 2008), pp. 1709–1723.

[26] Jian Yang et al. "Common SNPs explain a large proportion of the heritability for human height." *Nature Genetics* 42.7 (July 2010), pp. 565–569.

[27] Jonathan Flint and Kenneth S Kendler. "The genetics of major depression." *Neuron* 81.3 (Feb. 2014), pp. 484–503.

[28] Jian Yang et al. "GCTA: a tool for genome-wide complex trait analysis". *The American Journal of Human Genetics* (2011).

[29] Francesco Paolo Casale et al. "Joint genetic analysis using variant sets reveals polygenic gene-context interactions". *PLoS Genetics* 13.4 (Apr. 2017), e1006693.

[30] Vincent Laville et al. "VarExp: Estimating variance explained by Genome-Wide GxE summary statistics". *BioRxiv* (Nov. 2017), p. 224634.

[31] Rachel Moore et al. "A linear mixed model approach to study multivariate gene-environment interactions". *BioRxiv* (Feb. 2018), p. 270611.

[32] David Steinsaltz, Andrew Dahl, and Kenneth W Wachter. "Statistical properties of simple random-effects models for genetic heritability". *Electronic Journal of Statistics* 12.1 (2018), pp. 321–358.

[33] Hyun Min Kang et al. "Variance component model to account for sample structure in genome-wide association studies." *Nature Genetics* 42.4 (Apr. 2010), pp. 348–354.

[34] Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies." *Nature Genetics* 44.7 (July 2012), pp. 821–824.

[35] Matthew Stephens and David J Balding. "Bayesian statistical methods for genetic association studies." *Nature Reviews Genetics* 10.10 (Oct. 2009), pp. 681–690.

[36] Kenneth S Kendler et al. "A Swedish National Twin Study of Lifetime Major Depression". *American Journal of Psychiatry* 163.1 (Jan. 2006), pp. 109–114.

[37] Converge Consortium. "Sparse whole genome sequencing identifies two loci for major depressive disorder". *Nature* 523.7562 (July 2015), pp. 588–591.

[38] Peter McGuffin et al. "A Hospital-Based Twin Register of the Heritability of DSM-IV Unipolar Depression". *Archives of General Psychiatry* 53.2 (Feb. 1996), pp. 129–136.

[39] Doug Speed et al. "Improved Heritability Estimation from Genome-wide SNPs". *The American Journal of Human Genetics* 91.6 (Dec. 2012), pp. 1011–1021.

[40] Jae Hoon Sul et al. "Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models". *PLoS Genetics* 12.3 (Mar. 2016), e1005849.

[41] Sang Hong Lee et al. "Estimating missing heritability for disease from genome-wide association studies." *American journal of human genetics* 88.3 (Mar. 2011), pp. 294–305.

[42] David Golan, Eric S Lander, and Saharon Rosset. "Measuring missing heritability: inferring the contribution of common variants." *Proceedings of the National Academy of Sciences of the United States of America* 111.49 (Dec. 2014), E5272–81.

[43] Omer Weissbrod, Jonathan Flint, and Saharon Rosset. "Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics". *The American Journal of Human Genetics* 103.1 (July 2018), pp. 89–99.

[44] Lorin Crawford, Sayan Mukherjee, and Xiang Zhou. "Detecting Epistasis in Genome-wide Association Studies with the Marginal EPIstasis Test". *BioRxiv* (July 2016), p. 066985.

[45] Buhm Han and Eleazar Eskin. "Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies". *The American Journal of Human Genetics* 88.5 (May 2011), pp. 586–598.

[46] Xiaoquan Wen and Matthew Stephens. "Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions". *The Annals of Applied Statistics* 8.1 (Mar. 2014), pp. 176–203.

Appendix: Marginalizing random interaction coefficients

The point of the proposition is that the set of Gaussian GxE coefficients with covariance $D \otimes V$ describes essentially the same distributions as the set Hadamard kinship matrices $\left(GDG^T\right) \circ \left(ZVZ^T\right)$. The former are the natural model for polygenic GxE; the latter are easily fit by REML and have already been studied in special cases.

Proposition 1. *Assume that* $\gamma \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^{P \times L}$*, that* $G \in \mathbb{R}^{N \times L}$ *has continuous, random entries, and that* $Z \in \mathbb{R}^{N \times P}$ *has rank* P*, with* $P < N$*. Write* $K := GDG^T$ *for some fixed* D*. Then*

$$\mathbb{V}\left((G * Z)\,\gamma\right) = K \circ \Omega \iff (\Sigma, \Omega) = \left(D \otimes W, ZWZ^T\right) \text{ for some } W$$

where $*$ *is the column-wise Khatri-Rao product;* $\circ$ *is Hadamard; and* $\otimes$ *is Kronecker.*

Proof. First, note the identity:

$$\left(G_{i,} \otimes Z_{i,}\right)(D \otimes W)\left(G_{j,} \otimes Z_{j,}\right)^T = \left(G_{i,} D G_{j,}^T\right)\left(Z_{i,} W Z_{j,}^T\right) = K_{ij}\Omega_{ij}$$

Now, right to left is easy:

$$\mathbb{V}\left((G * Z)\,\gamma\right)_{ij} = \left(G_{i,} \otimes Z_{i,}\right)\mathbb{V}(\gamma)\left(G_{j,} \otimes Z_{j,}\right)^T = K_{ij}\Omega_{ij}$$

The other direction assumes decomposition of the variance into Hadamard products holds, so

$$\left(G_{i,} \otimes Z_{i,}\right)(D \otimes W)\left(G_{j,} \otimes Z_{j,}\right)^T = K_{ij}\Omega_{ij} = \left(G_{i,} \otimes Z_{i,}\right)\Sigma\left(G_{j,} \otimes Z_{j,}\right)^T$$

Using the standard identity $(A \otimes B)\,\mathrm{vec}(C) = \mathrm{vec}\left(B^T C A\right)$, where $\mathrm{vec}(\cdot)$ concatenates columns of a matrix, the above equality can be written as

$$\left(G_{j,} \otimes Z_{j,} \otimes G_{i,} \otimes Z_{i,}\right)\mathrm{vec}\left(\Sigma - D \otimes W\right) = 0$$

By assumption, this identity holds $\forall i, j$ and almost all genotypes $G_{i,}, G_{j,} \in R^L$. The span of such pure four-way tensors is $\mathbb{R}^{L^2P^2}$, so its kernel has dimension zero and thus $\Sigma - D \otimes W = 0$. □

A similar result can be obtained if N grows large with fixed L.