# Gene-environment interactions using a Bayesian whole genome regression model

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

The contribution of gene-environment (GxE) interactions for many human traits and diseases is poorly characterised. We propose a method, LEMMA, that estimates an interpretable environmental score (ES) that interacts with genetic markers throughout the genome. When applied to body mass index, systolic, diastolic and pulse pressure in the UK Biobank we estimate that $9.3\%$, $3.9\%$, $1.6\%$ and $12.5\%$ of phenotypic variance is explained by GxE interactions, and that rare variants explain most of this variance. We also identify 3 loci that interact with the estimated environmental scores ($-\log_{10} p > 7$).

## Introduction

Despite longstanding interest in gene-by-environment (GxE) interactions[1], this facet of genetic architecture remains poorly characterised in humans. Detection of GxE interactions is inherently

more difficult than detection of additive genetics in genome wide association studies (GWAS). One difficulty is that of sample size; a commonly cited rule of thumb suggests that detection of interaction effects requires a sample size at least four times larger than that required to detect a main effect of comparable effect size [2]. Another is that an individual's environment, which occurs through time, is very hard to measure in a comprehensive way, and is inherently high dimensional. Also, there are many environmental variables that could plausibly interact with the genome and many ways to combine them, and typically these factors were not all present in the same dataset. The recently released UK Biobank[3] dataset, a large population cohort study with deep genotyping and sequencing and extensive phenotyping, offers a unique opportunity to explore GxE effects [4–10].

Models that consider environmental variables jointly can be advantageous, particularly 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. StructLMM[7] models the environmental similarity between individuals (over multiple environments) as a random effect, and then tests each SNP independently for GxE interactions, but does not model the genome wide contribution of all the markers, which is often a major component of phenotypic variance.

Advances in methods applied to detect genetic main effects in standard GWAS have shown that linear mixed models (LMMs) can reduce false positive associations due to population structure, and improve power by implicitly conditioning on other loci across the genome [11–13]. Often these methods model the unobserved polygenic contribution as a multivariate Gaussian with covariance structure proportional to a genetic relationship matrix (GRM) [14–16]. This approach is

mathematically equivalent to a whole genome regression (WGR) model with a Gaussian prior over SNP effects [11]. More flexible approaches have been proposed in both the animal breeding [17,18] and human literature [19–21] to allow different prior distributions that better capture SNPs of small and large effects. The BOLT-LMM method[13] uses a mixture of Gaussians (MoG) prior and shows this can increase power to detect associated loci in some (but not all) complex traits, using a method with $\mathcal{O}(N^{1.5}M)$ computational complexity.

Here we present a new method called Linear Environment Mixed Model Analysis (LEMMA) which aims to combine the advantages of WGR and modelling GxE with multiple environments. Instead of assuming that the GxE effect over multiple environments is independent at each variant, as StructLMM does, we learn an environmental score (ES) which is a single linear combination of environmental variables, that has a common role in interaction effects genome wide. One advantage of this approach is that the ES provides a readily interpretable way to examine the combined effect of many environmental variables. The ES is estimated within a Bayesian WGR model, that uses an MoG prior on both main effects and GxE effects (see **Online Methods**). We use variational inference to fit the model that is tractable for GxE analyses of biobank scale datasets with tens of environmental variables.

A LEMMA analysis has several distinct steps. First, the model is fitted using a large set of SNPs genome-wide, for example all the SNPs that have been directly assayed on a genotyping chip. The estimated ES is then used to estimate the proportion of phenotypic variability that is explained by GxE interactions (SNP GxE heritability), using new randomised Haseman-Elston

(RHE) regression on UK Biobank scale datasets [22,23]. This heritability analysis can be run on genotyped or imputed SNPs and stratified by MAF and LD to better interrogate the genetic architecture of GxE interactions. The ES is also used to test for GxE interactions one variant at a time, typically at a larger set of imputed SNPs in the dataset. We use "robust" standard errors when testing each variant for a GxE interaction, which helps to control for the conditional heteroskedasticity caused by GxE interactions. We also suggest checks and solutions for the situation where environmental variables are themselves heritable and have a non-linear relationship to the trait of interest (see **Online Methods**).

We compared LEMMA to existing approaches such as StructLMM and F tests using simulated data, and applied the approach to UK Biobank data for body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP) and pulse pressure (PP).

## Results

**Performance on simulated data** **Figure 1** shows the performance of different methods to detect GxE in simulations where a single true ES interacts with SNPs across the genome. **Figure S1** shows the false positive rate (FPR) to detect main effects. We compared our default version of LEMMA, which uses robust standard errors, StructLMM, a simple F test of interaction and an F test that uses robust standard errors (see **Online Methods**). The simulations vary GxE heritability, the total number of environmental variables and sample size. When sample size is large (N=100k), all the methods have reasonable control of FPR and LEMMA controls FPR at least as well as

other methods across the range of simulations. When sample size is smaller (N=25k) the robust F-test performs less well as the number of environments grows (**Figure 1a**) and the F-test and StructLMM perform less well as the amount of GxE variance increases (**Figure 1b**).

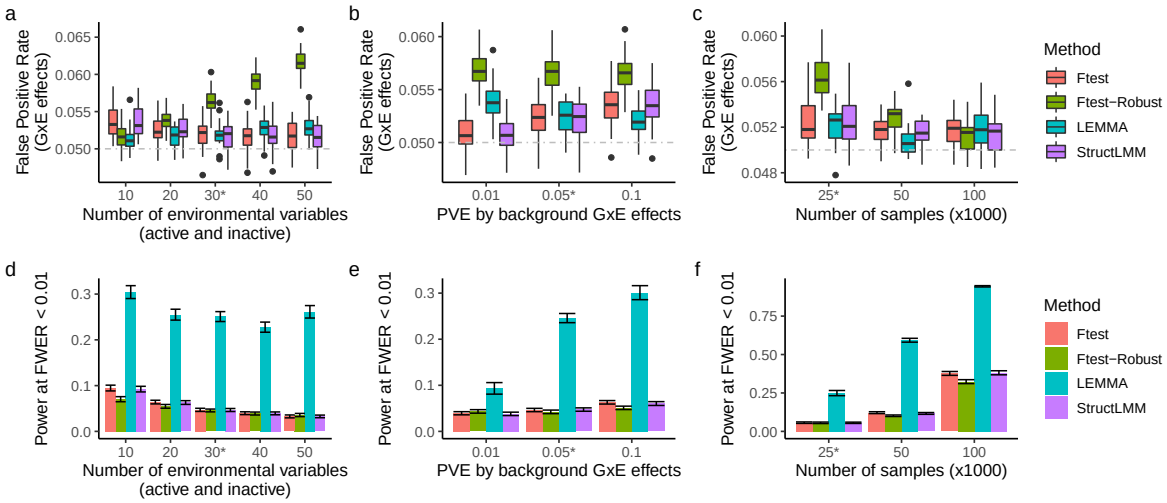



Figure 1: **Type I error and Power of tests to detect GxE effects in simulation.** (a-c) Comparison of false positive rate as the number of environments increases (a), as phenotype variance explained by GxE effects increases (b) or as the number samples increases (c). (e-f) Analagous comparison of the power to detect GxE interations. Simulations used genotypes subsampled from the UK Biobank and by default contained $N = 25K$ samples, $M = 100K$ SNPs, 6 environmental variables that contributed to the ES and 24 that did not (default parameters denoted by stars). We assess power (at Family Wise Error Rate; FWER $< 0.01$) to detect 60 causal SNPs whose GxE effect each explained $0.00016\%$ of trait variance. See **Online methods** for full details of phenotype construction.

When there is a single true ES involved in GxE interactions we found that LEMMA provided a substantial power increase (**Figure 1**, **Figure S2**). StructLMM and F tests have similar power in these simulations, although previous work suggests that StructLMM may outperform the F test in

small samples [7].

When estimating the GxE heritability of the LEMMA ES using randomised HE regression with a single SNP component (RHE-SC) we observed some upward bias as the number of environments increases, but this effect is ameliorated by increasing sample size (see **Figure S3**) suggesting that the influence of overfitting in our Biobank analyses is mild. In twenty simulations with $L = 30$ environmental variables, $N = 100$k samples and true GxE heritability of $5\%$ we observed mean GxE heritability of $5.2\%$. **Figure S4** further illustrates the ES estimation accuracy of LEMMA.

Finally, we ran LEMMA on two sets of simulated datasets (N=25k) with causal SNPs chosen either randomly, or chosen to be rare (MAF<0.1). We used the ES estimated from each simulated dataset to estimate $h^2_G$ and $h^2_{GxE}$ using RHE, with SNPs stratified by MAF and LD (RHE-LDMS), and then without any stratification (RHE-SC). **Figure S5** shows that stratifying by MAF and LD results in accurate heritability estimates irrespective of the MAF distribution of causal SNPs, and suggests that this method can be used to interrogate the MAF distribution of GxE component of a trait using LEMMA. However when causal SNPs are rare, not stratifying by MAF and LD results in underestimation of $h^2_G$.

**Controlling for heritable environmental variables** Previous work has shown that misspecifying the functional form of an environmental variable can induce heteroskedasticity into tests for GxE interactions and hence that the robust F-test will still be well calibrated, if the environment is independent of genotypes[24]. Independence between genotypes and the misspecified environment is important because it means that the least squares estimator is still unbiased.

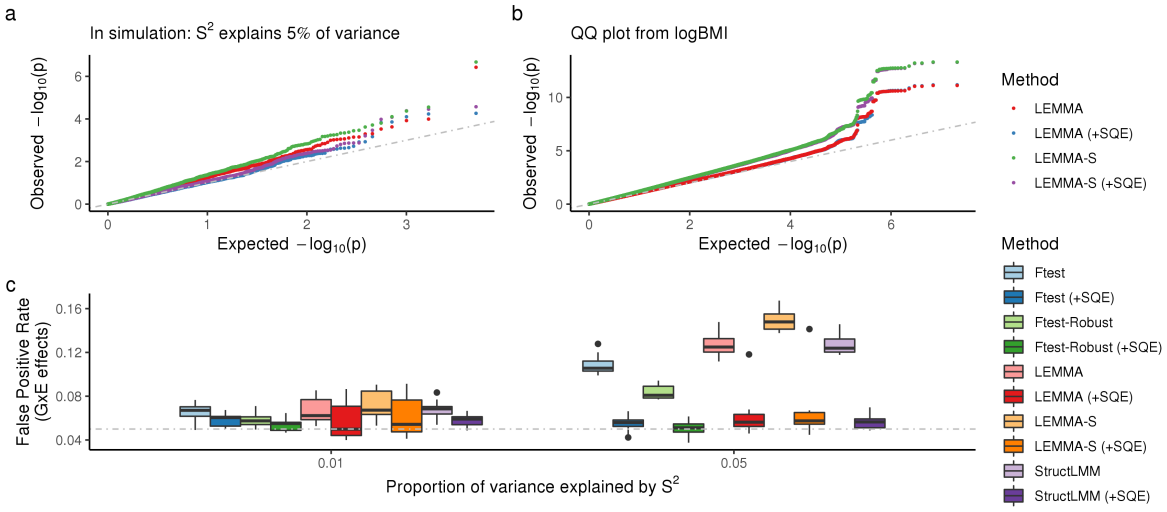



Figure 2: **Bias from model misspecification of a heritable environmental variable.** (a) Comparison of GxE association test statistics from a single simulation where non-linear dependence on the confounder explains $5\%$ of trait variance. FPR at heritable sites of the misspecified environment only. (b) Comparison of GxE association test statistics from an analysis of $\log$BMI in $281,149$ participants from the UK Biobank. (c) False Positive Rate at heritable sites of the misspecified environment whilst the strength of squared dependence varies. $20$ repeats per scenario. Abbreviations are as follows: LEMMA-S, LEMMA with non-robust variances used to compute test statistics; (+SQE), significant squared environmental variables (Bonferroni correction) included as additional covariates.

However, environmental variables themselves often have a genetic basis. We therefore performed simulations where the phenotype depended on the non-linear (squared) effects of a heritable environmental variable. In simulation (**Figure 2a,c**) we observed that misspecification of the environmental variable can cause substantial inflation in GxE test statistics at heritable sites of the confounding environment. Relatively smooth non-linearities, such as squared effects, are easily

detected by regression modelling before using LEMMA (see **Online Methods**) and can then be included as covariates (indicated in **Figure 2** by +SQE). This procedure produced well calibrated test statistics for all methods in simulation (**Figure 2c**).

In **Figure 2b** we compare the GxE association test statistics from our analysis of logBMI in the UK Biobank, with and without adjusting for detected squared effects. Although we detected squared effects for 30 of the 42 environmental variables (significance level 0.01; Bonferroni correction for multiple testing), the ES obtained from the two analyses was almost identical (Pearson $r^2 > 0.999$). As the additional variance explained collectively by the squared effects was negligible (incremental $R^2 < 0.00001$) it would be surprising if this was not the case. Negative $\log_{10}(P)$-values from the two analyses were also highly correlated (Pearson $r^2 = 0.961$), although there were small changes in the $p$-values at the *FOXO3* locus (which remained genome-wide significant in both analyses) and at the *SNAP25* locus (which was genome wide significant in the (-SQE) analysis only). We therefore conclude that the influence from this form of confounding in our analysis of logBMI was minor. However as the cost to this procedure is small, LEMMA uses the (+SQE) strategy by default for all analyses of UK Biobank traits.

**GxE interaction analysis in the UK Biobank** We applied LEMMA to characterise GxE interactions in logBMI, SBP, DBP and PP using a set of 42 environmental variables similar to those used in previous analyses [7,8,25], including data on smoking, hours of TV watched, Townsend Index, physical exercise and alcohol consumption (see **Online methods** and **Table S1**).

Previous studies have established that estimating heritability using a single SNP component

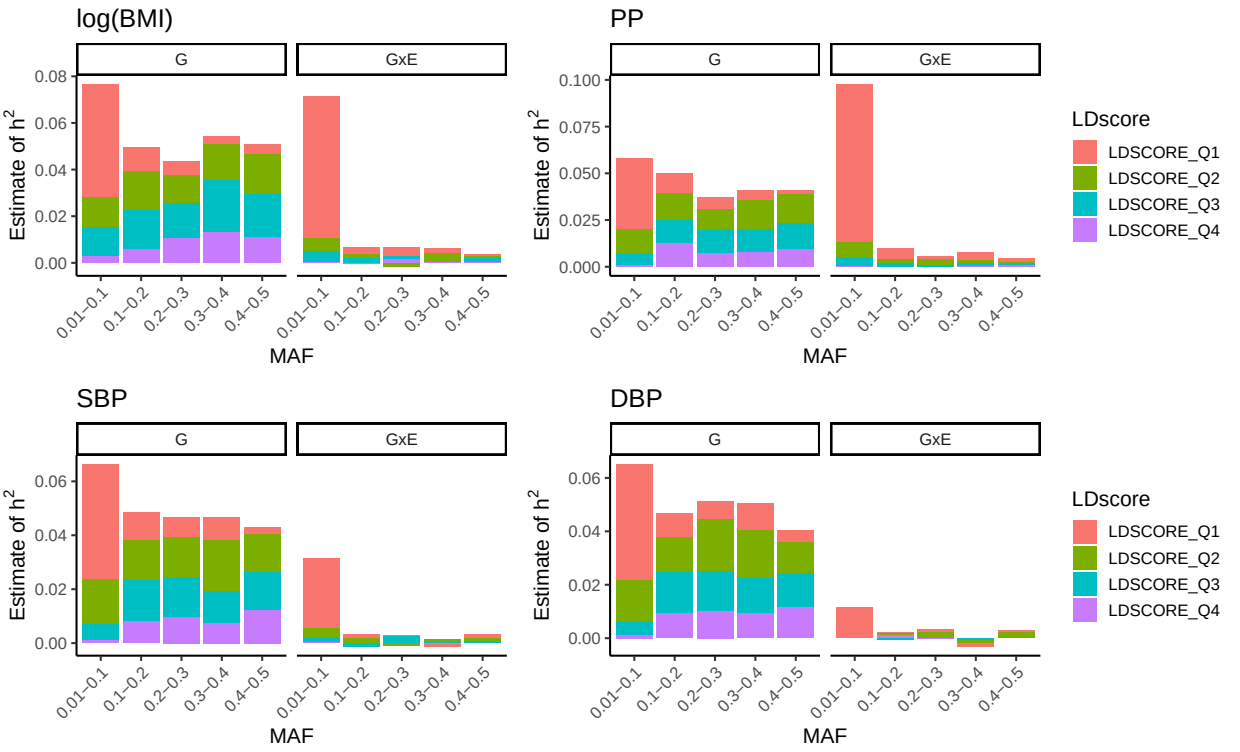



Figure 3: **Partitioned heritability estimates for four quantitative traitsin the UK Biobank**. Heritability estimates partitoned into additive genetic and multiplicative GxE interaction effects for four quantitative traits in the UK Biobank using approximately $280,000$ unrelated white British individuals (see **Table S1**) and $M = 10,270,052$ common imputed SNPs (MAF $> 0.01$ in the full UK Biobank cohort). Multiplicative GxE iteractions were computed using the $ES$ from each model fit. Heritability estimation was performed using a multi-component implementation of randomised HE-regression[22,23] with SNPs stratified into $20$ components ($5$ MAF bins and $4$ LD score quantiles).

makes assumptions about the relationship between MAF, LD and trait architecture that may not hold up in practice[26,27]. In comparison, stratifying SNPs into bins according to MAF and LDscore (the LDMS approach) is has been shown to be relatively unbiased[27–29]. We analysed GxE heri-

tability due to multiplicative effects with the ES using both $M = 639,005$ genotyped SNPs and $M = 10,270,052$ common imputed SNPs (MAF $< 0.01$ in the full UK Biobank cohort), stratified by MAF and LDscore into $20$ components. Using imputed SNPs we estimated GxE heritability of $9.3\%$, $12.5\%$, $3.9\%$ and $1.6\%$ for $\log$BMI, PP, SBP and DBP respectively (see **Table 1**). On genotyped SNPs the GxE heritability estimates were slightly lower for $\log$BMI and PP ($h^2_{\text{GxE}} = 8.6\%$ and $h^2_{\text{GxE}} = 11.1\%$ respectively) and almost identical for SBP and DBP (see **Table S2**). For all traits the heritability of additive SNP effects were slightly higher on imputed data, consistent with previous results [28].

Previous work on models of natural selection has suggested that the variance explained by additive SNP effects should be uniformly distributed as a function of MAF in a neutral evolutionary setting[30], and that enrichment of the variance explained by rare SNPs is evidence for negative selection. For all four traits we found that variance explained by the additive genetic effects of rare SNPs (MAF $< 0.1$) was slightly elevated, consistent with previous observations of negative selection[31] (**Figure 3**). Additionally the distribution of additive genetic effects by MAF for $\log$BMI was qualitatively similar to that found by GREML-LDMS is a previous study[28]. In contrast we found that variance explained by GxE effects was overwhelming attributed to rare SNPs (MAF $< 0.01$), especially those with low LD. However we are not aware of any evolutionary theory that has been extended to model the MAF distribution of GxE effects.

For $\log$BMI we estimated an ES that put high weight on alcohol intake frequency, Townsend Index and physical activity measures (**Figure 4c**). Almost all of the non-dietary environmental

exposures had a higher effect in women than in men, with smoking status being the one exception. This is reflected in the facts that (a) the ES having much higher variance in women and (b) those with a negative ES were almost all female ($97\%$) (see **Figure 4b**). When comparing the characteristics of those in the bottom $5\%$ of the ES to the whole cohort (using the mean for continuous variables and the mode for categoric), we found that those in the bottom $5\%$ were predominantly female ($100\%$ vs $53\%$), younger (51 vs 56), had higher Townsend deprivation index (0.91 vs -1.74), drank less often ('Special occasions' vs 'Once or twice a week') and watched more TV (3.28 vs 2.69 daily hours of TV) (**Table S5**). We note that positive values of the Townsend index indicate material deprivation, whereas negative values indicate relative affluence.

Previous cross-sectional studies have reported GxE interactions between a linear predictor formed from BMI-associated SNPs and alcohol intake frequency [32], Townsend Index [32], physical activity measures [5,32,33] and time watching TV [32,33], all of which are upweighted in the LEMMA lifestyle score. An alternative approach from Robinson *et al.*[6] binned samples according to their environmental exposure (eg. age) and tested for significant differences in SNP heritability using a likelihood ratio test. They reported strong interaction effects with age in a cohort of $43,407$ individuals whose ages spanned $18-80$, but only reported significant interactions with smoking in the UK Biobank interim release. This suggests that we might expect age to play a more dominant role in the $\log$BMI ES in a cohort that included younger individuals. Finally, one category that is notably downweighted is the contribution from dietary variables. Although significant interactions with fried food consumption [34] and sugar sweetened drinks [35] have previously been reported in a cohort of US health professionals, these dietary variables were not included in the diet question-

| Trait | $h^2_{\mathrm{G}}$ (s.e) | $h^2_{\mathrm{GxE}}$ (s.e) |
|---|---|---|
| log BMI | 0.274 (0.056) | 0.093 (0.028) |
| PP | 0.228 (0.051) | 0.125 (0.028) |
| SBP | 0.251 (0.05) | 0.039 (0.023) |
| DBP | 0.254 (0.05) | 0.016 (0.02) |

Table 1: **Partitioned heritability estimates for four quantitative traits in the UK Biobank**. Heritability estimates obtained using common imputed SNPs (MAF$> 0.01$ in the full UK Biobank cohort) with RHE-LDMS. GxE heritability estimates were were obtained using the ES from each model fit. All analyses controlled for the same covariates used in the WGR analysis (including the top 20 principal components). Abbreviations; s.e, standard error estimated using the block jacknife (see **Online Methods**); $h^2_{\mathrm{G}}$, heritability due to additive genetic effects; $h^2_{\mathrm{GxE}}$, heritability due to multiplicative GxE effects; RHE, randomised HE-regression[22,23]; LDMS, SNPs stratified by minor allele frequency and LDscore (20 components).

naire used by the UK Biobank.

The ES for PP was dominated by the effects of age and gender (age, age$^2$, age-x-gender, gender together explained $94.9\%$ of variance in the ES). The magnitude of the ES was strongly associated with increased age whilst the sign of the ES was strongly associated with gender, implying that GxE effects were stronger in the elderly but acted in the opposite direction in men and women (**Figure S7**).

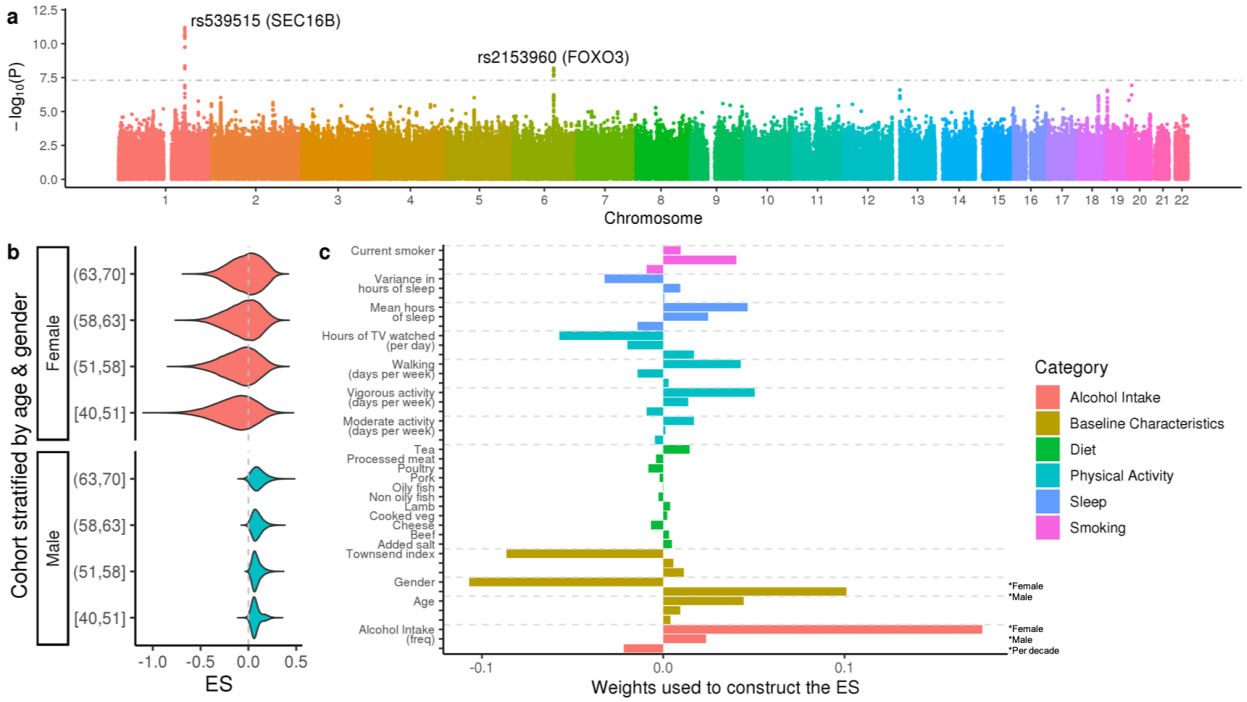



Figure 4: **GxE analysis of** log**BMI in the UK Biobank.** (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes $(p = 5 \times 10^{-8})$, $p$-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female specific weights is computed for each gender separately. Age is computed as the number of decades aged from 40. See **Online Methods** for details.

Similarly we observed that variance of the ES increased with age in both SBP and DBP, but instead of age itself being highly weighted we found that age interactions with other environmental variables were most important for explaining variation in the ES. Specifically for SBP we found that age interactions with smoking, Townsend index and alcohol frequency explained $86\%$ of variance in the ES (**Figure S8**) When compared to the cohort average, we found that participants in the top $5\%$ of the SBP ES were older (63 vs 58), had higher Townsend deprivation index (1.2 vs -1.74) and were more likely to smoke ($59\%$ vs $9\%$) whereas those in the bottom $5\%$ were also older ($65$ vs $58$), predominantly female ($91.5\%$), rarely drank alcohol ($43.9\%$ drank "Never") and had low Townsend deprivation index (-2.9 vs -1.74) (**Table S6**).

Finally we observed notably higher variance in the ES for DBP amongst men, most of which appeared to be driven by high gender-specific weights for smoking status and alcohol frequency (**Figure S9**). We further observed that alcohol frequency and smoking status became increasingly influential with age. The total SNP-GxE heritability for this ES however was quite low.

When testing for significant GxE interactions between the estimated ESs and imputed markers across the genome we observed that use of the robust standard errors made a noticeable difference to the calibration of LEMMA (**Figure S6**; **Table S4**). We identified two loci for $\log$BMI (**Figure 4a**), one locus for DBP (**Figure S9a**) and zero loci for SBP and PP, using a threshold of $5 \times 10^{-8}$ for genome wide significance (**Table 2**). For $\log$BMI, LEMMA identified GxE interactions at rs2153960 ($p = 6.5 \times 10^{-9}$; **Figure S10**) and at rs539515 ($p = 6.5 \times 10^{-12}$; **Figure S11**). rs2153960 is an intron in the *FOXO3* gene and has been previously associated with

Insulin-like growth factor 1 (IGF-I) concentration in a cohort of $10k$ middle aged Europeans [36]. IGF-I is known to be a a central mediator of metabolic, endocrine and anabolic effects of growth hormone and is also involved in carbohydrate homeostasis [36]. rs539515 is located 6kb downstream of *SEC16B*. Multiplicative GxE interactions have been reported at *Sec16B* with multiple environmental variables in a similar analysis in the UK Biobank[7], and with physical activity separately in Europeans [5] ($p = 0.025$) and in Hispanics [37] ($p = 8.1 \times 10^{-5}$). Highly significant variance effects ($p = 3.88 \times 10^{-17}$), which can be indicative of GxE, have also been reported at the *Sec16B* locus using $N = 456,422$ Eurpoeans in the UK Biobank[38]. GxE interactions have been reported at *Sec16B* with multiple environmental variables in a similar analysis in the UK Biobank[7] and with physical activity separately in Europeans [5] ($p = 0.025$) and in Hispanics [37] ($p = 8.1 \times 10^{-5}$). *SEC16B* transcribes one of the two mammalian homologues of the Sec16 protein, which has a key role in organizing endoplasmic reticulum exit sites by interacting with COPII components [39]. Although several GWASs have identified associations between *SEC16B* [40,41], the relevance of *SEC16B* to BMI is not well characterised [42]. Some evidence exists to suggest that *SEC16B* has role in the transport of peroxisome biogenesis factors; peroxisomes being an organelle involved in the catabolism of long chain fatty acids found ubiquitously in eukaryotic cells. Previous authors[41] have also speculated that the *SEC16B* might play a role in the transport of appetite regulatory peptides, however we are not aware of any evidence for this theory.

The DBP associated SNP is rs8090962 which is located within an enhancer, approximately 100KB downstream of the *SEC11C* gene and 50KB upstream of *ZNF532*. Neither gene has previously been associated with blood pressure traits, consistent with our analysis that shows no SNPs

with genome-wide significant main effects within 250KB (**Figure S12**)

**Comparison of methods on UK Biobank data** To compare LEMMA with existing single SNP methods we also ran StructLMM, the F-test and the robust F-test on logBMI using the same set of environmental variables as used by LEMMA (but not including the significant squared environments as covariates). Manhattan plots are displayed in **Figure S13**. Test statistics from both the F-test ($\lambda_{\text{GC}} = 1.37$) and StructLMM ($\lambda_{\text{GC}} = 1.235$) were substantially inflated when compared to the robust F-test and LEMMA ($\lambda_{\text{GC}} = 1.03$ and $\lambda_{\text{GC}} = 1.062$ respectively; see Table S3), suggesting that StructLMM does not properly control for heteroskedasticity. There are clear differences between the 4 methods, especially amongst SNPs with suggestive evidence of GxE Interaction results (**Figure S14a**). LEMMA did not find find the FTO locus, StructLMM and F-test did not find the SEC16B locus, and the robust F-test only found the FTO locus.

LEMMA relies on the assumption that all GxE interaction effects for a single trait share a common ES, and we have shown in simulation that when this assumption holds LEMMA achieves substantial increases in power. However we would expect LEMMA to have little power to detect SNPS which interact with a combination of environments that is not well correlated with the genome-wide ES estimated by LEMMA. The *FTO* seems to be one clear example of this. We extracted an estimate of the SNP specific interaction profile at *FTO* using the robust F-test (**Online Methods**), we found that it's correlation with LEMMA's ES was low (Pearson $r^2 = 0.3$). In comparison, a similar analysis at *SEC16B* and *FOXO3* yielded much higher correlations (Pearson $r^2 = 0.725$ and $r^2 = 0.713$ respectively).

## Discussion

In this study we proposed a new method, LEMMA, that uses whole genome regression methodology to estimate a single environmental score (ES) that interacts with SNPs across the genome. In simulation we have demonstrated that the ES can be used to compute well calibrated p-values of the multiplicative interaction effect at each SNP and to quantify the contribution of MAF and LD stratified multiplicative GxE interaction effects to trait heritability. In analyses of four quantitative traits in the UK Biobank, we have demonstrated that GxE effects amongst common imputed SNPs make a non-trivial contribution to the heritability of $\log$BMI and PP ($9.3\%$ and $12.5\%$ respectively). Our stratified heritability analysis has suggested that GxE interactions for these traits are mostly driven by low frequency variants. Our analysis identified three loci with statistically significant GxE interaction effects. Two of these loci, rs539515 (*FOXO3*) and rs8090962, are novel and for the other, rs539515 (*SEC16B*), we show stronger evidence for than in the previous study[7].

Robinson *et al.*[6] have previously attempted to quantify the contribution of GxE interactions to the heritability of BMI in a study performed on imputed SNPs from the interim UK Biobank release. Using the GCI-GREML model implemented in GCTA[43] and eight environmental variables that included measures of smoking, hours of TV watched and alcohol frequency, Robinson *et al.*[6] reported that only smoking had significant GxE heritability ($4.0\%$). In contrast, the ES estimated for $\log$BMI in our analysis had upweighted contributions from many environmental variables, including hours of TV watched and smoking, suggesting that multiple environmental variables can influence on the genetic predisposition to BMI. Modelling these environmental variables jointly

allowed LEMMA to capture a combination whose GxE interactions explained $9.3\%$ of heritability.

We have also evaluated the performance of three existing single SNP methods (StructLMM, the F-test and a robust F-test), both in simulation and on $\log$BMI from this same dataset. In simulation with large datasets we observed that StructLMM and the F-test had similar performance; an observation that also held in our analysis of $\log$BMI. Both of these methods appeared vulnerable to heteroskedasticity, which we showed is likely to occur in traits with non-trivial GxE heritability. A simple adjustment, using 'robust' or Huber-White variance estimators, solved this problem. The two F-test methods further benefit from a wealth of existing theory[44] and, being theoretically simpler than StructLMM, could be easily implemented as an R-plugin with PLINK[45] (for example [46]). In our opinion, the robust F-test is therefore the most appropriate of the three single SNP methods to model GxE effects with tens of environments in biobank scale datasets.

Although LEMMA represents a method with increased power to detect GxE interaction effects, our approach does have some caveats. First the gain in power is dependent on a strong assumption on the underlying genetic architecture. Whilst our analysis suggests that this does hold to some extent for PP and $\log$BMI, this may not be the case for other traits. Second, despite much effort to provide an efficient implementation, the LEMMA algorithm is still a computationally demanding. Using randomised HE-regression to estimate an improved initialisation of the interactions weights may help to reduce runtime, and is a avenue that we are currently pursuing. Finally, for simplicity LEMMA currently searches only for GxE interactions with a single linear combination of environments. Generalising the LEMMA approach to several orthogonal linear

combinations or using functional annotation to restrict the SNPs that each ES interacts with, may yet yield more power to identify interactions in complex traits and elucidate unknown biology.

Table 2: Loci with genome-wide significant GxE interaction effects with the ES

| | | | | | | | | | p-values | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Trait** | **SNP** | **Chr** | **BP** | **A0** | **A1** | **AF** | **GxE beta** | **GxE SE** | **Main effect** | **GxE** | **Nearest gene** |
| log(BMI) | rs539515 | 1 | 177,889,025 | A | C | 0.208 | -0.01166 | 0.00170 | $1.6 \times 10^{-60}$ | $6.5 \times 10^{-12}$ | *SEC16B* |
| log(BMI) | rs2153960 | 6 | 108,988,184 | G | A | 0.713 | -0.00984 | 0.00170 | $1.6 \times 10^{-8}$ | $6.5 \times 10^{-9}$ | *FOXO3* |
| DBP | rs8090962 | 18 | 56,694,404 | A | G | 0.443 | 0.00867 | 0.00157 | $4.5 \times 10^{-1}$ | $3.6 \times 10^{-8}$ | *OACYLP / SEC11C* |

Independent loci with genome-wide significant ($P < 5 \times 10^{-8}$) GxE interaction effects with the environmental score (ES). Loci at least 0.5cM apart were judged to be independent. SNP locations follow the GrCh37 human genome assembly. Abbreviations are as follows; A0, reference allele; A1, alternative allele; AF, reference allele frequency. All loci had IMPUTE INFO score $> 0.99$.

## Online Methods

**Linear Environment Mixed Model Analysis (LEMMA)** The standard LMM used in genome wide association studies is written as

$$y = C\alpha + u + \epsilon,$$

where $y$ is the centred and scaled $N \times 1$ vector of phenotypes, $C$ is an $N \times L\prime$ matrix of covariates with fixed effects $\alpha$, and $u$ and $\epsilon$ are $N$-vectors of unobserved polygenic and residual effects vectors respectively. Typically $u$ is modelled as gaussian with mean zero and covariance matrix $\sigma_g^2 K$. Specification of the kinship matrix $K$ is an area of active research [47–50] but the most simple approach is to let $K = XX^T/M$ where $X$ is the $N \times M$ genotype matrix and columns of $X$ (which usually correspond to SNPs) are normalised to have mean zero and variance one. This can equivalently be written as a Bayesian whole genome regression (WGR) model

$$y = C\alpha + X\beta + \epsilon,$$

where

$$\beta \sim \mathcal{N}(0, \sigma_g^2/M).$$

Here $\beta$ is a $M$-vector modelling the random effect of each SNP. This form corresponds to the so called infinitesimal model where every SNP is allowed to have a small but non-zero effect on a given trait. To generalise the model to a non-infinitesimal genetic architecture, we model SNP effects with a mixture of gaussians prior. This approach has been applied previously in human genetics [13,21] and by the 'Bayesian alphabet' of genomic prediction methods in the animal breeding literature [17,51,52].

We extend this setup to model GxE interactions genome-wide with a linear combination of multiple environmental variables using

$$Y = C\alpha + X\beta + Z\gamma + \epsilon$$

where

$$Z = \eta \odot X,$$

$$\eta = Ew,$$

$$w \sim \mathcal{N}(0, 1)$$

where $E$ is an $N \times L$ matrix of environmental variables that could potentially be involved in GxE interactions, and $\eta$ is the linear combination of those environments that is learnt in tandem with SNP effects. We note that all environmental variables contained in $E$ are also be contained in $C$. We chose to model the interaction weights $w$ with a gaussian prior, but in theory one could consider sparser priors such as a spike and slab. We use the notation $\eta \odot X$ for the element-wise product of $\eta$ with each of the columns of $X$. In effect, $Z$ contains all of the multiplicative interaction terms of $\eta$ with all of the genetic variants.

We use mixture of normals priors on both the main effects and the interaction effects, such that

$$\beta_j | \sigma_e^2, \lambda_\beta, \sigma_{\beta,1}^2, \sigma_{\beta,2}^2 \sim \lambda_\beta \mathcal{N}(0, \sigma_e^2 \sigma_{\beta,1}^2) + (1 - \lambda_\beta)\mathcal{N}(0, \sigma_e^2 \sigma_{\beta,2}^2),$$

$$\gamma_j | \sigma_e^2, \lambda_\gamma, \sigma_{\gamma,1}^2, \sigma_{\gamma,2}^2 \sim \lambda_\gamma \mathcal{N}(0, \sigma_e^2 \sigma_{\gamma,1}^2) + (1 - \lambda_\gamma)\mathcal{N}(0, \sigma_e^2 \sigma_{\gamma,2}^2)$$

and standard normal priors on the covariate and error terms.

$$\alpha|\sigma_\alpha^2 \sim \mathcal{N}(0, \sigma_\alpha^2),$$

$$\epsilon|\sigma_e^2 \sim \mathcal{N}(0, \sigma_e^2).$$

**Variational Inference** For notational convenience we define $\theta = \{\alpha, \beta, \gamma, w\}$ as the set of latent variables, $\mathcal{D} := \{X, E\}$ the genetic and environmental data and $\phi$ as the set of hyperparameters. Then the posterior $p(\theta|y, \mathcal{D}, \phi)$ is given by

$$p(\theta|y, \mathcal{D}, \phi) \propto p(y|\theta, \mathcal{D}, \phi) \prod_c p(\tau_c|\phi) \prod_l p(w_l) \prod_j p(\beta_j, u_j|\phi) \prod_j p(\gamma_j, v_j|\phi).$$

To evaluate the posterior we use the variational inference framework; approximating the true posterior $p(\theta|y, \mathcal{D}, \phi)$ with a tractable alternative distribution $q(\theta; \nu)$ governed by (variational) parameters $\nu$. To make inference tractable we use the standard Mean Field assumption so that $q(\theta; \nu)$ factorises

$$q(\theta; \nu) = \prod_c q(\tau_c) \prod_l q(w_l) \prod_j q(\beta_j, u_j) \prod_j q(\gamma_j, v_j). \tag{1}$$

To make $q(\theta; \nu)$ a close approximation of the true posterior we minimise the KL Divergence between $q(\theta; \nu)$ and $p(\theta|y, \mathcal{D}, \phi)$ with respect to variational parameters $\nu$. In this manner, the problem has been transformed from one of computing posterior distributions into one of optimization. We can show that minimising the KL Divergence is equivalent to maximising a lower bound on

the marginal log likelihood by observing

$$\begin{aligned} KL(q||p) &= -\mathbb{E}_q\left[\log \frac{p(\theta|y,\mathcal{D},\phi)}{q(\theta;\nu)}\right], \\ &= -\mathbb{E}_q\left[\log \frac{p(\theta,y|\mathcal{D},\phi)}{q(\theta;\nu)}\right] + \mathbb{E}_q\left[\log p(y|\mathcal{D},\phi)\right], \\ &= -\mathbb{E}_q\left[\log \frac{p(\theta,y|\mathcal{D},\phi)}{q(\theta;\nu)}\right] + \log p(y|\mathcal{D},\phi), \end{aligned}$$

thus we can write

$$\mathcal{F}(\nu;\phi) := \mathbb{E}_q\left[\log \frac{p(\theta,y|\mathcal{D},\phi)}{q(\theta;\nu)}\right] \leq \log p(y|\mathcal{D},\phi).$$

Here $\mathcal{F}(\nu;\phi)$ is commonly referred to as the Evidence Lower Bound (ELBO). Due to the factorised form of Equation (1) we can cyclically update the approximate distribution for each latent variable in turn until we reach convergence.

Our model depends on a set of eight hyperparameters $\phi = \{\sigma_e^2, \{\sigma_{\beta,i}^2\}_{i=1}^2, \{\sigma_{\gamma,i}^2\}_{i=1}^2, \lambda_\beta, \lambda_\gamma, \sigma_\alpha^2\}$. We set $\sigma_\alpha^2$ to a large constant to create a flat prior on the covariates, leaving seven unknowns. Similar methods have used cross-validation to estimate the hyperparameters [13] (choosing the best based on Mean Squared Error) or performed a grid search over hyperparameter values [20] (choosing the best based on a lower bound on the log likelihood). For LEMMA a grid search would be computationally demanding, both because the set of hyperparameters is larger and because we cannot efficiently perform multiple runs in parallel as done by Loh *et al.*[13]. Instead we maximise a lower bound on the approximate log likelihood (the so called Evidence Lower Bound or ELBO) with respect to the hyperparameters. In this manner our approach can be viewed either as a variational expectation maximisation algorithm [53,54].

Similar to the EM algorithm, the hyperparameter maximization step can lead to slow exploration of the hyperparameter space and thus to slow convergence of the LEMMA algorithm. We use an accelerator, SQUAREM [55], to speed up convergence. Given two estimates of the hyperparameters $\phi_{t-2}$ and $\phi_{t-1}$ we can adjust the maximised estimate $\phi_t$ with

$$\tilde{\phi}_t(v_t) = \phi_{t-2} - 2v_t\Delta\phi_{t-1} + v_t^2\Delta^2\phi_t,$$

where $\Delta\phi_{t-1} = \phi_{t-1} - \phi_{t-2}$ and $\Delta^2\phi_t = \phi_t - 2\phi_{t-1} + \phi_{t-2}$. Thus the new adjusted estimate $\tilde{\phi}_t(v_t)$ is a continuous function of the step size $v_t$, which yields the original estimate $\phi_t$ for $v_t = -1$. As recommended by [55] we set $v_t = \min(-1, -||\Delta\phi_{t-1}||_2^2/||\Delta^2\phi_t||_2^2)$. Occassionally this yields an estimate that is either outside of the domain of $\phi$ or leads to a state with worse ELBO than the previous state. For the first issue we use a simple backtracking method of halving the distance between $v_t$ and $-1$, and for the second we simply judge model convergence when the absolute change in ELBO drops below a given threshold. We judge the algorithm to have converged when a full pass through all latent variables yields an absolute change of less than $0.01$ in the aproximate log-likelihood (ELBO).

**Identifying GxE associated loci** After convergence of the LEMMA variational inference algorithm, we obtain posterior mean estimates of $\hat{\beta}, \hat{\gamma}$ and $\hat{\eta} = E\hat{w}$. From these we construct residualised phenotypes following a Leave One Chromosome Out (LOCO) scheme;

$$y_{\text{resid-LOCO}} = y - C\hat{\alpha} - X_{\text{LOCO}}\hat{\beta}_{\text{LOCO}} - \hat{\eta} \odot X_{\text{LOCO}}\hat{\gamma}_{\text{LOCO}}.$$

$X_{\text{LOCO}}$ denotes the genotype matrix excluding SNPs on the same chromosome of the test SNP, and $\beta_{\text{LOCO}}$ and $\gamma_{\text{LOCO}}$ are constructed similarly. Using a LOCO scheme has been shown to increase

power in LMMs as the effect of the test SNP is conditioned on the effects on a large proportion of the rest of the genome [12,15].

For each imputed SNP, we then perform hypothesis tests $\beta_{\text{test}} \neq 0$ and $\gamma_{\text{test}} \neq 0$ using the linear model

$$y_{\text{resid-LOCO}} = x_{\text{test}}\beta_{\text{test}} + (\hat{\eta} \odot x_{\text{test}})\gamma_{\text{test}} + \epsilon,$$

$$= H\tau + \epsilon.$$

Assuming that $\epsilon$ has mean zero and covariance matrix $\Omega$ we can use the standard OLS estimator

$$\hat{\tau} = (H^T H)^{-1} H^T y,$$

which (under certain regularity conditions) is asymptotically normally distributed with mean $\tau$ and variance $\text{Var}\,(\hat{\tau})$. By assuming the residual phenotype is homoskedastic, that is that $\Omega = \hat{\sigma}_e^2 I$, we can obtain the usual variance estimator given by

$$\text{Var}\,(\hat{\tau}) = \hat{\sigma}_e^2 (H^T H)^{-1}.$$

[46] have previously observed that GxE interaction tests are likely to suffer from conditional heteroskedasticity, and hence the homoskedastic variance estimator is likely to underestimate the true variance [56]. We explain this phenomenon in detail in the **Supplementary Material**.

To overcome this we use robust standard errors, alternatively called Huber-White, sandwich or heteroskedastic-consistent errors [57,58], that are standard tools in economics [44] and have previously been proposed for use in GxE interaction studies [24,46,59]. We further include a small adjustment that reduces bias in small samples [60]. This yields the variance estimator

$$\text{Var}\,(\hat{\tau}) = (H^T H)^{-1} H^T \hat{\Sigma} H (H^T H)^{-1},$$

where $\hat{\Sigma}$ is a diagonal matrix with $\hat{\Sigma}_{ii} = \frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}$, where $\hat{\epsilon} = y - H\hat{\tau}$ and $h = H(H^TH)^{-1}H^T$. Hence our GxE test statistic is given by

$$\frac{\hat{\gamma}_{\text{test}}^2}{\text{Var}\left(\hat{\gamma}_{\text{test}}\right)}$$

and under the null hypothesis is asympototically distributed as $\chi_1^2$. As main effects tests are not sensitive to assumptions of heteroskedacity in the same way that GxE tests are [46], we use a simple t-test to test the hypothesis $\beta_{\text{test}} \neq 0$.

**Heritability estimation** Previous genome wide regression methods [20,31,61] have shown that it is possible to rearrange the model hyperparameters to gain an estimate of trait heritability. We find that in our variational framework, this approach underestimates trait heritability due to the tendency of mean field variational inference to underestimate the posterior variance of each parameter. Instead we treat the posterior mean $\hat{\eta}_{\text{LEMMA}}$ as a fixed effect, and use randomised HE-regression[22,23,62] to estimate heritability with a single SNP component[22] (RHE-SC) and multiple SNP components[23] (RHE-LDMS). When using the multi-component model, SNPs are stratified into a total of 20 bins; using 5 MAF bins ($\leq 0.1$, $0.1 <$ MAF $\leq 0.2$, $0.2 <$ MAF $\leq 0.3$, $0.3 <$ MAF $\leq 0.4$, $0.4 <$ MAF $\leq 0.5$) and 4 LD score quantiles.

The single component model is given by

$$y \sim \mathcal{N}\left(E\alpha, \sigma_\beta^2 K + \sigma_\gamma^2 \hat{V} + \sigma_e^2 I\right),$$

where $K = XX^T/M$, $\hat{V} = Z(\hat{\eta})Z(\hat{\eta})^T/M$ and $Z(\hat{\eta}) = \text{diag}(\hat{\eta})X$. HE-regression is a method of moments estimator that fits the variance components $(\sigma_\beta^2, \sigma_\gamma^2, \sigma_e^2)$ to minimise the difference

between the empirical and expected covariances. This is mathematically equivalent to solving the following linear system

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

Wu *et al.*[22] showed that this system can be solved in $\mathcal{O}(NMB)$ time (for small $B$) without ever forming the kinship matrices $K$ and $V$ using Hutchinson's estimator, and that covariates can be efficiently projected out of the phenotype, genotypes and interaction matrix $Z$ with minimal additional cost. Pazokitoroudi *et al.*[23] give an extension to multiple components and showed that variance estimates can be obtained with the block jackknife.

Speed *et al.*[26] show that the usual form for $h_G^2$, the proportion of trait variance explained by additive genetic effects, given by

$$\hat{h}_G^2 = \frac{\hat{\sigma}_\beta^2}{\hat{\sigma}_\beta^2 + \hat{\sigma}_\gamma^2 + \hat{\sigma}_e^2},$$

holds only when genotype matrix $X$ is standardised to have column mean zero and column variance one. Whilst this is true in expectation for $\hat{Z}$ (assuming that $\text{Cov}(\hat{\eta}, X_j) = 0, \quad \forall j \in \{1, M\}$), this is not guaranteed. To obtain column mean zero we include an intercept of ones amongst the covariates that are projected out of the phenotype, genotypes and interaction matrix. To account for columns having variance not equal to one, we use a more general form of the heritability estimator (see Speed *et al.*[26] for details)

$$\hat{h}_{\text{GxE}}^2 = \frac{\hat{\sigma}_\gamma^2 \text{tr}\left(\hat{V}\right)/N}{\sigma_\beta^2 + \hat{\sigma}_\gamma^2 \text{tr}\left(\hat{V}\right)/N + \hat{\sigma}_e^2}.$$

**Implementation and computational efficiency** We provide software implementing the LEMMA algorithm in C++ from `https://jmarchini.org/lemma/`. We implement a number of steps to improve computational and memory efficiency including vectorization using SIMD extensions, compressed data formats, pre-computing quantities, parallel computing with OpenMPI, use of the well optimised Intel Math Kernel Library. Full details are given in the **Supplementary Material**.

**Detecting squared environmental dependance** By default, each of the $L$ environmental variables is tested against the phenotype for significant squared effects. To do this LEMMA tests the hypothesis $\beta_l \neq 0$ using the following linear model

$$y = \mathbf{1}\alpha_0 + C\alpha + E_l^2\beta_l + \epsilon.$$

The squared effect of any environmental variables with a p-value less than $0.01$ (Bonferroni correction for $L$ multiple tests) are added to the matrix of covariates $C$.

**Controlling for covariates** Unlike in BOLT-LMM [13], it is not possible to efficiently project covariates out of the model $(y, X, Z)$, because the multiplicative interaction matrix $Z$ changes after each pass through the data. Instead the LEMMA software package can either regress covariates out of the phenotype or model the covariates as random effects in the variational framework. For our analyses of the UK Biobank we included all covariates within the variational model.

**Comparison to existing GxE methods** We compare LEMMA to three other single SNP methods that jointly model interactions with multiple environments. The first comparison method,

StructLMM [7], is a method that uses a random effects term $u$ to model environmental similarity instead of genetic similarity. Specifically, StructLMM uses the model

$$y \sim \mathcal{N}(C\alpha + x_{\text{test}}\beta, \sigma^2_{\text{GxE}}\text{diag}\,(x_{\text{test}})\,\Sigma\text{diag}\,(x_{\text{test}}) + \sigma^2_e\Sigma + \sigma^2_n I),$$

to test the hypothesis $\sigma^2_{\text{GxE}} \neq 0$. Here $C$ is the matrix of covariates with fixed effects $\alpha$, $x_{\text{test}}$ is the focal variant and $\Sigma = EE^T$ is the environmental similarity matrix (where $E$ is an $N \times L$ matrix of environmental variables). Although StructLMM both an interaction test and a joint test that looks for non-zero main and interaction effects at each SNP, we use only the interaction test in our comparisons. Finally we note that StructLMM recommends 'gaussianizing' the phenotype as a preprocessing step; however we just center and scale the phenotype for consistency with our other methods.

Our second and third comparison methods use equivalent information to StructLMM in a fixed effects framework. Consider the linear model

$$\begin{aligned} y &= C\alpha + E\alpha' + x_{\text{test}}\beta_{\text{test}} + x_{\text{test}} \odot E\gamma + \epsilon, \\ &= H\tau + \epsilon, \end{aligned}$$

where $H$ is formed from column-wise concatenation of $[C, E, x_{\text{test}}, \text{diag}(x_{\text{test}})E]$ and $\tau$ is the corresponding vector of fixed effects. Let $R$ be the indicator matrix such that $R\tau = \gamma$. We wish to test the null hypothesis $H_0 : \gamma = 0$. Assuming that $\epsilon$ has mean zero and covariance matrix $\Omega$ we can use the standard OLS estimator $\hat{\tau} = (H^T H)^{-1} H^T y$ which (under certain regularity conditions) is asympototically distributed as normal with mean $\tau$ and variance given by $\text{Var}\,(\hat{\tau}) = (H^T H)^{-1} H^T \Omega H (H^T H)^{-1}$. Assuming homoskedacity yields the standard F test statis-

tic

$$F_{\text{test}} = \frac{(R\hat{\tau})^T (R(H^T H)^{-1} R^T)^{-1} (R\hat{\tau})/L}{\hat{\sigma}_e^2},$$

which follows an $F_{d_1-d_0,N-d_1}$ distribution under the null hypothesis, where $d_1$ is the column rank of $H$ and $d_0$ is the column rank of H under the null hypothesis. Alternatively we can use the same robust standard error used in the LEMMA test statistic

$$F_{\text{robust}} = (R\hat{\tau})^T (R(H^T H)^{-1} H^T \hat{\Omega} H (H^T H)^{-1} R^T)^{-1} (R\hat{\tau}),$$

where $\hat{\Omega}$ is a diagonal matrix with $\hat{\Omega}_{ii} = \frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}$, $\hat{\epsilon} = y - H\hat{\tau}$ and $h = H(H^T H)^{-1} H$. Then $F_{\text{robust}}$ is asympototically distributed as $\chi^2_{d_3}$ where $d_3$ is the rank of $HR^T$. In our simulations we refer to this as the robust F-test.

**SNP specific interaction profile** The SNP specific interaction profile is defined as $\eta_{LS} = Ew_{LS}$ where $w_{LS}$ is the least squares parameter estimate of $w$ in the single SNP model

$$y = C\alpha + x_{\text{test}}\beta_{\text{test}} + x_{\text{test}} \odot E\boldsymbol{w} + \epsilon$$

and $y, C$ and $E$ are the data matrices defined in **Online Methods**. The correlation between $\eta_{LS}$ for a given SNP and the ES estimated by LEMMA can be viewed as a proxy for how well LEMMA captures the GxE interactions at that locus.

**UK Biobank analysis** We used real genotype and phenotype data from the UK Biobank, which is a large prospective cohort study of approximately $500,000$ individuals living in the UK [3]. To account for potential confounding effects of population structure, we first subset down to the white British subset of 344,068 individuals used by Bycroft *et al.*[3] in a GWAS on human height. This

represents unrelated individuals who self-report white British ethnicity and whose genetic data projected onto principal components lies within the white British cluster[3]. After subsetting down to individuals who had complete data across the phenotype, covariates and environmental factors (see below) we were left with approximately $280,000$ samples per trait (**Table S1**). Finally we filtered genetic data based on minor allele frequency ($\geq 0.01$) and IMPUTE info score ($\geq 0.3$), leaving approximately 642,000 variants per trait (**Table S1**). Association testing was performed across $10,295,038$ imputed SNPs. For each trait we included age$^3$, age$^2\times$ gender, age$^3\times$ gender, a binary indicator for the genotype chip and the top $20$ genetic principal components as additional covariates.

BMI was derived from height and weight measurements made during the first assessment visit (instance '0' of field 21001). We set BMI readings that were more than six standard deviations from the population mean to missing, and then applied a log transformation.

After calculating the mean SBP and DBP using automated blood pressure readings from the first assessment visit (fields 4080 and 4079), we adjusted for medication usage by adding $15$mmHg and $10$mmHg to SBP and DBP respectively [63]. Data from manual measurements (fields 93 and 94) was used in the rare instance that no automated reading was available. Blood pressure readings more than four standard deviations from the mean were set to missing. PP was then calculated as SBP minus DBP.

For our GxE analyses, we made use of 42 environmental variables from the UK Biobank, similar to those used in previous GxE analyses of BMI in the UK Biobank [7,25]. From the data pro-

vided by the UK Biobank we included 7 continuous environmental variables ("Age when attended assessment centre", "Sleep duration", "Time spent watching television", "Number of days/week walked 10+ minutes", "Number of days/week of moderate physical activity 10+ minutes", "Number of days/week of vigorous physical activity 10+ minutes", "Townsend deprivation index at recruitment"), 1 ordinal environmental variable ("Alcohol intake frequency"), 9 dietary ordinal variables ("Salt added to food", "Oily fish intake", "Non-oily fish intake", "Processed meat intake", "Poultry intake", "Beef intake", "Lamb intake", "Pork intake", "Cheese intake") and 2 dietary continuous variables ("Tea intake", "Cooked vegetable intake"). We further derived 1 categoric variable ("Is Current Smoker" from the responses given in the UK Biobank field "Smoking status") and 1 continuous variable ("Sleep sd"; the number of standard deviations from the population mean sleep duration). For analyses of blood pressure, we additionally included one further continuous variable "Waist circumference". This left 11 dietary variables and 10 non-dietary variables (11 for blood pressure traits). In addition we included multiplicative interactions between participants age and gender with all non-dietary variables, and included the main effect of gender, giving the data matrix $E$ a total of 42 columns (45 for blood pressure traits). Before running LEMMA each column was standardised as

$$E_{ij} = \frac{E_{ij} - \text{mean}(E_{:,j})}{\text{sd}(E_{:,j})}.$$

In all cases, where participants responded with "Prefer not to answer", "Do not know" or "None of the above" we set the value to missing. For 3 continuous variables ("Time spent watching television", "Tea intake", "Cooked vegetable intake") we removed the 99'th percentile and for "Sleep duration" we removed both the 1st and 99th percentiles.

After running LEMMA, we found it convenient to interpret weights corresponding to a rescaled data matrix $E_1$. Assuming the column space of $E$ and $E_1$ is the same, weights $\boldsymbol{w_1}$ that correspond to $E_1$ can be extracted from the ES using least squares

$$\boldsymbol{w}_1 = (E_1^T E_1)^{-1} E_1^T \hat{\eta}_{\text{LEMMA}},$$

where $\hat{\eta}_{\text{LEMMA}}$ represents the ES. We note that although multivariate linear regression is invariant to a rescaling of the design matrix, ridge regression is not due to the penalisation place on the magnitude of the learned parameters. However, as the magnitude of the weights from our UK Biobank analysis is typically small (less than 0.2) compared to the standard deviation of our gaussian prior (1) in this case the rescaling makes minimal difference.

Recoded data matrix $E_1$ was formed with one column for each of the 11 dietary variables (normalised to have mean zero and variance one), and three columns for each of the 10 (11 for blood pressure traits) non-dietary variables; the first augmented by a binary male indicator vector, the second by a binary female indicator vector and the third by a continuous vector of participant age. Columns augmented by male and female binary indicator vectors where normalised to have mean zero and variance one (not including zeros due to augmentation), apart from age (scaled to represent the number of decades aged past 40). Columns augmented by age were normalised first and then multiplied by age on the per decade scale. We further included indicator columns for men and women, which can be interpreted as gender specific intercepts and is equivalent to including an intercept and a binary column for only one gender (men or women). We note this leaves 43 (46) columns where the extra column comes from including an intercept within the column space of $E_1$ and is necessary because some columns have mean not equal to zero. Thus the column space of

$E_1$ is equivalent to $E$ under the constraint that the ES has mean zero.

**Simulation studies** Genetic data was subsampled from the UK Biobank, using $N = 25K$ unrelated individuals of mixed ancestry and $M = 100K$ genotyped SNPs, and $L$ environmental variables were simulated from a standard gaussian. We constructed phenotypes with $2500$ causal main effects and $1250$ causal interaction effects explaining $20\%$ and $5\%$ of trait variance respectively. For each phenotype we constructed a weighted average of the environmental variables which we used to simulate multiplicative interaction effects. Environments with a non-zero weight are referred to as active. All non-zero effects were drawn from SNPs in the first half of each chromosome, allowing us to test the calibration of each method on 'null' SNPs from the second half of each chromosome. To allow direct power comparisons across different scenarios we included an additional $60$ SNPs with standardised effect sizes, that together accounted for $1\%$ of trait variance with their main effects and $1\%$ of trait variance with their interaction effects. Finally, a further $1\%$ of trait variance was modelled using the first genetic principal component (PC). For all methods we included the first genetic PC as a covariate. For each method we calculated power as the proportion of the SNPs of standardised effect identified at a threshold of $p < 0.01$.

In simulations used to test HE-regression, phenotypes were constructed with $10000$ causal main effects explaining $20\%$ of trait variance and, in simulations with non-zero GxE heritability, $10000$ causal SNPs with interaction effects.

**Model misspecification** We simulated a scenario where a diease trait $Y$ depends non-linearly on an environmental factor $S$ which has it's own genetic basis. More explicity suppose that $X$ is the

centered and scaled genotype matrix so that columns have mean zero and variance one, that $S$ is modelled as

$$S = X\tau + \epsilon_s,$$

where $\epsilon_s \sim \mathcal{N}(0, (1 - h^2_\tau)I)$, $\tau$ models random SNP effects for $S$ and trait $Y$ is given by

$$Y|a = aS^2 + X\beta + \epsilon.$$

Here $a$ is a constant that we use to control the strength of the contribution of $S^2$ to $Y$, $\epsilon \sim \mathcal{N}(0, (1 - h^2_\beta))$ and $\beta$ is the random SNP effects for $Y$. For simulation we suppose that $\tau$ and $\beta$ have spike and slab priors

$$\tau_j|v_j \sim v_j\mathcal{N}\left(0, \frac{h^2_\tau}{P\lambda_\tau}\right) + (1 - v_j)\delta_0(\tau_j)$$
$$\beta_j|u_j \sim u_j\mathcal{N}\left(0, \frac{h^2_\beta}{P\lambda_\beta}\right) + (1 - u_j)\delta_0(\beta_j),$$

$$v_j \sim \text{Ber}(\lambda_\tau).$$

$$u_j \sim \text{Ber}(\lambda_\beta).$$

**Data availability**

The genetic and phenotype datasets generated by UK Biobank analysed 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, without preferential or exclusive access for any person. More details are available here http://www.ukbiobank.ac.uk/register-apply/

## URLs

UK Biobank: http://www.ukbiobank.ac.uk

StructLMM as implemented in LIMIX 2.0.0: https://github.com/limix/limix

## 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 Kevin Sharp and David Steinsaltz for discussions about this work.

## Author contributions

J.M. and M.K. conceived the ideas for the model and methods development. M.K. conducted all analyses, derived the variational Bayes updates and developed the software that implemented the methods with guidance from J.M. M.K. and J.M. wrote the manuscript. J.M. carried out this work while affiliated with the University of Oxford.

## References

1. Hunter, D. J. Gene-environment interactions in human diseases. *Nature Reviews Genetics* **6**, 287–298 (2005). URL http://www.nature.com/doifinder/10.1038/nrg1578.

2. Smith, PG and Day, N. The design of case-control studies: The influence of confounding and interaction effects. *International Journal of Epidemiology* **13**, 356–365 (1984).

3. Bycroft, C. *et al.* Genome-wide genetic data on ˜ 500 , 000 UK Biobank participants. *bioRxiv* (2017).

4. Kilpeläinen, T. O. *et al.* Physical activity attenuates the influence of FTO variants on obesity risk: A meta-analysis of 218,166 adults and 19,268 children. *PLoS Medicine* **8** (2011). URL https://doi.org/10.1371/journal.pmed.1001116.

5. Ahmad, S. *et al.* Gene x Physical Activity Interactions in Obesity: Combined Analysis of 111,421 Individuals of European Ancestry. *PLoS Genetics* **9**, 1–9 (2013).

6. Robinson, M. R. *et al.* Genotype-covariate interaction effects and the heritability of adult body mass index. *Nature Genetics* **49**, 1174–1181 (2017). URL http://www.nature.com/doifinder/10.1038/ng.3912.

7. Moore, R. *et al.* A linear mixed model approach to study multivariate gene-environment interactions. *Nat Genet* 270611 (2018). URL https://www.nature.com/articles/s41588-018-0271-0.

8. Young, A. I., Wauthier, F. L. & Donnelly, P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nature Genetics (in press)* **50** (2018). URL http://www.nature.com/articles/s41588-018-0178-9.

9. de Leeuw, C. A., Stringer, S., Dekkers, I. A., Heskes, T. & Posthuma, D. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. *Nature Communications* **9** (2018).

10. Burgoine, T., Sarkar, C., Webster, C. J. & Monsivais, P. Examining the interaction of fast-food outlet exposure and income on diet and obesity: Evidence from 51,361 UK Biobank participants. *International Journal of Behavioral Nutrition and Physical Activity* **15**, 1–12 (2018).

11. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525–526 (2012). URL http://www.nature.com/articles/nmeth.2037.

12. 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). URL https://doi.org/10.1038/ng.2876. NIHMS150003.

13. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015). URL https://doi.org/10.1038/ng.2876. 15334406.

14. Eskin, E. *et al.* Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709–1723 (2008). 1305.3240.

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

16. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012). URL https://doi.org/10.1038/ng.2310. arXiv:1305.4366v2.

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

18. 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).

19. 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).

20. 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). arXiv:1208.4400v1.

21. Zhou, X. *et al.* Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics* **9**, e1003264 (2013). URL http://dx.plos.org/10.1371/journal.pgen.1003264.

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). URL https://www.biorxiv.org/content/10.1101/522003v2.

24. Tchetgen, E. J. T. & Kraft, P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. *Epidemiology* **22**, 257–261 (2011).

25. 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). URL https://doi.org/10.1038/ncomms12724.

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

27. 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). URL http://dx.doi.org/10.1038/s41588-018-0108-x.

28. 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). URL http://dx.doi.org/10.1038/ng.3390. 15334406.

29. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics* **51**, 1244–1251 (2019). URL https://doi.org/10.1038/s41588-019-0465-0.

30. Visscher, P. M., Goddard, M. E., Derks, E. M. & Wray, N. R. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Molecular Psychiatry* **17**, 474–485 (2012).

31. Powell, J. E. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics* **50**, 746–753 (2018). URL http://dx.doi.org/10.1038/s41588-018-0101-4.

32. Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLOS Genetics* **13**, e1006977 (2017). URL http://dx.plos.org/10.1371/journal.pgen.1006977.

33. Qi, Q. *et al.* Television watching, leisure time physical activity, and the genetic predisposition in relation to body mass index in women and men. *Circulation* **126**, 1821–1827 (2012).

34. Qi, Q. *et al.* Fried food consumption, genetic risk, and body mass index: Gene-diet interaction analysis in three US cohort studies. *BMJ (Online)* **348**, 1–12 (2014). URL http://dx.doi.org/doi:10.1136/bmj.g1610.

35. Qi, Q. *et al.* Sugar-Sweetened Beverages and Genetic Risk of Obesity From the Departments of Nutrition (Q. *NEJM.org. N Engl J Med* **15**, 1387–96 (2012). URL `https://www.nejm.org/doi/pdf/10.1056/NEJMoa1203039`.

36. Kaplan, R. C. *et al.* A genome-wide association study identifies novel loci associated with circulating IGF-I and IGFBP-3. *Human Molecular Genetics* **20**, 1241–1251 (2011).

37. Richardson, A. S. *et al.* Moderate to vigorous physical activity interactions with genetic variants and body mass index in a large US ethnically diverse cohort. *Pediatric Obesity* **9**, e35–e46 (2014).

38. Wang, H. *et al.* Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the uk biobank. *Science Advances* **5** (2019).

39. Bhattacharyya, D. & Glick, B. S. Two Mammalian Sec16 Homologues Have Nonredundant Functions in Endoplasmic Reticulum (ER) Export and Transitional ER Organization. *Molecular biology of the cell* **18**, 986–994 (2007).

40. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature genetics* **41**, 18–24 (2009).

41. Hotta, K. *et al.* Association between obesity and polymorphisms in SEC16B, TMEM18, GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population. *Journal of Human Genetics* **54**, 727–731 (2009). URL `http://dx.doi.org/10.1038/jhg.2009.106`.

42. Schmid, P. M. *et al.* Expression of fourteen novel obesity-related genes in zucker diabetic fatty rats. *Cardiovascular Diabetology* **11**, 1–11 (2012).

43. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).

44. Greene, W. H. *Econometric Analysis 5th edition* (Pearson Education India, 2003).

45. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

46. Almli, L. M. *et al.* Correcting systematic inflation in genetic association tests that consider interaction effects application to a genome-wide association study of posttraumatic stress disorder. *JAMA Psychiatry* **71**, 1392–1399 (2014).

47. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015). URL `http://www.nature.com/doifinder/10.1038/ng.3404. 15334406`.

48. 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). URL `http://dx.doi.org/10.1038/ng.3865`.

49. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* **49**, 1421–1427 (2017). URL `http://dx.doi.org/10.1038/ng.3954`.

50. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics* **51**, 277–284 (2019). URL `http://dx.doi.org/10.1038/s41588-018-0279-5`.

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

52. Erbe, M. *et al.* Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science* **95**, 4114–4129 (2012).

53. Murphy, K. P. & Bach, F. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).

54. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research* **14**, 1303–1347 (2013).

55. Varadhan, R. & Roland, C. Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics* **35**, 335–353 (2008).

56. Wilcox, R. R. *Introduction to robust estimation and hypothesis testing* (Academic press, 2011).

57. Huber, P. J. *et al.* The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, 221–233 (University of California Press, 1967).

58. White, H. *et al.* A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica* **48**, 817–838 (1980).

59. Voorman, A., Lumley, T., McKnight, B. & Rice, K. Behavior of QQ-plots and Genomic Control in studies of gene-environment interaction. *PLoS ONE* **6** (2011).

60. Long, J. S. & Ervin, L. H. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician* **54**, 217–224 (2000).

61. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics* **5**, 1780–1815 (2011). 1110.6019.

62. Haseman, J. & Elston, R. The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics* **2**, 3–19 (1972).

63. Tobin, M. D., Sheehan, N. A., Scurrah, K. J. & Burton, P. R. Adjusting for treatment effects in studies of quantitative traits: Antihypertensive therapy and systolic blood pressure. *Statistics in Medicine* **24**, 2911–2935 (2005).

64. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer-Verlag New York, 2006), 1 edn.

65. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **112**, 859–877 (2017). URL https://doi.org/10.1080/01621459.2017.1285773. 1601.00670.

66. Band, G. & Marchini, J. Bgen: a binary file format for imputed genotype and haplotype data. *BioRxiv* 308296 (2018).

67. Fernando, R. L., Dekkers, J. C. & Garrick, D. J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics Selection Evolution* **46**, 1–13 (2014).

68. Wang, H. *et al.* Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *bioRxiv* 1–24 (2019).

## Supplementary Material

**Derivation of Variational Bayes updates** We use Coordinate Ascent Variational Inference to optimize the ELBO $\mathcal{F}(\nu;\phi)$ [64].

Using the fact that the variational distributions factorises, we can write the ELBO as

$$\begin{aligned}
\mathcal{F}(\nu;\phi) &= \mathbb{E}_q\left[\log p(y,\theta|\mathcal{D},\phi) - \log q(\theta)\right], \\
&= \mathbb{E}_q\left[\log p(y|\mathcal{D},\phi) + \log p(\theta|y,\mathcal{D},\phi) - \log q(\theta)\right], \\
&= \log p(y|\mathcal{D},\phi) + \mathbb{E}_q\left[\log p(\theta|y,\mathcal{D},\phi) - \log q(\theta)\right], \\
&= \log p(y|\mathcal{D},\phi) + \mathbb{E}_q\left[\log p(\theta|y,\mathcal{D},\phi)\right] - \sum_j \mathbb{E}_q\left[\log q(\theta_j)\right].
\end{aligned}$$

Hence we can consider $\mathcal{F}$ as a function of $q(\theta_j)$

$$\begin{aligned}
\mathcal{F}_j &\propto \mathbb{E}_q\left[\log p(\theta_j|y,\theta_{-j},\mathcal{D},\phi)\right] - \mathbb{E}_q\left[\log q(\theta_j)\right], \\
&\propto \int \prod_i q(\theta_i)\left\{\log p(\theta_j|y,\theta_{-j},\mathcal{D},\phi)\right\} d\theta - \int q(\theta_j)\log q(\theta_j) d\theta_j, \\
&\propto \int q(\theta_j)\left\{\int \prod_{i\neq j} q(\theta_i)\log p(\theta_j|y,\theta_{-j},\mathcal{D},\phi) d\theta_{-j}\right\} d\theta_j - \int q(\theta_j)\log q(\theta_j) d\theta_j, \\
&\propto \int q(\theta_j)\left\{\mathbb{E}_{-j}\left[\log p(\theta_j|y,\theta_{-j},\mathcal{D},\phi)\right]\right\} d\theta_j - \int q(\theta_j)\log q(\theta_j) d\theta_j.
\end{aligned}$$

If we recognize that the last line is proportional to the KL divergence between $\log q(\theta_j)$ and $\mathbb{E}_{-j}\left[\log p(\theta|y,\phi)\right]$ then it is clear that to maximise $\mathcal{F}_j$ we minimise the KL divergence, which occurs at

$$\begin{aligned}
\log q^*(\theta_j) &\propto \mathbb{E}_{-j}\left[\log p(\theta_j|y,\theta_{-j},\mathcal{D},\phi)\right], \\
&\propto \mathbb{E}_{-j}\left[\log p(\theta_j,y|\theta_{-j},\mathcal{D},\phi)\right]
\end{aligned}$$

**Updates for $q(\beta_j)$**

We perform a CAVI update for $q(\beta_j)$.

$$\log q^*(\beta_j) \propto \mathbb{E}_{-\beta_j}\left[\log p(y|\beta_j, \theta_{-\beta_j}, \mathcal{D}, \phi) + \log p(\beta_j|\phi)\right]$$

The prior on $\beta_j$

$$p(\beta_j|\phi) = \frac{\lambda_\beta}{\sqrt{2\pi\sigma_e^2\sigma_{\beta,1}^2}}\exp\left(-\frac{\beta_j^2}{2\sigma_e^2\sigma_{\beta,1}^2}\right) + \frac{1-\lambda_\beta}{\sqrt{2\pi\sigma_e^2\sigma_{\beta,2}^2}}\exp\left(-\frac{\beta_j^2}{2\sigma_e^2\sigma_{\beta,2}^2}\right),$$

and the component of the expected conditional log-likelihood that depends on $\beta_j$

$$\begin{aligned}\mathbb{E}_{-\beta_j}\left[\log p(y|\beta_j, \theta_{-\beta_j}, \mathcal{D}, \phi)\right] &\propto -\frac{\beta_j^2}{2\sigma_e^2}||X_j||_2^2 + \frac{\beta_j}{\sigma_e^2}X_j^T\left(y - \mathbb{E}_{-\beta_j}\left[C\alpha + X_{-j}\beta_{-j} + \eta \odot X\gamma\right]\right), \\ &= -\frac{\beta_j^2}{2\sigma_e^2}||X_j||_2^2 + \frac{\beta_j}{\sigma_e^2}X_j^T\mathbb{E}_{-\beta_j}\left[y_{\text{resid},-\beta_j}\right]\end{aligned}$$

where $y_{\text{resid},-\beta_j} = y - C\alpha - X_{-j}\beta_{-j} - \eta \odot X\gamma$. Thus $q^*(\beta_j)$ can be written as

$$q^*(\beta_j) \propto p(\beta|\phi)\exp\left(-\frac{\beta_j^2}{2\sigma_e^2}||X_j||_2^2 + \frac{\beta_j}{\sigma_e^2}X_j^T y_{\text{resid},-\beta_j}\right),$$

and by expanding out the terms we can see that $q^*(\beta_j)$ is also a mixture of gaussians

$$\begin{aligned}q(\beta_j) \propto &\frac{\lambda_\beta}{\sqrt{2\pi\sigma_e^2\sigma_{\beta,1}^2}}\exp\left(-\frac{\beta_j^2}{2\sigma_e^2}\left(||X_j||_2^2 + \frac{1}{\sigma_{\beta,1}^2}\right) + \frac{\beta_j}{\sigma_e^2}X_j^T\mathbb{E}_{-\beta_j}\left[y_{\text{resid},-\beta_j}\right]\right) + \\ &\frac{(1-\lambda_\beta)}{\sqrt{2\pi\sigma_e^2\sigma_{\beta,2}^2}}\exp\left(-\frac{\beta_j^2}{2\sigma_e^2}\left(||X_j||_2^2 + \frac{1}{\sigma_{\beta,2}^2}\right) + \frac{\beta_j}{\sigma_e^2}X_j^T\mathbb{E}_{-\beta_j}\left[y_{\text{resid},-\beta_j}\right]\right)\end{aligned}$$

with parameters

$$\begin{aligned}s_{j,i}^\beta &= \frac{\sigma_e^2}{||X_j||_2^2 + 1/\sigma_{\beta,i}^2}, && \text{for } i = 1, 2 \\ \mu_{j,i}^\beta &= \frac{s_{j,i}^\beta}{\sigma_e^2}X_j^T\mathbb{E}_{-\beta_j}\left[y_{\text{resid},-\beta_j}\right], && \text{for } i = 1, 2 \\ \psi_j^\beta &= \text{sigmoid}\left(\text{logit}\left(\lambda_\beta\right) - \frac{1}{2}\log\left(\frac{\sigma_{\beta,1}^2 s_{j,2}^\beta}{s_{j,1}^\beta\sigma_{\beta,2}^2}\right) + \frac{(\mu_{j,1}^\beta)^2}{2s_{j,1}^\beta} - \frac{(\mu_{j,2}^\beta)^2}{2s_{j,2}^\beta}\right)\end{aligned}$$

**Updates for $q(\gamma_j)$**

We perform a CAVI update for $q(\gamma_j)$.

$$\log q^*(\gamma_j) \propto \mathbb{E}_{-\gamma_j}\left[\log p(y|\gamma_j, \theta_{-\gamma_j}, \mathcal{D}, \phi) + \log p(\gamma_j|\phi)\right]$$

The prior on $\gamma_j$ is

$$p(\gamma_j|\phi) = \frac{\lambda_\gamma}{\sqrt{2\pi\sigma_e^2\sigma_{\gamma,1}^2}}\exp\left(-\frac{\gamma_j^2}{2\sigma_e^2\sigma_{\gamma,1}^2}\right) + \frac{1-\lambda_\gamma}{\sqrt{2\pi\sigma_e^2\sigma_{\gamma,2}^2}}\exp\left(-\frac{\gamma_j^2}{2\sigma_e^2\sigma_{\gamma,2}^2}\right) \tag{2}$$

To obtain the component of the expected conditional log-likelihood that depends on $\gamma_j$, we first note that

$$\begin{aligned}\log p(y|\gamma_j, \theta_{-\gamma_j}, \mathcal{D}, \phi) &\propto -\frac{1}{2\sigma_e^2}||y - C\alpha - X\beta - Z\gamma||_2^2, \\ &\propto -\frac{1}{2\sigma_e^2}\left(-2\gamma_j z_j^T(y - C\alpha - X\beta - Z_{-j}\gamma_{-j}) + \gamma_j^2||Z_j^T||_2^2\right).\end{aligned}$$

Hence the expected conditional log-likelihood is

$$\begin{aligned}\mathbb{E}_{-\gamma_j}\left[\log p(y|\gamma_j, \theta_{-\gamma_j}, \mathcal{D}, \phi)\right] &\propto \frac{\gamma_j^2}{\sigma_e^2}\mathbb{E}_{-\gamma_j}\left[||Z_j||_2^2\right] - \\ &\quad \frac{1}{2\sigma_e^2}\gamma_j X_j^T\mathbb{E}_{-\gamma_j}\left[\eta \odot (y - C\alpha X - \beta) - \eta^2 \odot X_{-j}\gamma_{-j}\right], \\ &\propto \frac{\gamma_j^2}{\sigma_e^2}\mathbb{E}_{-\gamma_j}\left[||Z_j||_2^2\right] - \frac{1}{2\sigma_e^2}\gamma_j X_j^T\mathbb{E}_{-\gamma_j}\left[\eta \odot y_{\text{resid},-\gamma_j}\right],\end{aligned} \tag{3}$$

where $y_{\text{resid},-\gamma_j} = y - C\alpha - X\beta - \eta \odot X_{-j}\gamma_{-j}$.

Finally if we substitute (2) and (3) into

$$q^*(\gamma_j) \propto p(\gamma|\phi) \exp\left(-\frac{\gamma_j^2}{2\sigma_e^2}\mathbb{E}_{-\gamma_j}\left[||Z_j||_2^2\right] + \frac{\gamma_j}{\sigma_e^2}X_j^T\mathbb{E}_{-\gamma_j}\left[\eta \odot y_{\text{resid},-\gamma_j}\right]\right).$$

then we can see that $q^*(\gamma_j)$ is a mixture of two gaussians with parameters

$$s_{j,i}^{\gamma} = \frac{\sigma_e^2}{\mathbb{E}_{-\gamma_j}\left[||Z_j||_2^2\right] + 1/\sigma_{\gamma,i}^2}, \qquad \text{for } i = 1, 2$$

$$\mu_{j,i}^{\gamma} = \frac{s_{j,i}^{\gamma}}{\sigma_e^2}X_j^T\mathbb{E}_{-\gamma_j}\left[\eta \odot y_{\text{resid},-\gamma_j}\right], \qquad \text{for } i = 1, 2$$

$$\psi_j^{\gamma} = \text{sigmoid}\left(\text{logit}\left(\lambda_\gamma\right) - \frac{1}{2}\log\left(\frac{\sigma_{\gamma,1}^2 s_{j,2}^{\gamma}}{s_{j,1}^{\gamma}\sigma_{\gamma,2}^2}\right) + \frac{(\mu_{j,1}^{\gamma})^2}{2s_{j,1}^{\gamma}} - \frac{(\mu_{j,2}^{\gamma})^2}{2s_{j,2}^{\gamma}}\right)$$

Note that

$$\begin{aligned}
\mathbb{E}_{-\gamma_j}\left[||Z_j||_2^2\right] &= X_{\cdot j}^T \text{diag}\left(\mathbb{E}_q\left[\eta^2\right]\right) X_{\cdot j}, \\
&= \sum_{l,m}\mathbb{E}_q\left[w_m\right]\mathbb{E}_q\left[w_l\right]\underbrace{\left(X^T\text{diag}\left(E_l \odot E_{\cdot m}\right)X\right)_{jj}}_{\text{precomputed}} \\
&\quad + \sum_l \text{Var}_q(w_l)\underbrace{\left(X^T\text{diag}\left(E_l^2\right)X\right)_{jj}}_{\text{precomputed}}
\end{aligned}$$

is an $O(L^2)$ operation.

**Updates for** $q(w_l)$

We first rewrite the conditional log-likelihood to make its dependence on $w$ clear.

$$\mathbb{E}_q\left[\log p(y|w,\theta_{-w},\mathcal{D},\phi)\right] \propto -\frac{1}{2\sigma_e^2}\mathbb{E}_q\left[||y - C\alpha - X\beta - \eta \odot X\gamma||_2^2\right],$$
$$\propto -\frac{1}{2\sigma_e^2}\mathbb{E}_q\left[||A - Bw||_2^2\right]$$

where $A = y - C\alpha - X\beta$ and $B = \text{diag}\left(X\gamma\right)E$. For convenience we denote the $l$'th column of $B$ as $B_l$. Therefore extracting the component of the conditional log-likelihood that depends on $w_l$ yields

$$\mathbb{E}_q\left[\log p(y|w_l,\theta_{-w_l},\mathcal{D},\phi)\right] \propto -\frac{1}{2\sigma_e^2}\left(w_l^2\mathbb{E}_{-w_l}\left[||B_l||_2^2\right] - 2w_l\mathbb{E}_{-w_l}\left[B_l^T(a - B_{-l}w_{-l})\right]\right).$$

The prior on $w_l$ is a standard gaussian, so $q^*(w_l)$ is also gaussian with parameters

$$s_l^w = \frac{\sigma_e^2}{\sigma_e^2 + \mathbb{E}_{-w_l}\left[||B_l||_2^2\right]},$$
$$= \frac{\sigma_e^2}{\sigma_e^2 + y_X^T\text{diag}\left((E_{.l}^*)^2\right)y_X + \sum_{k=1}^P \text{Var}\left(\gamma_k\right)\left(X^T\text{diag}\left((E_{.l}^*)^2\right)X\right)_{kk}},$$

$$\mu_l^w = \frac{s_l^w}{\sigma_e^2}\mathbb{E}_{-w_l}\left[B_l^T(a - B_{-l}w_{-l})\right],$$
$$= \frac{s_l^w}{\sigma_e^2}\left((y - \hat{y}_M)^T\text{diag}\left(E_l^*\right)\hat{y}_X - \hat{y}_X^T\text{diag}\left(E_l^* \odot \mathbb{E}\left[\eta_{-l}\right]\right)\hat{y}_X\right)$$
$$- \frac{s_l^w}{\sigma_e^2}\left(\sum_k \text{Var}\left(\gamma_k\right)\left(X^T\text{diag}\left(\mathbb{E}\left[\eta_{-l}\right] \odot E_l^*\right)X\right)_{kk}\right).$$

We now evaluate the quantities $\mathbb{E}_{-w_l}\left[||B_l||_2^2\right]$ and $\mathbb{E}_{-w_l}\left[B_l^T(a - B_{-l}w_{-l})\right]$.

$$\begin{aligned}
\mathbb{E}_{-w_l}\left[||B_l||_2^2\right] &= \mathbb{E}\left[\sum_i E_{il}^2\left(\sum_j X_{ij}^2\gamma_j\right)^2\right], \\
&= \sum_i E_{il}^2\left(\sum_j X_{ij}^2\mathbb{E}\left[\gamma_j\right]\right)^2 + \sum_i E_{il}^2\sum_j X_{ij}^2\mathrm{Var}\left(\gamma_j\right), \\
&= y_X^T\mathrm{diag}\left(E_l^2\right)y_X + \sum_j \mathrm{Var}\left(\gamma_j\right)\sum_i E_{il}^2X_{ij}^2.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{-w_l}\left[B_l^T(a - B_{-l}w_{-l})\right] &= (y - \hat{y}_M)^T\mathrm{diag}\left(E_l^*\right)\hat{y}_X - E_l^*\mathrm{diag}\left(\hat{y}_X^2\right)\mathbb{E}\left[\eta_{-l}\right] \\
&\quad - \sum_k \mathrm{Var}\left(\gamma_k\right)\sum_{m\neq l}\mathbb{E}\left[w_m\right]\sum_i X_{ij}^2E_{il}E_{im}.
\end{aligned}$$

Recomputing quantities $\sum_i E_{il}^2X_{ij}^2$ and $\sum_i X_{ij}^2E_{il}E_{im}$ for $1 \leq j \leq M$ and $1 \leq l \leq m \leq L$ every iteration would impose prohibitive computational costs (and cause our algorithm to scale as $\mathcal{O}(NPL)$. Instead we precompute a $M \times L^2$ matrix where entry $(j, (m \times L + l)) = \sum_i X_{ij}^2E_{il}E_{im}$. This is easily parallelised over variants and/or environments, so for biobank scale datasets we recommend that users precompute this quantity beforehand and provide to LEMMA at runtime in a text file.

**Updates for** $q(\alpha_c)$

$$\begin{aligned}
s_c^\alpha &= \frac{\sigma_e^2}{1/\sigma_\alpha^2 + (N-1)}, \\
\mu_c^\alpha &= \frac{s_c^\alpha}{\sigma_e^2}E_c^T\mathbb{E}_{-\alpha_c}\left[y_{\text{resid},-\alpha_c}\right]
\end{aligned}$$

where $y_{\text{resid},-\alpha_c} = y - E_{-c}\alpha_{-c} - X\beta - \eta \odot X\gamma$.

**Evidence lower bound** Variational inference involves maximising a lower bound $\mathcal{F}(\phi;\nu)$ on the model log-likelihood $\log p(y|\mathcal{D},\phi)$. This is given by

$$\mathcal{F}(\phi;\nu) = \mathbb{E}_q\left[\log p(y|\theta,\mathcal{D},\phi)\right] - \mathbf{KL}(q(\theta;\nu)||p(\theta|\phi),$$

$$\begin{aligned}
\mathcal{F}(\phi;\nu) = &-\frac{N}{2}\log(2\pi\sigma_e^2) \\
&-\frac{1}{2\sigma_e^2}\left(||y - C\mathbb{E}_q\left[\alpha\right] - X\mathbb{E}_q\left[\beta\right] - \mathbb{E}_q\left[\eta\right]\odot X\mathbb{E}_q\left[\gamma\right]||_2^2\right) \\
&-\frac{1}{2\sigma_e^2}\left(\mathbb{E}_q\left[\gamma\right]^T X^T \mathbf{diag}\left(\mathbb{E}_q\left[\eta^2\right]\right) X\mathbb{E}_q\left[\gamma\right] - ||\mathbb{E}_q\left[\eta\right]\odot X\mathbb{E}_q\left[\gamma\right]||_2^2\right) \\
&-\frac{N-1}{2\sigma_e^2}\sum_l \mathrm{Var}_q\left(\alpha_l\right) \\
&-\frac{N-1}{2\sigma_e^2}\sum_k \mathrm{Var}_q\left(\beta_k\right) \\
&-\frac{1}{2\sigma_e^2}\sum_k \mathrm{Var}_q\left(\gamma_k\right)\left(X^T\mathbf{diag}\left(\mathbb{E}_q\left[\eta^2\right]\right)X\right)_{kk} \\
&+\frac{P_c}{2}(1-\log(\sigma_e^2\sigma_\alpha^2)) + \frac{1}{2}\sum_m^{P_c}\left(\log(s_m^C) - \frac{s_m^C + (\mu_m^C)^2}{\sigma_e^2\sigma_\alpha^2}\right) \\
&-\sum_j\left[\psi_j^\beta\log\frac{\psi_j^\beta}{\lambda_\beta} + (1-\psi_j^\beta)\log\frac{1-\psi_j^\beta}{1-\lambda_\beta}\right] \\
&+\frac{1}{2}\sum_j\left[1 + \psi_j^\beta\left(\log\frac{s_{j,1}^\beta}{\sigma_e^2\sigma_{\beta,1}^2} - \frac{s_{j,1}^\beta + (\mu_{j,1}^\beta)^2}{\sigma_e^2\sigma_{\beta,1}^2}\right) + (1-\psi_j^\beta)\left(\log\frac{s_{j,2}^\beta}{\sigma_e^2\sigma_{\beta,2}^2} - \frac{s_{j,2}^\beta + (\mu_{j,2}^\beta)^2}{\sigma_e^2\sigma_{\beta,2}^2}\right)\right] \\
&-\sum_j\left[\psi_j^\gamma\log\frac{\psi_j^\gamma}{\lambda_\gamma^\gamma} + (1-\psi_j^\gamma)\log\frac{1-\psi_j^\gamma}{1-\lambda_\gamma}\right] \\
&+\frac{1}{2}\sum_j\left[1 + \psi_j^\gamma\left(\log\frac{s_{j,1}^\gamma}{\sigma_e^2\sigma_{\gamma,1}^2} - \frac{s_{j,1}^\gamma + (\mu_{j,1}^\gamma)^2}{\sigma_e^2\sigma_{\gamma,1}^2}\right) + (1-\psi_j^\gamma)\left(\log\frac{s_{j,2}^\gamma}{\sigma_e^2\sigma_{\gamma,2}^2} - \frac{s_{j,2}^\gamma + (\mu_{j,2}^\gamma)^2}{\sigma_e^2\sigma_{\gamma,2}^2}\right)\right] \\
&+\frac{P_e}{2} + \frac{1}{2}\sum_l\left(\log(s_l^w) - (s_l^w + (\mu_l^w)^2)\right)
\end{aligned}$$

**Derivation of hyperparameter maximisation** For the maximisation step we set $\phi = \hat{\phi}$ where $\nabla_\phi F(\phi; \nu) = 0$. For ease of notation we perform the following change of variables

$$\tilde{\sigma}^2_{\beta,1} = \sigma^2_e \sigma^2_{\beta,1} \quad \rightarrow \frac{\partial}{\partial \tilde{\sigma}^2_{\beta,1}} = \frac{1}{\sigma^2_e} \frac{\partial}{\partial \sigma^2_{\beta,1}},$$
$$\tilde{\sigma}^2_{\beta,2} = \sigma^2_e \sigma^2_{\beta,2} \quad \rightarrow \frac{\partial}{\partial \tilde{\sigma}^2_{\beta,2}} = \frac{1}{\sigma^2_e} \frac{\partial}{\partial \sigma^2_{\beta,2}},$$

$$\tilde{\sigma}^2_{\gamma,1} = \sigma^2_e \sigma^2_{\gamma,1} \quad \rightarrow \frac{\partial}{\partial \tilde{\sigma}^2_{\gamma,1}} = \frac{1}{\sigma^2_e} \frac{\partial}{\partial \sigma^2_{\gamma,1}},$$
$$\tilde{\sigma}^2_{\gamma,2} = \sigma^2_e \sigma^2_{\gamma,2} \quad \rightarrow \frac{\partial}{\partial \tilde{\sigma}^2_{\gamma,2}} = \frac{1}{\sigma^2_e} \frac{\partial}{\partial \sigma^2_{\gamma,2}}.$$

This makes the derivation easier as all the partial derivatives become decoupled. Partial derivatives with respect to each hyperparameter are given by

$$\frac{\partial F}{\partial \lambda_\beta} = \sum_j \left( \frac{\psi_j^\beta}{\lambda_\beta} - \frac{(1-\psi_j^\beta)}{1-\lambda_\beta} \right),$$
$$\frac{\partial F}{\partial \lambda_\gamma} = \sum_j \left( \frac{\psi_j^\gamma}{\lambda_\gamma} - \frac{(1-\psi_j^\gamma)}{1-\lambda_\gamma} \right),$$

$$\frac{\partial F}{\partial \tilde{\sigma}^2_{\beta,1}} = \sum_j \frac{\psi_j^\beta}{2} \left( -\frac{1}{\tilde{\sigma}^2_{\beta,1}} + \frac{s^\beta_{j,1} + (\mu^\beta_{j,1})^2}{(\tilde{\sigma}^2_{\beta,1})^2} \right),$$
$$\frac{\partial F}{\partial \tilde{\sigma}^2_{\beta,2}} = \sum_j \frac{1-\psi_j^\beta}{2} \left( -\frac{1}{\tilde{\sigma}^2_{\beta,2}} + \frac{s^\beta_{j,2} + (\mu^\beta_{j,2})^2}{(\tilde{\sigma}^2_{\beta,2})^2} \right),$$

$$\frac{\partial F}{\partial \tilde{\sigma}^2_{\gamma,1}} = \sum_j \frac{\psi_j^\gamma}{2} \left( -\frac{1}{\tilde{\sigma}^2_{\gamma,1}} + \frac{s^\gamma_{j,1} + (\mu^\gamma_{j,1})^2}{(\tilde{\sigma}^2_{\gamma,1})^2} \right),$$
$$\frac{\partial F}{\partial \tilde{\sigma}^2_{\gamma,2}} = \sum_j \frac{1-\psi_j^\gamma}{2} \left( -\frac{1}{\tilde{\sigma}^2_{\gamma,2}} + \frac{s^\gamma_{j,2} + (\mu^\gamma_{j,2})^2}{(\tilde{\sigma}^2_{\gamma,2})^2} \right),$$

$$\frac{\partial F}{\partial \sigma_e^2} = -\frac{N}{2\sigma_e^2} + \frac{1}{2(\sigma_e^2)^2} \mathbb{E}_q \left[ ||y - C\alpha - X\beta - \text{diag}(\eta) X\gamma||_2^2 \right] - \frac{M}{2\sigma_e^2} + \frac{1}{2(\sigma_e^2)^2 \sigma_C^2} \sum_m \left( s_m^C + (\mu_m^C)^2 \right)$$

Hence the maximization steps are

$$\hat{\lambda}_\beta = \frac{1}{P}\sum_j \psi_j^\beta, \tag{4}$$

$$\hat{\lambda}_\gamma = \frac{1}{P}\sum_j \psi_j^\gamma, \tag{5}$$

$$\tag{6}$$

$$\hat{\sigma}^2_{\beta,1} = \frac{\sum_j \psi_j^\beta (s^\beta_{j,1} + (\mu^\beta_{j,1})^2)}{\hat{\sigma}_e^2 \sum_j \psi_j^\beta}, \tag{7}$$

$$\hat{\sigma}^2_{\beta,2} = \frac{\sum_j (1-\psi_j^\beta)(s^\beta_{j,2} + (\mu^\beta_{j,2})^2)}{\hat{\sigma}_e^2 \sum_j (1-\psi_j^\beta)}, \tag{8}$$

$$\tag{9}$$

$$\hat{\sigma}^2_{\gamma,1} = \frac{\sum_j \psi_j^\gamma (s^\gamma_{j,1} + (\mu^\gamma_{j,1})^2)}{\hat{\sigma}_e^2 \sum_j \psi_j^\gamma}, \tag{10}$$

$$\hat{\sigma}^2_{\gamma,2} = \frac{\sum_j (1-\psi_j^\gamma)(s^\gamma_{j,2} + (\mu^\gamma_{j,2})^2)}{\hat{\sigma}_e^2 \sum_j (1-\psi_j^\gamma)}, \tag{11}$$

$$\tag{12}$$

$$\hat{\sigma}^2 = \frac{\mathbb{E}_q\left[||y - C\alpha - X\beta - \text{diag}\left(\eta\right) X\gamma||_2^2\right] + \frac{1}{\sigma_C^2}\sum_m \left(s_m^C + (\mu_m^C)^2\right)}{N+M}. \tag{13}$$

As an aside we note that one could use the maximised hyperparameters (after converence) to obtain a point estimate of $\text{Var}\left(\beta\right)$. However, by substituting in Equations (4) to (11) we can see that this

is equivalent to $\sum_j \mathbb{E}_q\left[\beta_j^2\right]/M$.

$$\begin{aligned}
\mathrm{Var}\left(\beta\right) &= \lambda_\beta \sigma_{\beta,1}^2 + (1-\lambda_\beta)\sigma_{\beta,2}^2, \\
&\approx \hat{\lambda}_\beta \hat{\sigma}_{\beta,1}^2 + (1-\hat{\lambda}_\beta)\hat{\sigma}_{\beta,2}^2, \\
&= \frac{1}{M}\sum_j \left(\psi_j^\beta (s_{j,1}^\beta + (\mu_{j,1}^\beta)^2) + (1-\psi_j^\beta)(s_{j,2}^\beta + (\mu_{j,2}^\beta)^2)\right), \\
&= \frac{1}{M}\sum_j \mathbb{E}_q\left[\beta_j^2\right].
\end{aligned}$$

As the mean field assumption tends to cause variational inference algorithms to underestimate the variance of latent variables [65], this is likely to produce an underestimate of $\mathrm{Var}\left(\beta\right)$. We can observe the same result for $\mathrm{Var}\left(\gamma\right)$ with an analogous argument.

**Compressed genotype data** To reduce RAM usage, LEMMA stores a compressed version of the genotype matrix using $NM$ bytes. To do this LEMMA splits the interval $[0, 2]$ into $2^8$ segments and stores the index of the segment that each dosage falls into, as well as the mean and variance for each SNP. Then when operating on a SNP, LEMMA reconstructs the centered and scaled dosages for that SNP. This approach results in a small loss of accuracy, but is more flexible that assuming dosages are in $\{0, 1, 2\}$ as it allows LEMMA to run on imputed SNPs [66].

**Computational efficiency** Using mean field variational inference, estimation of the posterior means of the latent variables $\beta, \gamma, w$ can be reduced to an iterative algorithm that cycles through the variables sequentially, updating each conditional on the values of the others. Taking the main

effect of the $j$'th SNP as an example, the update scheme for $\beta_j$ can be written heuristically as

$$\tilde{\beta}_j = X_j^T y_{\text{resid}}, \tag{14}$$

$$\hat{\beta}_j^t = \text{regularise}(\tilde{\beta}_j; \phi^t), \tag{15}$$

$$y_{\text{resid}} = y_{\text{resid}} - (\hat{\beta}_j^t - \hat{\beta}_j^{t-1}) X_j^T. \tag{16}$$

In Equation (14) we compute the correlation between the $j$ SNP and the residual phenotype vector. In Equation (15) we compute the posterior mean of $\beta_j$ which depends on the correlation with the residual phenotype, the prior on $\beta_j$ and the current hyperparameters. Finally in Equation (16) we update the residual phenotype vector.

The majority of computational time is spent on the dot product in Equation (14) and updating the residual phenotype in Equation (16). Both are BLAS Level 1 operations, which implies that memory access is often the principle bottleneck rather than the number of cores available. It is possible to step up to BLAS Level 2 by updating a block of SNPs in parallel [13], however this is still a memory bound operation. Instead we use a parallel computing strategy suggested by [67] for use in genome wide regression, and subsequently used by [31], to compute the dot product and perform the residual update in parallel using OpenMPI. Briefly, we partition the samples such that blocks of rows of the phenotype $y$, genotypes $X$ and environmental variables $E$ are assigned to each core. For a given update step, each core calculates the dot product for the locally held block of samples and then shares the local dot product with the rest of the network. From this the dot product for the entire cohort can be reconstructed cheaply. After computing the poesterior mean, each core then updates the residual phenotype for the block of samples stored locally. We observed that

a distributed algorithm using OpenMPI was faster than the same algorithm using multithreaded matrix-vector operations with the Intel MKL Library even on a single node with multiple cores. However using OpenMPI has the additional advantage of allowing users to utilise cores from across a cluster rather than being restricted to a single node. **Figure S15** shows LEMMA scales with increasing sample size.

**Precomputed quantities** To aid computational efficiency we precompute a $P \times L(L+1)$ matrix $W$ where

$$W_{:,m\times L+l} = \text{diag}((E_l^* \odot X)^T (E_m^* \odot X)), \quad \text{for } 1 \geq l < m \geq L,$$

and is used in the update steps of $q(w_l)$. LEMMA can compute this internally, incurring a one off cost of $\mathcal{O}(NPL^2)$, or is able to read from a text file at runtime. As this is easily computed in parallel over batches of variants and/or environment, we recommend that for biobank scale datasets users should precompute this quantity beforehand using a separate tool that we have provided and provide to LEMMA at runtime.

**Parameter Initialisation** We start the variational mean estimates of $q(\beta)$ and $q(\gamma)$ at zero. To initialise mean estimates of the interaction weights $q(w)$ we have two options; the first of which is simply to use a uniform weighting over all environments. For the second we apply an F-Test independently at each SNP and use the learned coefficients from the test with the lowest p-value as the initial values of the interaction weights. We find that we often obtain similar results from both options, so for simplicity we use a uniform start point for our Biobank analyses. To initialise mean estimates of $q(\alpha)$ we use the least squares fit of $C$ on $y$.

To initialise the hyperparameters we draw a random samples

$$h^2_\beta \sim \mathcal{U}(0, 0.5),$$

$$h^2_\gamma \sim \mathcal{U}(0, 0.1),$$

$$-\log_{10}(\lambda_\beta) \sim \mathcal{U}([2, ..., 1 - \log_{10}(M)])$$

$$-\log_{10}(\lambda_\gamma) \sim \mathcal{U}([2, ..., 1 - \log_{10}(M)]).$$

We set $\lambda_\beta, \lambda_\gamma$ in this manner to reflect a belief that somewhere between ten and one in one-hundred SNPs should be part of the slab prior. We then set

$$\sigma^2_e = 1 - h^2_\beta - h^2_\gamma,$$
$$\sigma^2_{\beta,1} = \frac{1}{\lambda_\beta M} \frac{h^2_\beta}{1 - h^2_\beta - h^2_\gamma},$$
$$\sigma^2_{\gamma,1} = \frac{1}{\lambda_\gamma M} \frac{h^2_\gamma}{1 - h^2_\beta - h^2_\gamma},$$

and initialise the spike variances at

$$\sigma^2_{\beta,2} = \sigma^2_{\beta,1}/1000,$$

$$\sigma^2_{\gamma,2} = \sigma^2_{\gamma,1}/1000.$$

**Missing data** Samples with missing data in the phenotype, environmental variables or covariates are excluded. By default LEMMA imputes missing genetic data with the mean dosage of each SNP, however as LEMMA does not assume dosages are hard called with $\{0, 1, 2\}$ we recommend that useres first impute genetic data with standard imputation pipelines.

**Robust standard errors in GxE Studies** In **Figure S16** we demonstrate how a multiplicative GxE interaction effect on a quantitative trait can cause the conditional trait variance given an interacting SNP $\mathrm{Var}(Y|g_0)$ to differ according to the interacting SNPs genotype. This is known as conditional heteroscedasticity and is the key insight behind several recent methods to detect SNPs with non-zero GxE effects in the UK Biobank [8,68].

In the same figure, we can observe that the conditional trait variance given the environmental exposure $\mathrm{Var}(y|E)$ also displays signs of conditional heteroskedacity. Previous studies [46] have observed that methods that assume homoskedasticity can display substantial inflation when testing for GxE effects at SNPs where there is no true GxE effect. In our simulations we observed that inflation of GxE tests statistics from LEMMA-S and the F-test, both of which assune homoskdasticity, increased with SNP-GxE heritability. Below we give an explanation for this phenomenon.

Consider a polygenic quantitative trait $Y$ that has multiplicative GxE interactions with the same environmental exposure $E$ at multiple SNPs

$$y_i = \alpha E_i + \sum_{j=1}^{M} \beta_j G_{ij} + \sum_{j=1}^{M} \gamma_j E_i G_{ij} + \epsilon_i, \tag{17}$$

where $M$ is the number of SNPs and the coefficients represent true effects. For simplicity we assume that $E$ and SNPs $G_j$ are normalised to have mean zero and variance one, that the set of $E$ with all causal SNPs $\{E\} \cup \{G_j : \beta_j \neq 0\}$ is pairwise independent and that the influence from population structure is negligible.

Suppose we have identified $E$ as an environmental variable that may plausible have GxE

interations with our phenotype and we then conduct a GWAS for GxE effects. Then at the $k$'th SNP we wish to test the hypothesis $\gamma_k \neq 0$ in the following linear model

$$\begin{aligned} y &= \alpha E + G_k\beta_k + E \cdot G_k\gamma_k + u, \\ &= X\tau + u, \end{aligned}$$

where in the second line $\tau = (\alpha, \beta_k, \gamma_k)^T$, $X$ is the corresponding design matrix encapsulating all fixed effects and $u$ in an unobserved random effects capturing residual noise. Assuming that $\mathbb{E}[u|X] = 0$, the usual least squares estimate of $\tau$, $\hat{\tau} = (X^TX)^{-1}X^Ty$, has asymptotic distribution

$$\hat{\tau} \to \mathcal{N}(\tau, \mathrm{Var}(\hat{\tau})),$$

where

$$\begin{aligned} \mathrm{Var}(\hat{\tau}) &= \mathbb{E}_X[\mathrm{Var}(\hat{\tau}|X)] + \mathrm{Var}_X(\mathbb{E}[\hat{\tau}|X]), \\ &= \mathbb{E}_X[\mathrm{Var}(\hat{\tau}|X)] + \mathrm{Var}_X(\tau), \\ &= \mathbb{E}_X[\mathrm{Var}(\hat{\tau}|X)], \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(\hat{\tau}|X) &= \mathrm{Var}\left(\tau + (X^TX)^{-1}X^Tu|X\right), \\ &= (X^TX)^{-1}\mathrm{Var}\left(X^Tu|X\right)(X^TX)^{-T}, \\ &= (X^TX)^{-1}X^T\mathrm{Var}(u|X)\,X(X^TX)^{-T}. \end{aligned}$$

The usual approach is to assume that $\mathrm{Var}(u|X) = \sigma^2 I$ (ie homoskedasticity), which yields the standard variance estimator $\mathrm{Var}(\hat{\tau}|X) = \sigma^2(X^TX)^{-1}$. However, given the true generative model

for $y$ given in Equation (17), we can write $u$ as

$$u = \sum_{j \neq k} (G_j \beta_j + E G_j \gamma_j) + \epsilon. \tag{18}$$

Therefore we can evaluate the conditional variance of $u$ given $X$,

$$\begin{aligned}
\mathrm{Var}\left(u|X\right) &= \mathrm{Var}\left(u|E = e, G_k = g_k\right), \\
&= \mathrm{Var}\left(\sum_{j \neq k} (G_j \beta_j + e G_j \gamma_j) + \epsilon\right), \\
&= \sum_{j \neq k} \mathrm{Var}\left(\beta_j G_j\right) + \sum_{j \neq k} \mathrm{Var}\left(\gamma_j e G_j\right) + 2 \sum_{j \neq k} \mathrm{Cov}\left(\beta_j G_j, \gamma_j e G_j\right) + 1 \\
&\quad + \sum_{j \neq k, m \neq k} \mathrm{Cov}\left(\beta_j G_j, \beta_m G_m\right) + \sum_{j \neq k, m \neq k} \mathrm{Cov}\left(\gamma_j e G_j, \gamma_m e G_m\right) \\
&= \sum_{j \neq k} \left(\beta_j + e\gamma_j\right)^2 + 1,
\end{aligned}$$

where the covariances in the second line are all zero due to pairwise independance of the set $\{E\} \cup \{G_j : \beta_j \neq 0\}$. Thus the conditional trait variance will vary depending on the strength of environmental exposure either if there are a few SNPs with GxE interactions of large effect or if there are many SNPs with small yet non-zero interaction effects, and in either case homoskedasticity is unlikely to be an appropriate assumption.

Robust standard errors, alternatively called Huber-White, sandwich or heteroskedastic-consistent errors [57,58], are standard tools used in economics [44] to overcome this issue and have previously been proposed for use in GxE interaction studies [24,46,59]. We further include a small adjustment that reduces bias in small samples [60]. This yields the variance estimator

$$\mathrm{Var}\left(\hat{\tau}\right) = (H^T H)^{-1} H^T \hat{\Sigma} H (H^T H)^{-1},$$

where $\hat{\Sigma}$ is a diagonal matrix with $\hat{\Sigma}_{ii} = \frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}$, where $\hat{\epsilon} = y - H\hat{\tau}$ and $h = H(H^T H)^{-1} H^T$.

## Supplementary Figures

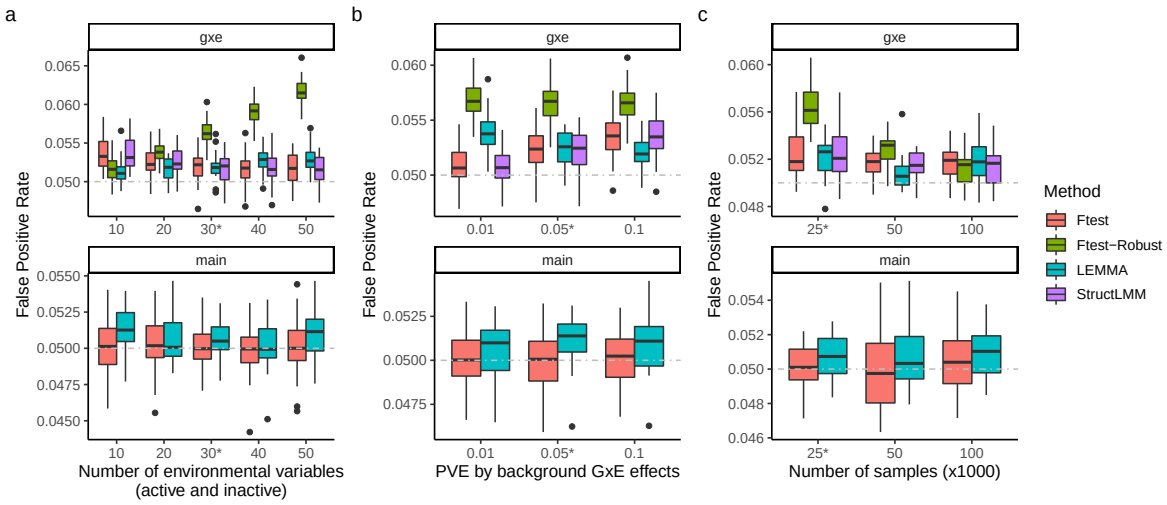



Figure S1: **False positive rate in simulation**. False positive rate (FPR) for SNP main effects tests (bottom) and SNP GxE interaction tests (top) at null SNPs in the second half of each chromosome, whilst varying (a) the number of environmental variables, (b) proportion of trait variance explained by background GxE effects and (c) sample size. The grey line denotes expected FPR. Simulations used genotypes subsampled from the UK Biobank and by default contained $N = 25K$ samples, $M = 100K$ SNPs, $6$ environmental variables that contributed to the ES and $24$ that did not (default parameters denoted by stars). We performed $20$ repeats for each scenario. See **Online methods** for full details of phenotype construction.

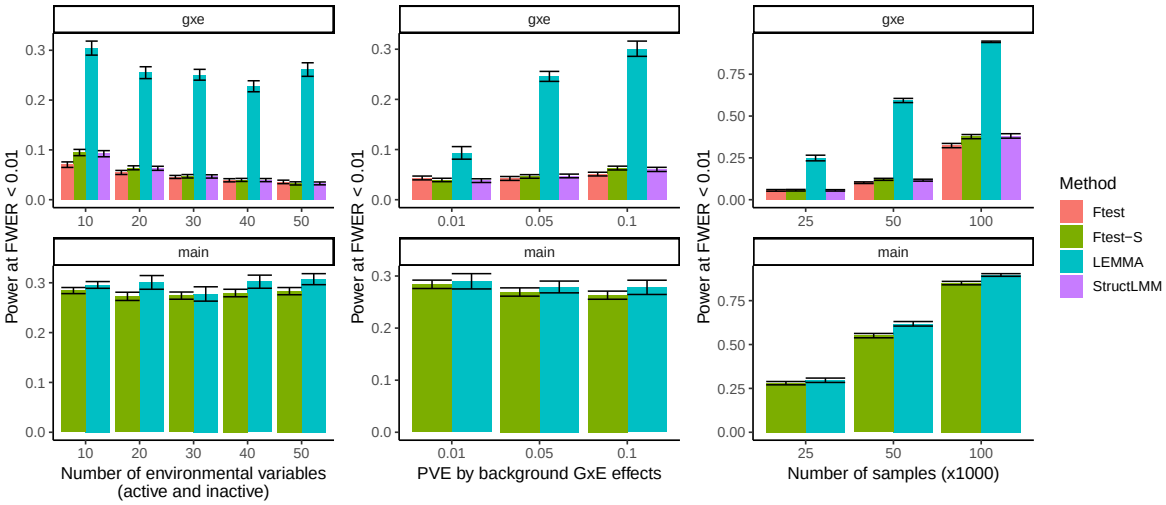



Figure S2: **Power to detect causal SNPs in simulation.** Power to detect SNP GxE interaction effects (top) and SNP main effects (bottom), whilst varying (a) the number of environmental variables, (b) proportion of trait variance explained by background GxE effects and (c) sample size. Power was assessed as the proportion of $60$ causal SNPs detected at $p < 0.01$ (Family Wise Error Rate; FWER $< 0.01$), where causal SNPs main and GxE interaction effects each explained $0.00016\%$ of trait variance. Simulations used genotypes subsampled from the UK Biobank and by default contained $N = 25K$ samples, $M = 100K$ SNPs, $6$ environmental variables that contributed to the ES and $24$ that did not (default parameters denoted by stars). We performed $20$ repeats for each scenario. See **Online methods** for full details of phenotype construction.

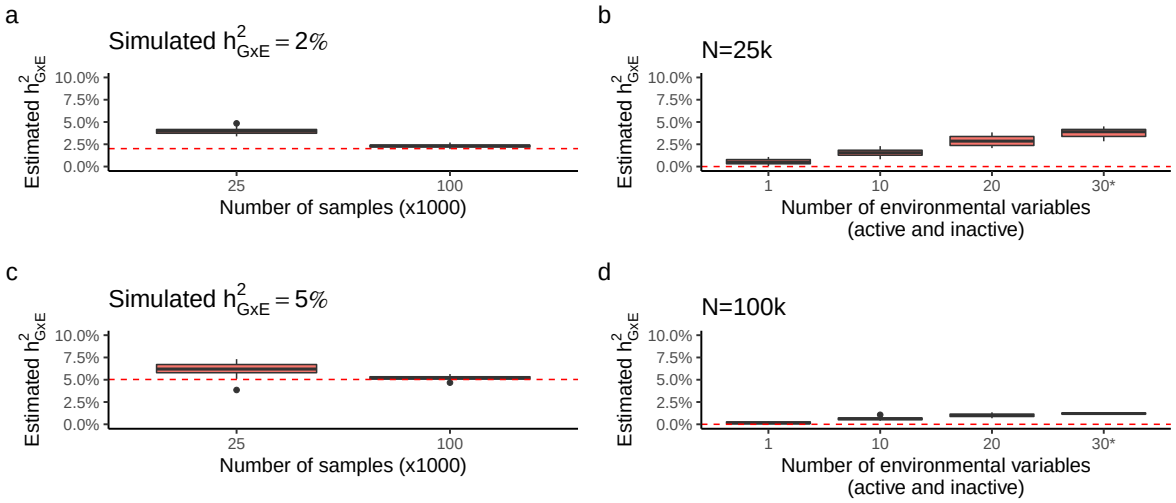



Figure S3: **Estimation of GxE heritability.** Estimates of SNP-GxE heritability whilst varying the number of environmental variables (b, d) and sample size (a, c). The red dotted line denotes the true SNP-GxE heritability used whilst constructing the simulation. We observed some upwards bias as the number of environmental variables increases (b, d), which is ameliorated with increased sample size (d). Phenotypes were constructed using $M = 100,000$ SNPs with $M_{\text{causal, main-effects}} = 80,000$ causal main effects and $M_{\text{causal, GxE-effects}} = 40,000$ causal interaction effects. See **Online methods** for full details of phenotype construction.

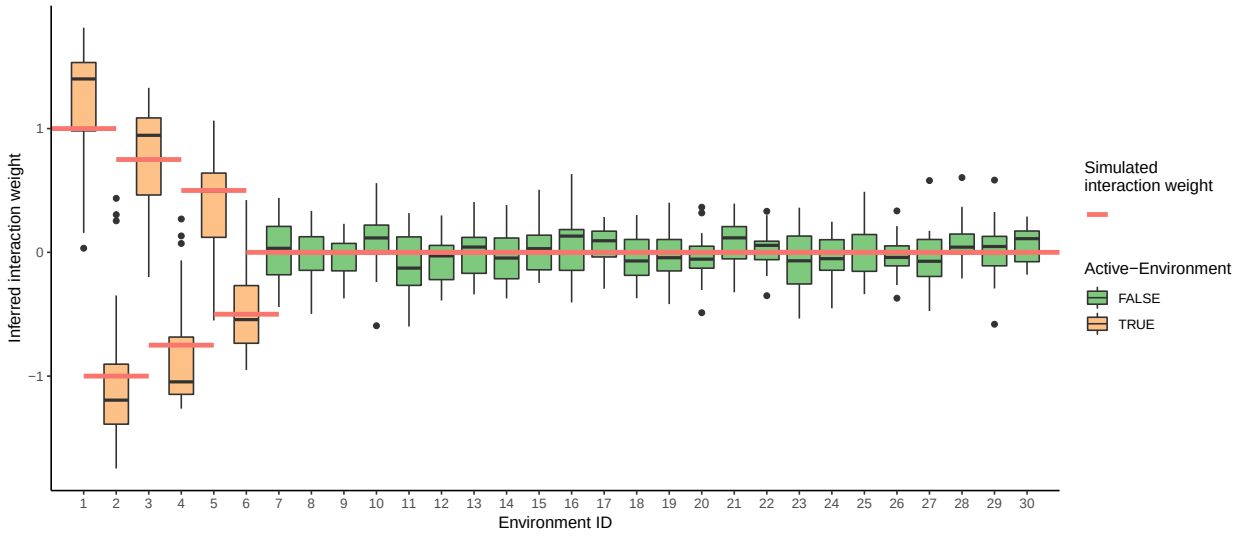



Figure S4: **Estimation of ES weights in simulation.** Boxplots of the environmental score (ES) weights estimated by LEMMA (left) over 20 simulations. Red lines denote true weights used to construct the simulated ES. Simulations performed with $N = 25k$ samples, $M = 100k$ SNPs and $L = 30$ environments (of which 6 were active). Phenotypes were constructed with $M_{\text{causal, main-effects}} = 5000$ SNPs explaining $20\%$ of trait variance and $M_{\text{causal, GxE-effects}} = 2500$ SNPs explaining $5\%$ of trait variance. LEMMA is invariant to a sign change in both the interaction weights and interaction SNP effects, so ES weights are automatically rescaled such that the largest weight is positive before plotting.

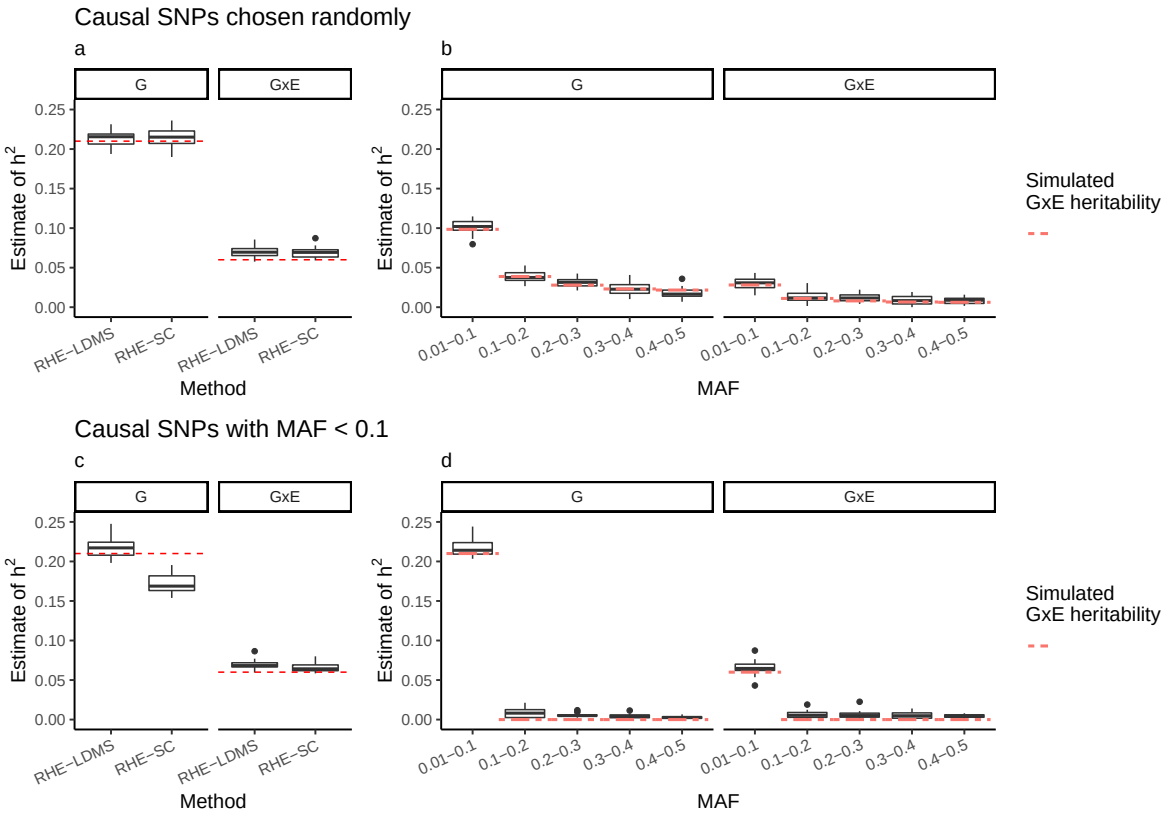



Figure S5: **Heritability estimates stratified by LD and MAF in simulation**. Comparison of heritability estimates using RHE-SC and RHE-LDMS when causal SNPs were drawn (a) at random or (c) only from rare (MAF ¡ 0.1) SNPs. Heritability estimates (using RHE-LDMS) stratified by MAF when causal SNPs were drawn (b) at random or (d) only from rare (MAF ¡ 0.1) SNPs. Simulations performed with $N = 25K$ samples, $M = 100K$ SNPs and the default simulation parameters described in **Online Methods**. Abbreviations; MAF, minor allele frequency; RHE-SC, randomised HE regression with a single SNP component[22]; RHE-LDMS, multi-component randomised HE-regression[23].

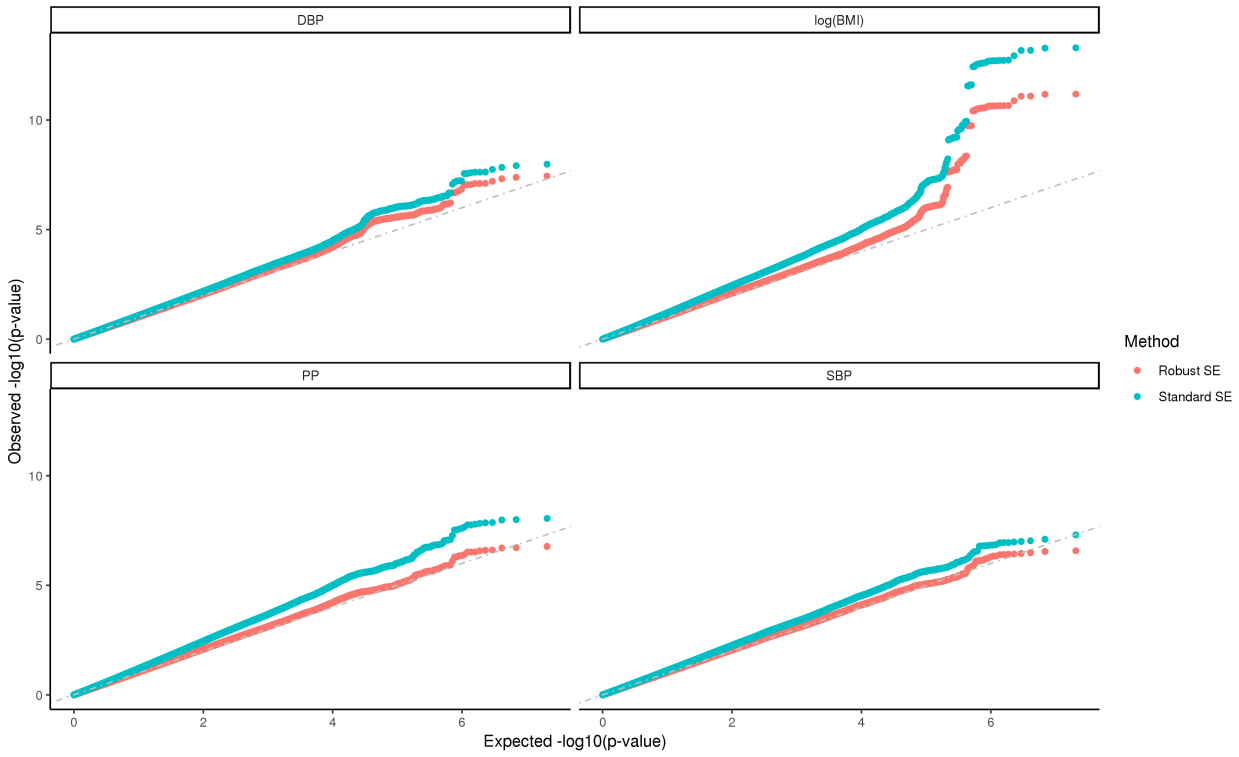



Figure S6: **Effect of using robust standard errors for GxE interaction tests in the UK Biobank.** QQ plots of the observed LEMMA $-\log_{10}(p)$ values for GxE interactions at imputed SNPs for four UK Biobank traits, with and without robust standard errors. The grey dotted line denotes expected $-\log_{10}(p)$-values under a null model. Association tests using 'Robust' standard errors are well calibrated in both homoskedastic and heteroskedastic regimes (see **Online Methods**) and are used in all follow up analysis. Genomic control statistics were $1.275, 1.271, 1.163, 1.111$ for $\log$BMI, PP, SBP and DBP respectively using homoskedastic standard errors and $1.062, 1.047, 1.037, 1.027$ for $\log$BMI, PP, SBP and DBP respectively using robust standard errors.

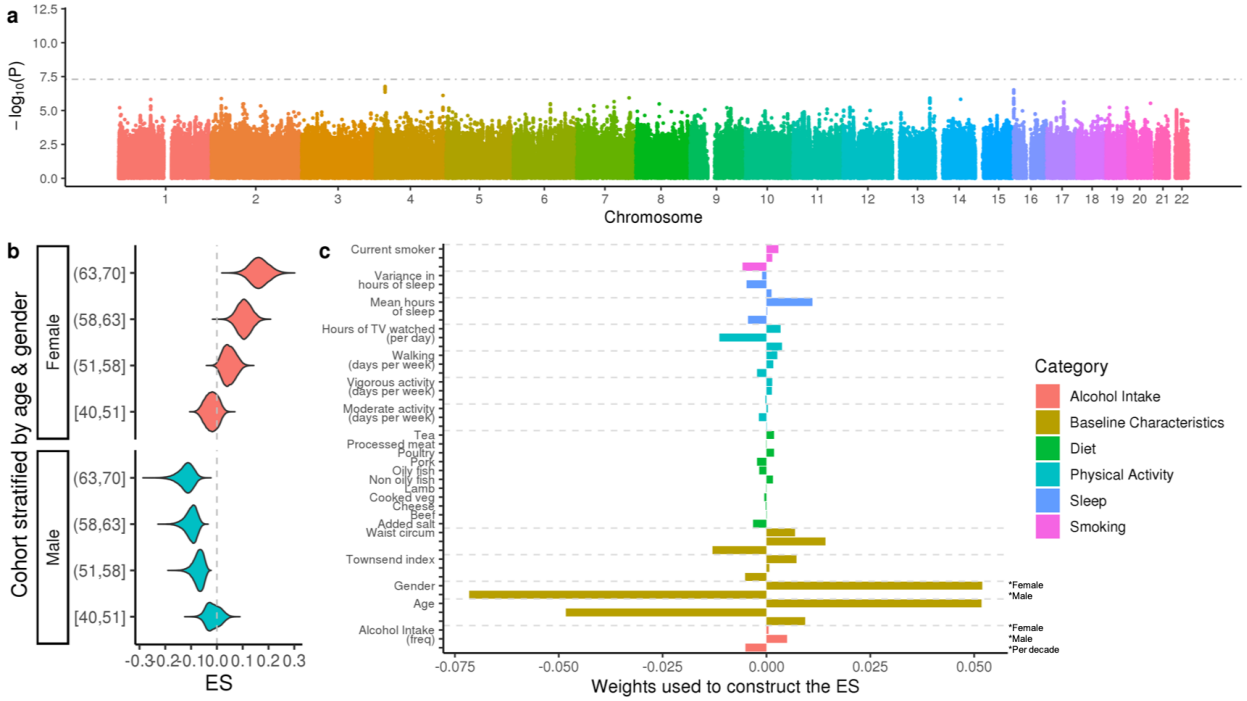



Figure S7: **GxE analysis of PP in the UK Biobank.** (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes $(p = 5 \times 10^{-8})$, $p$-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female specific weights is computed for each gender separately. Age is computed as the number of decades aged from 40. See **Online Methods** for details.

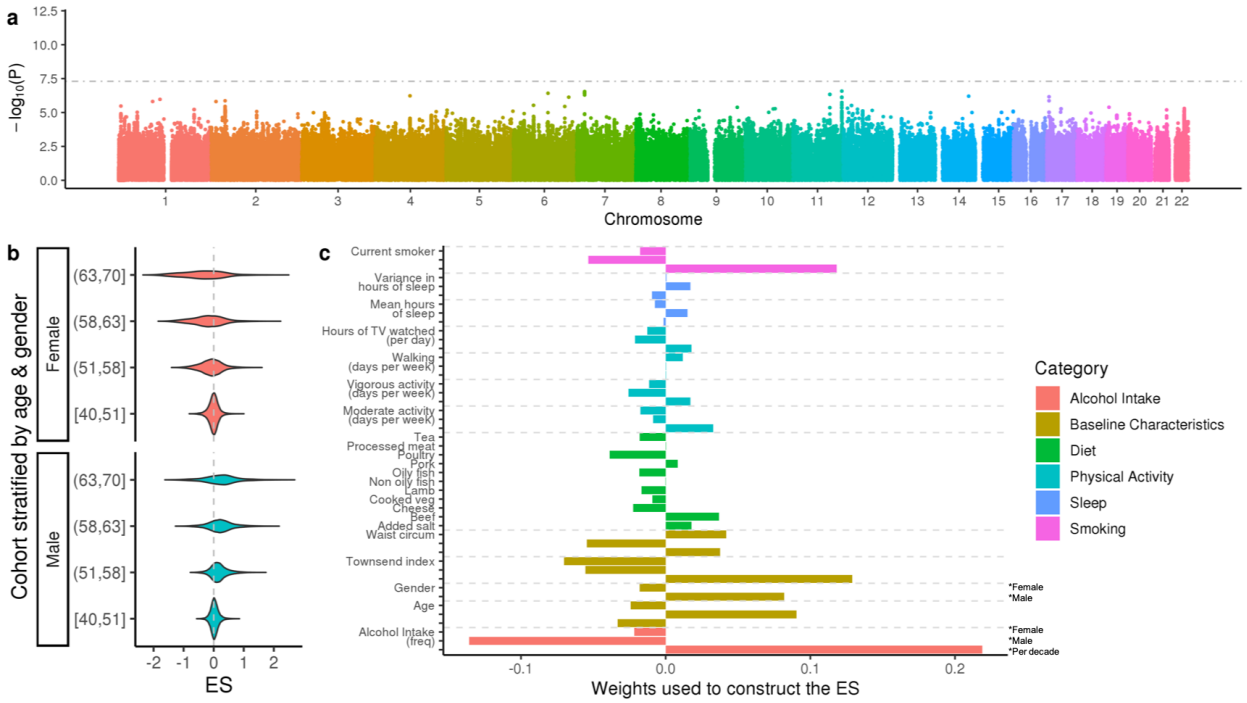



Figure S8: **GxE analysis of SBP in the UK Biobank.** (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes $(p = 5 \times 10^{-8})$, $p$-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female specific weights is computed for each gender separately. Age is computed as the number of decades aged from 40. See **Online Methods** for details.

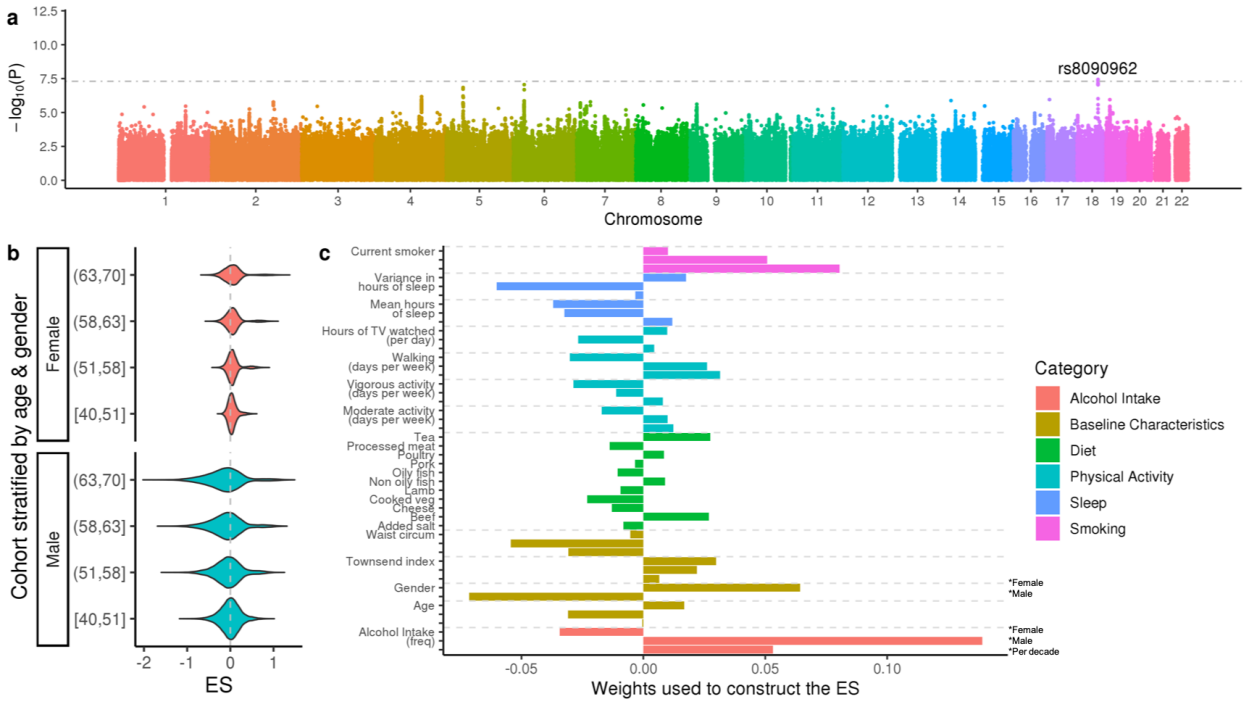



Figure S9: **GxE analysis of DBP in the UK Biobank.** (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes $(p = 5 \times 10^{-8})$, $p$-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female specific weights is computed for each gender separately. Age is computed as the number of decades aged from 40. See **Online Methods** for details.

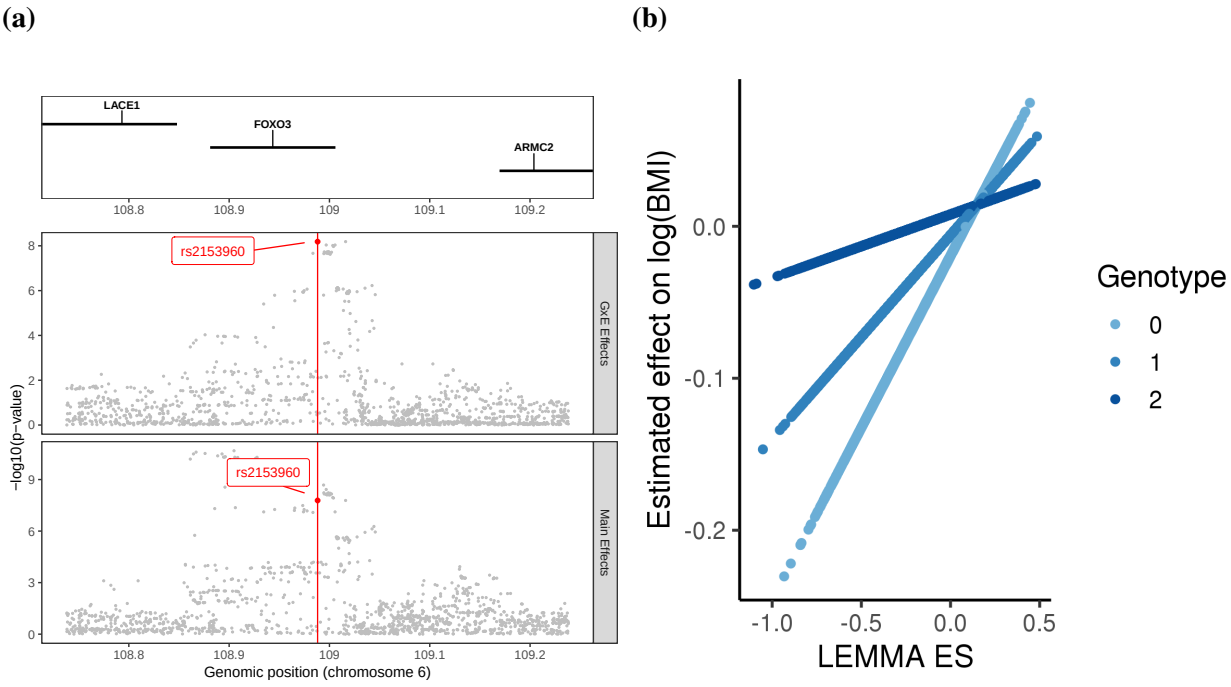



Figure S10: **Estimated GxE effect rs2153960 on** log**BMI.** (a) Regional plot of the main and interaction effects of SNPs within 250KB of rs2153960, (b) the estimated effect of rs2153960 on logBMI as a function of the environmental score (ES).

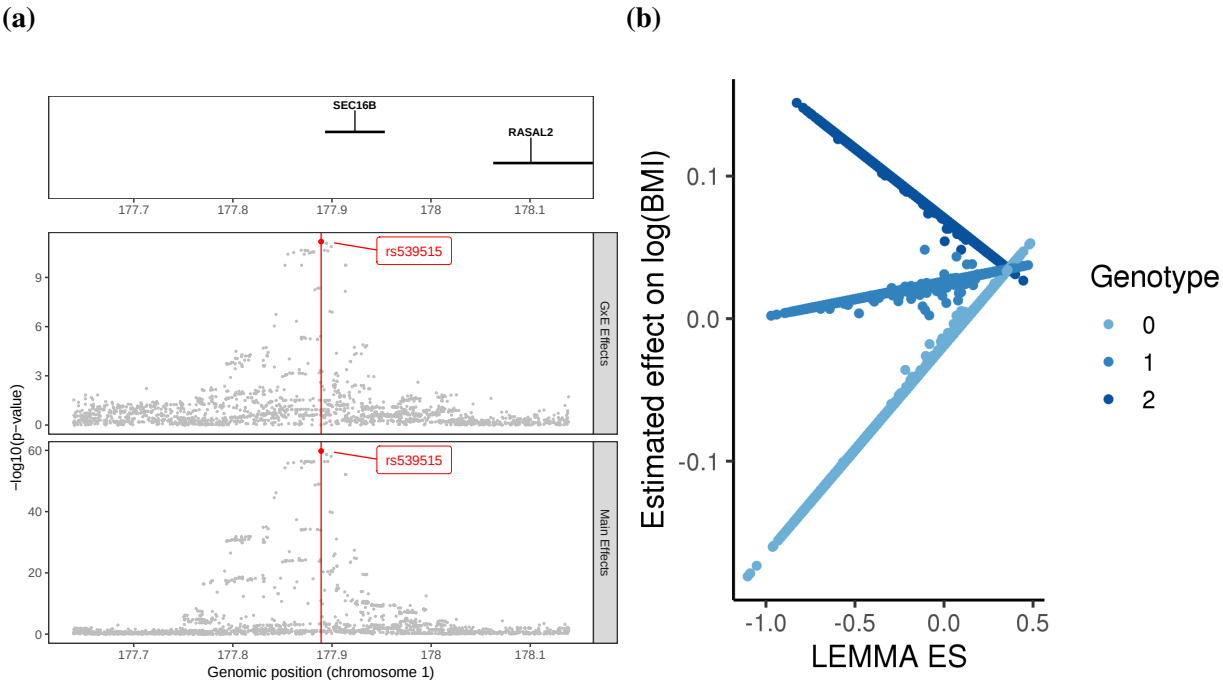



Figure S11: **Estimated GxE effect rs539515 on** log**BMI.** (a) Regional plot of the main and interaction effects of SNPs within 250KB of rs539515, (b) the estimaetd effect of rs539515 on logBMI as a function of the environmental score (ES).

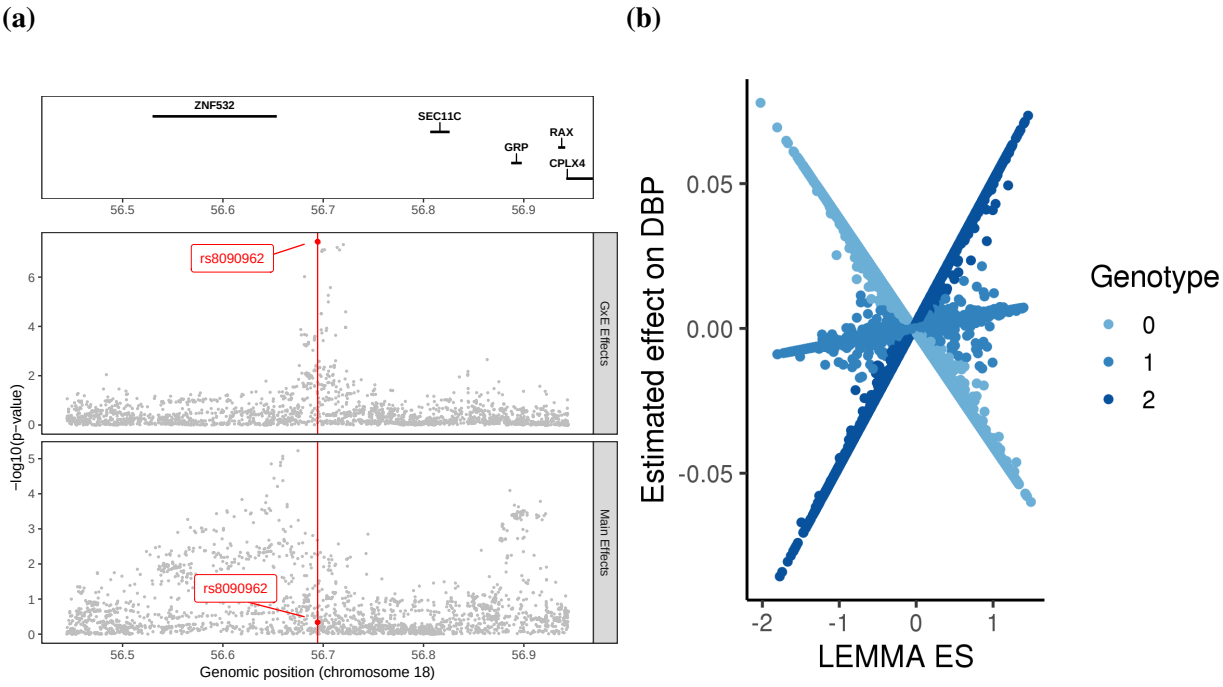
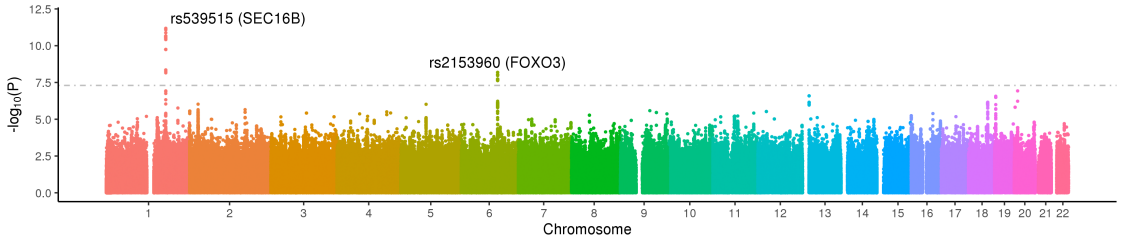
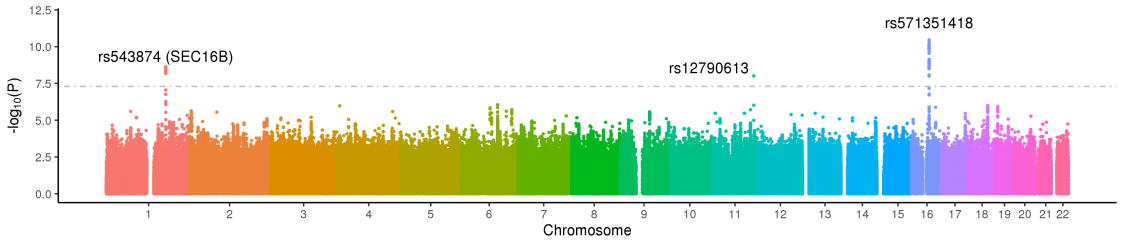



Figure S12: **Estimated GxE effect rs8090962 on DBP.** (a) Regional plot of the main and interaction effects of SNPs within 250KB of rs8090962, (b) the estimated effect of rs8090962 on DBP as a function of the environmental score (ES).

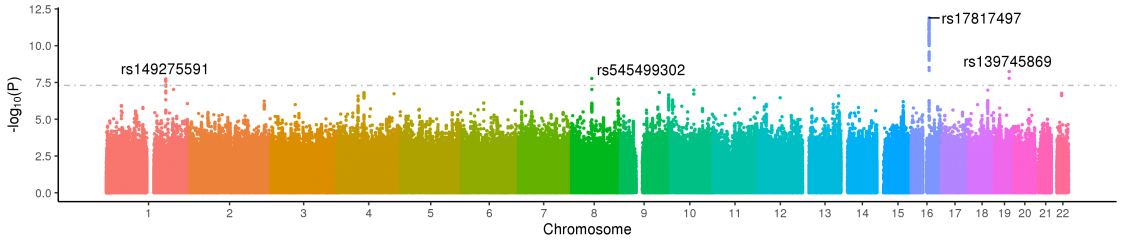
**(a)** LEMMA




**(b)** StructLMM




**(c)** F-test




**(d)** robust F-test

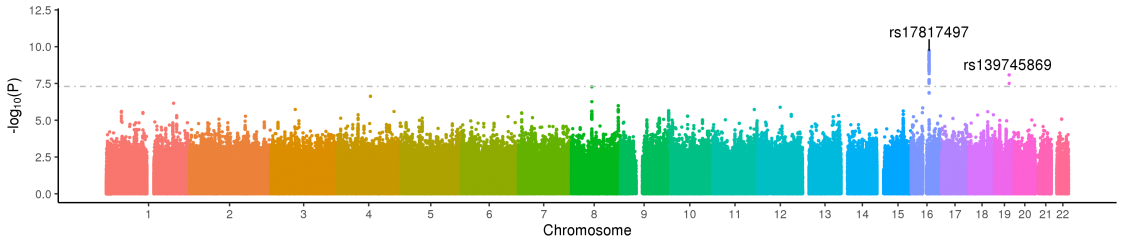



Figure S13: **GxE association statistics for** $\log$**BMI.** Manhattan plots displaying the negative $\log_1 0$ $p$ values from GxE interaction tests at $10,295,038$ imputed SNPs applied to $\log$BMI in the UK Biobank. GxE interaction tests were computed using (a) LEMMA, (b) StructLMM, (c) the F-test and (d) the robust F-test. The horizontal grey line denotes $(p = 5 \times 10^{-8})$.

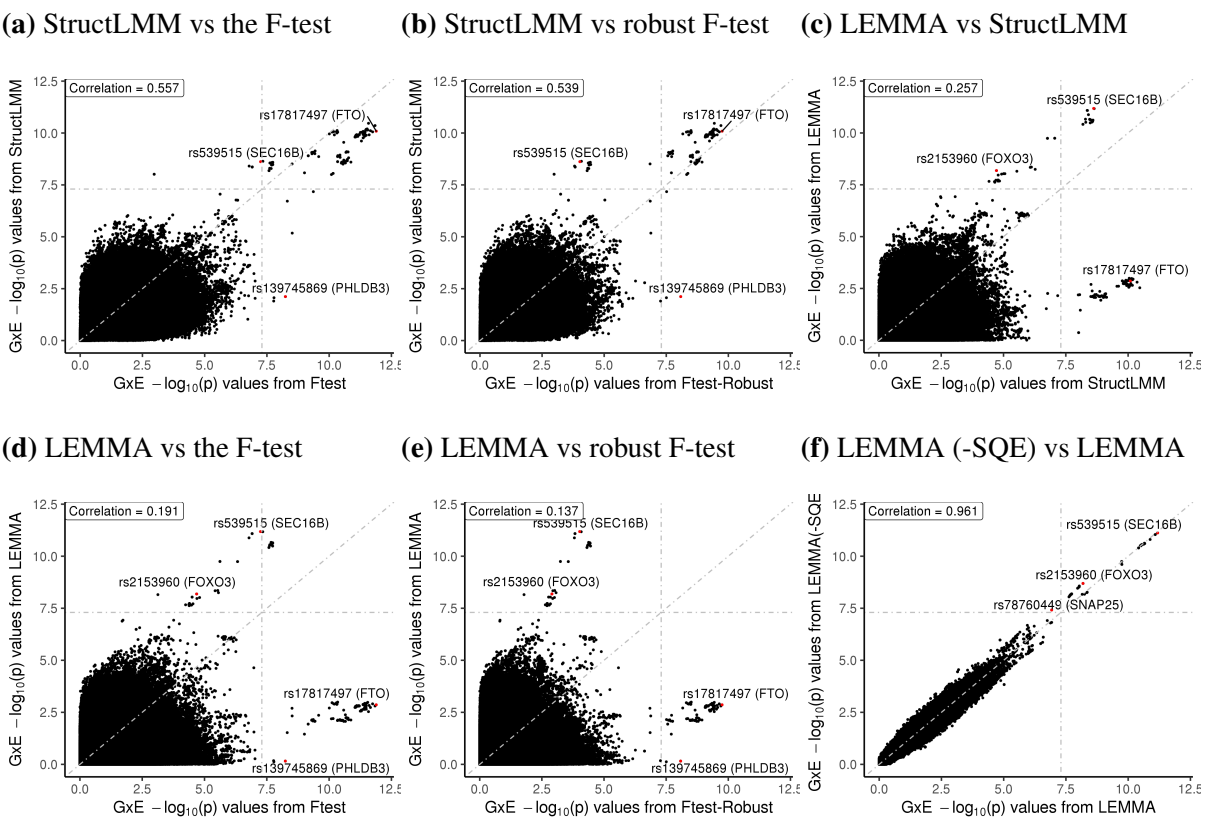



Figure S14: **Comparison of GxE association statistics for** log**BMI.** Comparison of negative $\log_{10}$ $p$ values obtained from LEMMA, StructLMM, the F-test and the robust F-test in an analysis of logBMI in the UK Biobank. Grey lines denote $(p = 5 \times 10^{-8})$ and the $y = x$ axis. Pearson correlation is shown in a label at the top left of each plot. Red points denote the sentinel SNP for each locus.

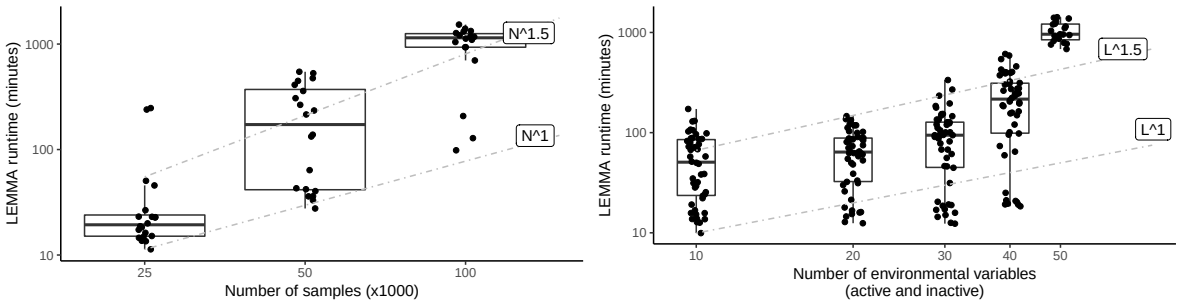



Figure S15: Log-log plots showing runtime of the variational bayes algorthim used to perform whole genome regression by LEMMA, as a function of sample size (left) and the number of environmental variables (right). Unless otherwise stated simulations were performed using $N = 25k$ samples, $M = 100k$ SNPs and $L = 30$ environmental variables. Phenotypes were constructed using $2500$ non-zero main effects expaining $20\%$ of variance, $1250$ nonzero interaction effects explaining $5\%$ of variance and $6$ active environmental variables. See **Online methods** for full details of phenotype construction.

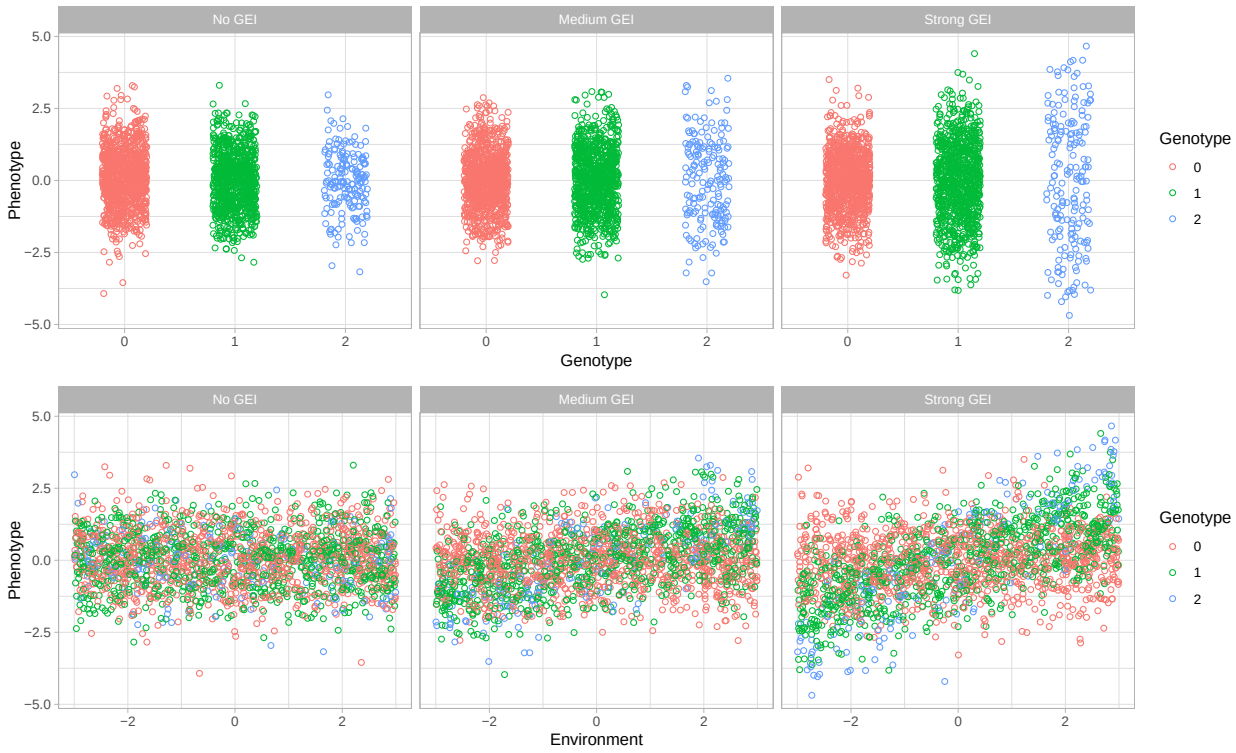



Figure S16: **Visualisation of differences in variance induced by a multiplicative Gene-x-Environment effect.** Differences in phenotypic variance by genotype group (top) and by strength of the environmental exposure (bottom). The phenotype was simulated using $2000$ individuals on the basis of a multiplication interaction between a single genotype (minor allele frequency $0.3$) and an environment (uniformly distributed over $[-2, 2]$). From left to right; the GxE interaction explained $0\%, 20\%, 40\%$ of trait variance.

Table S1: **Quality control and time to convergence of the WGR analyses**

| Trait | No. samples | No. SNPs | No. envs | No. envs-sq | No. other covars | No. covars total | Iterations for WGR converge | Time for WGR to converge* |
|---|---|---|---|---|---|---|---|---|
| log(BMI) | 281149 | 642095 | 42 | 30 | 24 | 96 | 1631 | 78 hours 18 mins |
| PP | 280749 | 642102 | 45 | 13 | 25 | 83 | 3821 | 183 hours 26 mins |
| SBP | 280749 | 642102 | 45 | 15 | 25 | 85 | 967 | 46 hours 25 mins |
| DBP | 280749 | 642102 | 45 | 15 | 25 | 85 | 646 | 31 hours 00 mins |

Time for the whole genome regression analysis to converge is reported for four quantitative traits in the UK Biobank, as well as the number of SNPs and samples passing quality control and the number of covariates controlled for. 'Other covariates' consisted of the top 20 genetic principal components as reported by the UK Biobank, age$^3$, age$^2\times$ gender, age$^3\times$ gender, a binary indicator for the genotype chip and (for blood pressure traits only) BMI. Environmental variables used (including lower orders of age and gender) are described in (**Online Methods**). To control for potential bias due to non-linear dependance between the phenotype and heritable environmental variables, we tested each environmental variable and included any significant squared effects as additional covariates (**Online Methods**) *based on the average per-iteration cost of 243 seconds, using 32 cores distributed across a cluster with Xeon E5-2667 v4 3.2Ghz processors.

| Trait | Genotyped (SC) | | Genotyped (LDMS) | | CommonImputed (LDMS) | |
|---|---|---|---|---|---|---|
| | $h^2_{\text{G}}$ (s.e) | $h^2_{\text{GxE}}$ (s.e) | $h^2_{\text{G}}$ (s.e) | $h^2_{\text{GxE}}$ (s.e) | $h^2_{\text{G}}$ (s.e) | $h^2_{\text{GxE}}$ (s.e) |
| log BMI | 0.259 (0.069) | 0.071 (0.009) | 0.237 (0.126) | 0.086 (0.024) | 0.274 (0.056) | 0.093 (0.028) |
| PP | 0.233 (0.039) | 0.075 (0.018) | 0.203 (0.084) | 0.111 (0.021) | 0.228 (0.051) | 0.125 (0.028) |
| SBP | 0.24 (0.053) | 0.033 (0.003) | 0.223 (0.095) | 0.038 (0.017) | 0.251 (0.05) | 0.039 (0.023) |
| DBP | 0.277 (0.034) | 0.014 (0.001) | 0.231 (0.079) | 0.016 (0.017) | 0.254 (0.05) | 0.016 (0.02) |

Table S2: **Partitioned heritability estimates for four quantitative traits in the UK Biobank**. Comparison of the heritability estimates obtained using genotyped SNPs with RHE-SC, genotyped SNPs with RHE-LDMS, and common imputed SNPs (MAF$>0.01$ in the full UK Biobank cohort) with RHE-LDMS. GxE heritability estimates were were obtained using the ES from each model fit. All analyses controlled for the same covariates used in the WGR analysis (including the top 20 principal components). Abbreviations; s.e, standard error estimated using the block jacknife (see **Online Methods**); $h^2_{\text{G}}$, heritability due to additive genetic effects; $h^2_{\text{GxE}}$, heritability due to multiplicative GxE effects; RHE, randomised HE-regression[22,23]; SC, single SNP component; LDMS, SNPs stratified by minor allele frequency and LDscore (20 components).

Table S3: **Comparison of the number of genome-wide significant GxE associations and genomics control statistics from a GxE analysis of** $\log$**BMI in the UK Biobank**

| Method | No. signals | Genomic control ($\chi^2$) | Genomic control (p-values)* |
|---|---|---|---|
| LEMMA | 2 | 1.062 | 1.038 |
| LEMMA-S | 5 | 1.275 | 1.164 |
| StructLMM (-SQE) | 3 | NA | 1.236 |
| F-test (-SQE) | 4 | NA | 1.372 |
| robust F-test (-SQE) | 2 | NA | 1.034 |
| LEMMA (-SQE) | 3 | 1.065 | 1.04 |
| LEMMA-S (-SQE) | 6 | 1.288 | 1.171 |

The number of independent loci (atleast 0.5cM apart) with genome-wide significant GxE interaction effects and genomic control statistics for seven different methods applied to $\log$BMI in the UK Biobank. Genomic control is computed from GxE interaction tests statistics from $10,295,038$ imputed SNPs. Abbreviations; LEMMA-S, LEMMA with a homoskedastic test statistic (see **Online Methods**); (-SQE), significant squared environmental variables (Bonferroni correction) not included as additional covariates.

*The test statistics from StructLMM, F-test and the robust F-test are not $\chi^2_1$ distributed. Hence for these methods we use $\lambda_{\text{GC}} = \log_{10}(m)/\log_1 0(0.5)$, where $m$ is the median $p$-value, to denote the genomic control statistic as suggested by Moore *et al.*[7].

Table S4: **Comparison of the number of genome-wide significant GxE associations and genomics control statistics from GxE analyses of four quantitative traits in the UK Biobank**

| Trait | Method | No. signals | Genomic control |
|---|---|---|---|
| log(BMI) | LEMMA | 2 | 1.062 |
| log(BMI) | LEMMA-S | 5 | 1.275 |
| PP | LEMMA | 0 | 1.047 |
| PP | LEMMA-S | 2 | 1.271 |
| SBP | LEMMA | 0 | 1.037 |
| SBP | LEMMA-S | 1 | 1.163 |
| DBP | LEMMA | 1 | 1.027 |
| DBP | LEMMA-S | 2 | 1.111 |

The number of independent loci (atleast 0.5cM apart) with genome-wide significant GxE interaction effects and genomic control statistics for four different in the UK Biobank. Genomic control is computed from GxE interaction tests statistics from $10,295,038$ imputed SNPs. Abbreviations; BMI, body mass index; PP, pulse pressure; SBP, systolic blood pressure; DBP, diastolic blood pressure; LEMMA-S, LEMMA with a homoskedastic test statistic (see **Online Methods**).