Genetics and Genomics

Tradeoffs in Modeling Context Dependency in Complex Trait Genetics

Reviewed Preprint
v2 • March 13, 2025
Revised by authors

Reviewed Preprint
v1 • July 17, 2024

Eric Weine, Samuel Pattillo Smith, Rebecca Kathryn Knowlton, Arbel Harpak

Department of Integrative Biology, The University of Texas at Austin, Austin, United States • Department of Population Health, The University of Texas at Austin, Austin, United States • Department of Human Genetics, University of Chicago, Chicago, United States • Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, United States

eLife Assessment

It is known from model organisms that genes' effects on traits are often modulated by environmental variables, but similar gene-by-environment (GxE) interactions have been difficult to detect using statistical analyses of genomic data, e.g., in humans. This study introduces a new framework to estimate gene-by-environment effects, treating it as a bias-variance tradeoff problem. The authors **convincingly** show that greater statistical power can be achieved in detecting GxE if an underlying model of polygenic GxE is assumed. This polygenic amplification model is a truly novel view with **fundamental** promise for the detection of GxE in genomic datasets, especially with continued development to detect more complex signals of amplification.

https://doi.org/10.7554/eLife.99210.2.sa2

Abstract

Genetic effects on complex traits may depend on context, such as age, sex, environmental exposures or social settings. However, it is often unclear if the extent of context dependency, or Gene-by-Environment interaction (GxE), merits more involved models than the additive model typically used to analyze data from genome-wide association studies (GWAS). Here, we suggest considering the utility of GxE models in GWAS as a tradeoff between bias and variance parameters. In particular, We derive a decision rule for choosing between competing models for the estimation of allelic effects. The rule weighs the increased estimation noise when context is considered against the potential bias when context dependency is ignored. In the empirical example of GxSex in human physiology, the increased noise of context-specific estimation often outweighs the bias reduction, rendering GxE models less useful when variants are considered independently. However, we argue that for complex traits, the joint consideration of context dependency across many variants mitigates both noise and bias. As a result, polygenic GxE models can improve both estimation and trait prediction. Finally, we exemplify (using GxDiet effects on longevity in fruit flies) how analyses based on independently ascertained “top hits” alone can be misleading, and that considering polygenic patterns of GxE can improve interpretation.

Introduction

In organisms and study systems where the environment can be tractably manipulated, gene-by-environment interactions (GxE) are the rule, not the exception [1–5]. Yet, in complex (polygenic) human traits, there are but a few cases in which models that incorporate GxE explain data—such as Genome-Wide Association Study (GWAS) data—better than parsimonious models that assume additive contributions of genetic and environmental factors [6–8]. This is true for both physical environments but also for other definitions of "E," broadly construed to be any context that modifies genetic effects, such as age, sex, or social setting [9–16].

GWAS commonly estimate marginal additive effects of an allele on a trait. The estimand here can be thought of as the average effect of the allele over a distribution of multidimensional contexts [17]. With this view, some differences in allelic effects across contexts are likely omnipresent, but may very well be small, such that the cost of including additional parameters (for context-specific effects) outweighs the benefit of measuring heterogeneous effects.

Here, we consider this problem and its connection to the currently underwhelming utility of GxE models in GWAS. First, we rigorously describe the statistical trade-off involved in estimating context-specificity at the level of a single variant. Then, we highlight ways in which this trade-off might change as we consider GxE in complex traits, involving numerous genetic variants simultaneously.

We begin by framing the problem of estimating context-specificity at an individual variant as a bias–variance trade-off. For example, consider the estimation of an allelic effect on lung cancer risk that depends on smoking status. When the allelic effect is estimated from a sample without considering smoking status, the estimate would be biased with respect to the true effect in smokers. We can estimate the effect separately in smokers and non-smokers to eliminate the bias, but the consideration of the additional parameters—smoking status-specific effects—has an associated cost of increasing the estimation variance, compared to an estimator that ignores smoking status. This bias–variance trade-off is closely related to the "signal-to-noise" ratio, where the signal of interest is the true difference in context-specific allelic effects. To demonstrate this tradeoff in real data, we consider sex-specific effects on physiological traits in humans. We show that for the majority of traits, it is typically unhelpful to model sex dependency for individual sites since the increase in noise vastly outweighs the signal.

We then consider the extension to GxE in complex traits. Complex trait variation is primarily due to numerous genetic variants of small effects distributed throughout the genome [18–21]. Simultaneously considering GxE across multiple variants may decrease estimation noise if the extent and mode of context-specificity is similar across numerous variants. This would tilt the scale in favor of context-dependent estimation. In addition, we show how conventional approaches for detecting and characterizing GxE, which focus on the most significant associations, may lead to erroneous conclusions. Finally, we discuss implications for complex trait prediction (with polygenic scores). We suggest a future focus on prediction methods that empirically learn the extent and nature of context dependency by simultaneously considering GxE across many variants.

Results and Discussion

Modeling context-dependent effect estimation as a bias–variance trade-off

The problem setup. We consider a sample of n + m individuals characterized as being in one of two contexts, A or B. n of the individuals are in context A with the remaining m individuals in context B. We measure a continuous trait for each individual, denoted by

$$\overbrace{y_1, \ldots, y_n}^{A}, \overbrace{y_{n+1}, \ldots, y_{n+m}}^{B}.$$

We begin by considering the estimation of the effect of a single variant on the continuous trait. We assume a generative model of the form

$$y_i \sim \begin{cases} \mathcal{N}(\alpha_A + \beta_A g_i, \sigma_A^2) & \text{if } i \in \{1, \ldots, n\} \\ \mathcal{N}(\alpha_B + \beta_B g_i, \sigma_B^2) & \text{if } i \in \{n+1, \ldots, n+m\}, \end{cases} \quad (1)$$

where β_A and β_B are fixed, context-specific effects of a reference allele at a biallelic, autosomal variant i, $g_i \in \{0, 1, 2\}$ is the observed reference allele count. α_A and α_B are the context-specific intercepts, corresponding to the mean trait for individuals with zero reference alleles in context A and B, respectively. σ_A^2 and σ_B^2 are context-specific observation variances. We would like to estimate the allelic effects β_A and β_B.

Estimation approaches

We compare two approaches to this estimation problem. The first approach, which we refer to as GxE estimation, is to stratify the sample by context and separately perform an ordinary least squares (OLS) regression in each sample. This approach yields two estimates, $\hat{\beta}_A$ and $\hat{\beta}_B$, the OLS estimates of β_A and β_B of the generative model in Eq. 1, respectively. This estimation model is equivalent to a linear model with a term for the interaction between context and reference allele count, in the sense that context-specific allelic effect estimators have the same maximum likelihood estimators in the two models (see Supplementary Materials).

The second approach, which we refer to as additive estimation, is to perform an OLS regression on the entire sample and use the allelic effect estimate to estimate both β_A and β_B. We denote this estimator as $\hat{\beta}_{A\cup B}$, to emphasize that the regression is run on all individuals from context A and context B. This estimation model posits that for i = 1, ..., n + m,

$$y_i \sim \mathcal{N}(\alpha_{A\cup B} + \beta_{A\cup B} g_i, \sigma_{A\cup B}^2), \quad (2)$$

where $\alpha_{A\cup B}$ is the mean trait value for an individual with zero reference alleles, $\beta_{A\cup B}$ is the additive allelic effect and $\sigma_{A\cup B}^2$ is the observation variance which is independent of context. Notably, this model differs from the generative model assumed above: $\beta_{A\cup B}$ may not equal β_A and β_B; in addition, this model ignores heteroskedasticity across contexts.

Error analysis

We focus on the mean squared error (MSE) of the additive and GxE estimators for the allelic effect in context A. The estimator minimizing the MSE may differ between contexts A and B, but the analysis for context B is analogous. When selecting between these two estimation approaches, a bias–variance decomposition of the MSE is useful.

Based on OLS theory ([22, Theorem 11.3.3]), under the model specified above we have

$$\hat{\beta}_A \sim \mathcal{N}(\beta_A, V_A),$$

where $V_A = \frac{\sigma_A^2}{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2}$ and $\bar{g}_A$ is the mean genotype of individuals in context A. The unbiasedness of the GxE estimator implies

$$MSE(\hat{\beta}_A, \beta_A) = V_A,$$

where $MSE(\hat{\beta}_A, \beta_A)$ is the mean squared error of estimating β_A with $\hat{\beta}_A$. The case of the additive estimator, $\hat{\beta}_{A\cup B}$, is a bit more involved. As we show in the **Methods** section, we can write

$$\hat{\beta}_{A\cup B} = \omega_A \hat{\beta}_A + \omega_B \hat{\beta}_B \tag{3}$$

for non-negative weights ω_A and ω_B (that need not sum to 1). Further, we show in Eq. 7 of the **Methods** section that $\omega_A \propto nH_A$ and $\omega_B \propto mH_B$, where H_A and H_B are the sample heterozygozities in contexts A and B, respectively. Using Eq. 3, we may write

$$\begin{aligned} MSE(\hat{\beta}_{A\cup B}, \beta_A) &= Bias^2(\hat{\beta}_{A\cup B}, \beta_A) + Var(\hat{\beta}_{A\cup B}) \\ &= \left((\omega_A - 1)\beta_A + \omega_B \beta_B\right)^2 + \omega_A^2 V_A + \omega_B^2 V_B, \end{aligned}$$

where V_B is defined analogously to V_A. Thus, with MSE as our metric for comparison, we prefer the GxE estimator in context A when

$$MSE(\hat{\beta}_{A\cup B}, \beta_A) > MSE(\hat{\beta}_A, \beta_A),$$

or, if and only if

$$\left((\omega_A - 1)\beta_A + \omega_B \beta_B\right)^2 + \omega_A^2 V_A + \omega_B^2 V_B > V_A. \tag{4}$$

We refer to Eq. 4 as the "decision rule," since it guides us on the more accurate estimator; to minimize the MSE, we will use the context-specific estimator if and only if the inequality is satisfied.

To gain some intuition about the important parameters here, we first consider the case of equal allele frequencies (and hence equal heterozygozities) in both contexts and equal estimation variance in both contexts. In this case, the GxE estimator is advantaged by larger context-specificity (larger $|\beta_A - \beta_B|$) and disadvantaged by larger estimation noise (larger $V_A = V_B$) (**Fig. 1**). In fact, the decision boundary (i.e. the point at which the two models have equal MSE) can be written as a linear combination of $|\beta_A - \beta_B|$ and $\sqrt{V_A}$ (**Fig. 1C**). In this special case, we show in the **Methods** section that **Eq. 4** is an equality when

$$\sqrt{\frac{m}{2n}}|\beta_A - \beta_B| - \sqrt{V_A} = 0. \tag{5}$$

More generally, in the case where $H_A = H_B$ but $V_A \neq V_B$, we show in the **Methods** section that we can write **Eq. 4** as

$$\frac{(\beta_A - \beta_B)^2}{V_A} > \frac{1 + \omega_A}{1 - \omega_A} - \frac{V_B}{V_A}. \tag{6}$$

This dimensionless re-parameterization of the decision rule makes explicit its dependence on three factors $\frac{(\beta_A-\beta_B)^2}{V_A}$ can be viewed as the "signal-to-noise" ratio: it captures the degree of context-specificity (the signal) relative to the estimation noise in the focal context, A. $A.\ \frac{1+\omega_A}{1-\omega_A}$ is the relative contribution to heterozygosity, which equals the relative contribution to variance in the independent variable of the OLS regression of **Eq. 2**. $\frac{V_B}{V_A}$ is the ratio of context-specific estimation noises. In the **Supplementary Materials**, we extend the decision rule for the case of a continuous context variable.

For a given trait and context, we can consider the behavior of the decision rule across variants with variable allele frequencies and allelic effects. The ratio of estimation noises, $r := \frac{V_A}{V_B}$, will not be constant. However, in some cases, considering a fixed *r* across variants is a good approximation. In GWAS of complex traits, each variant often explains a small fraction of trait variance. As a result, the estimation noise is effectively a matter of trait variance and heterozygosity alone. If per-site heterozygosity is similar in strata A and B, as it is, for example, for autosomal variants in biological males and females, *r* is approximately fixed across variants [9].

Fig. 2 illustrates the linearity of the decision boundary under the assumption that *r* is fixed across variants. It also shows that the slope of the decision boundary changes as a function of r. Intuitively, we are less likely to prefer GxE estimation for the noisier context. In fact, for sufficiently small values of *r* (e.g. $r < \frac{1}{3}$ for $\omega_A = \frac{1}{2}$), $\frac{1+\omega_A}{1-\omega_A} - \frac{V_B}{V_A}$ will be negative. This corresponds to the situation where $V_A \ll V_B$, in which case the additive estimator is never preferable to the GxE estimator in estimating β_A, as the signal-to-noise ratio is always non-negative. Typically, this will also imply that the additive estimator is greatly preferable for estimating β_B, as $\hat{\beta}_B$ will be extremely noisy.

It is natural to ask where the decision rule of **Eq. 4** falls with respect to empirical GWAS data. We considered the example of biological sex as the context (GxSex), and examined sex-stratified GWAS data across 27 continuous physiological traits in the UK Biobank [9, 23]. For each of nine million variants, we estimated the difference in sex-specific effects and the variance of each marginal effect estimator in males. Then, using an estimate of the ratio of sex-specific trait variances as a proxy for the ratio of estimation variances of males and females, we approximated the linear decision boundary between the additive and GxE estimators (**Fig. 3A,B**; **Text S1**). To demonstrate the accuracy of our decision rule, we employed a data-splitting technique where we estimate the MSE difference between estimators in a training set and evaluate the accuracy in a holdout set (**Fig. S1**).

For almost all traits examined, very few allelic effects in males are expected to be more accurately estimated using the male-specific estimator (usually between 0% and 0.1%). Notable exceptions to this rule are testosterone, sex hormone binding globulin (SHBG), and waist-to-hip ratio adjusted for body mass index (BMI), for which roughly .5% of allelic effects are expected to be better estimated with the GxE model (**Fig. 3B**). However, when considering only SNPs that are genome-wide significant in males (marginal p-value $< 5 \times 10^{-8}$ in males), many traits show a much larger proportion of effects that would be better estimated by the GxE model. At an extreme, for testosterone, all genome-wide significant SNPs are expected to be better estimated by the GxE model. In addition, a large fraction of genome-wide significant effects are better estimated with the GxE model for creatinine (62%), arm fat-free mass (24%), waist-to-hip ratio (19%) and SHBG (18%) as well (**Fig. 3,D**).

The decision rule we derived could potentially guide more accurate allelic effect estimation approaches. However, the consideration of GxE pattern sharing across many variants (polygenic GxE) can alter both bias and variance and therefore the tradeoff. In our discussion of complex

Figure 1

Bias-Variance tradeoff for single-site estimation with equal estimation noise and equal heterozygozity across contexts. The x-axis shows the difference in context-specific effects, while the y-axis shows the standard deviation of the context-specific estimators—both in raw measurement units. The color on the plot indicates the difference between the additive and GxE estimators in bias (A), variance (B) or MSE (C). Only the additive estimator is potentially biased. The bias is proportional to the difference in context-specific effects and independent of the estimation noise. **(B)** The difference in variance is is proportional to context-specific estimation noise and independent of the difference of context-specific effects. **(C)** The decision boundary is linear in both the estimation noise and the difference between context-specific effects.

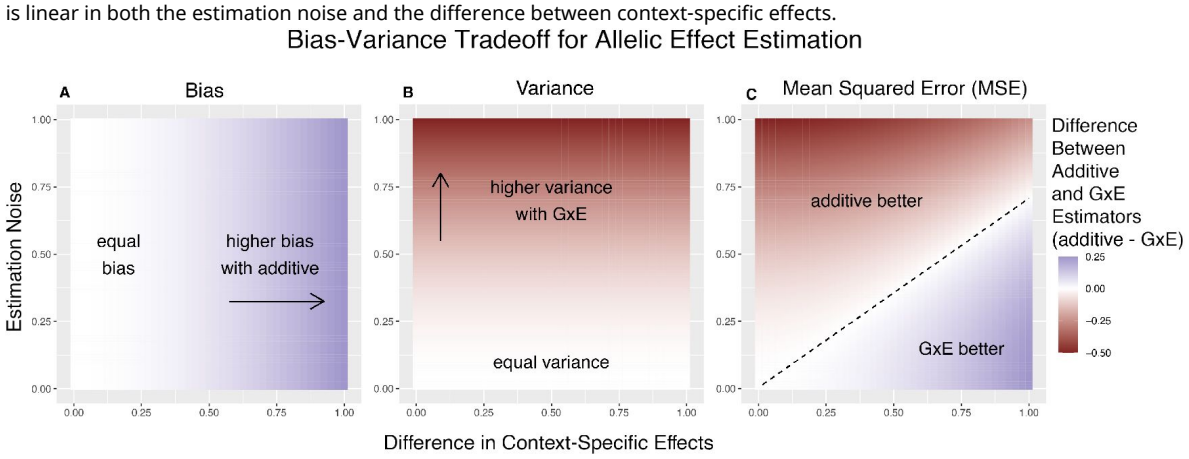
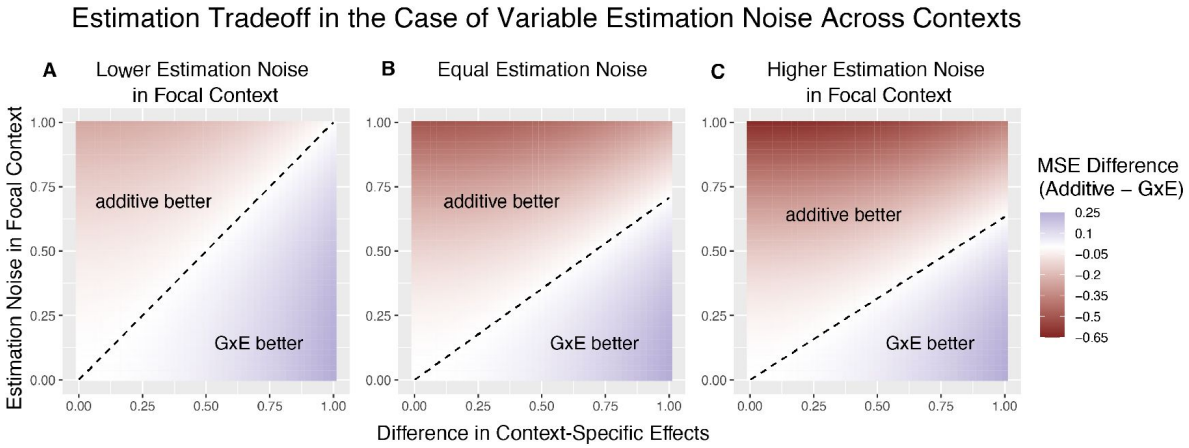


Figure 2

The decision boundary with different ratios of context-specific estimation noises. In all panels, the heterozygozity of the variant is assumed to be equal across contexts. The x and y axes are the same as in **Fig. 1**. Estimation noise in the focal context, A, is half that of the other context, B. **(B)** Estimation noise is equal in both contexts. **(C)** Estimation noise in focal context is double that of the other context.



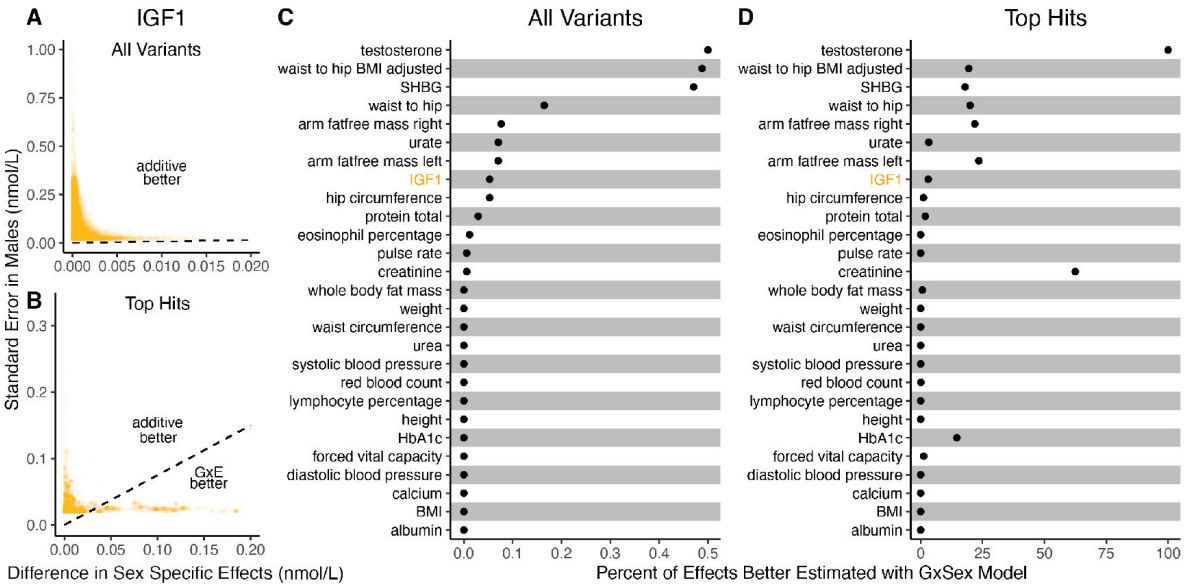


Figure 3

Applying the decision rule to sex-dependent effects on human physiological traits. **(A-B)** The x-axis shows the estimated absolute difference between the effect of variants in males and females. The y-axis shows the measured standard error for each variant in males, the focal context here. The dashed line shows the decision boundary for effect estimation in males. The difference in MSE between estimation methods increases linearly with distance from the dashed line, as in **Fig. 2**. If a variant falls above (below) the line, the additive (GxE) estimator has a lower MSE. (A) shows a random sample of 15K single nucleotide variants whereas (B) shows only variants with a marginal p-value less than 5×10^{-8} in males. **(C-D)** The percent of effects in males which would be better estimated by the GxE estimator, across continuous physiological traits. To estimate these percentages, one single nucleotide variant is sampled from each of 1,700 approximately independent autosomal linkage blocks, and this procedure is repeated 10 times. Shown are average percentages across the 10 iterations.

traits that follows, we therefore expand on the rule through qualitative consequences of polygenic GxE, and no longer stick to the analytical single variant rule.

Context dependency in complex traits

At the single variant level, and specifically when variants are considered independently from one another, we have discussed how the accurate estimation of allelic effects can be boiled down to a bias–variance tradeoff. For complex traits, genetic variance is often dominated by the contribution of numerous variants of small effects that are best understood when analyzed jointly [8, 20, 24–28]. It stands to reason that to evaluate context-dependence in complex traits, we would also want to jointly consider polygenic patterns, rather than just the patterns at the loci most strongly associated with a trait [3, 13, 29–32].

Motivated by this rationale, we recently inferred polygenic GxSex patterns in human physiology [9]. One pattern that emerged as a common mode of GxSex across complex physiological traits is "amplification": a systematic difference in the magnitude of genetic effects between the sexes. Moving beyond sex and considering any context, amplification can happen if, for example, many variants regulate a shared pathway that is moderated by a factor—and that factor varies in its distribution among contexts. Amplification is but one possible mode of polygenic GxE, but can serve as a guiding example for ways in which GxE may be pervasive but difficult to characterize with existing approaches [9, 16, 33, 34]. In what follows, we will therefore use the example of pervasive amplification (across causal effects) to illustrate the interpretive advantage of considering context dependency across variants jointly, rather than independently.

A focus on "top hits" may lead to mis-characterization of polygenic GxE

A common approach to the analysis of context dependency involves two steps. First, categorization of context dependency (or lack thereof) is performed for each variant independently. Second, variants falling under each category are counted and annotated across the genome. Some recent examples of this approach towards the characterization of GxE in complex traits include studies of GxSex effects on flight performance in *Drosophila* [35], GxSex effects on various traits in humans [36, 37] and GxDietxAge effects on body weight in mice [38].

Characterizing polygenic trends by summarizing many independent hypothesis tests may miss GxE signals that are subtle and statistically undetectable at each individual variant, yet pervasive and substantial cumulatively across the genome. To characterize polygenic GxE based on just the "top hits" may lead to ascertainment biases, with respect to both the pervasiveness and the mode of GxE across the genome. Much like the heritability of complex traits is thought to be due to the contribution of many small (typically sub-significant) effects [24, 26], when GxE is pervasive we may expect that the sum of many small differences in context-specific effects accounts for the majority of GxE variation.

For concreteness, we consider in more depth one recent study characterizing GxDiet effects on longevity in *Drosophila melanogaster* [39]. In this study, Pallares et al. tracked caged fly populations given one of two diets: a "control" diet and a "high-sugar" diet. Across 271K single nucleotide variants, the authors tested for association between alleles and their survival to a sampling point (thought of as a proxy for "lifespan" or "longevity") under each diet independently. Then, they classified variants according to whether or not their associations with survivorship were significant under each diet as follows:

1. significant under neither diet → classify as *no effect*.
2. significant when fed the high-sugar diet, but not when fed the control diet → classify as *high-sugar specific effect*.

3. significant when fed the control diet, but not when fed the high-sugar diet → classify as *control specific effect*.
4. significant under both diets → classify as *shared effect*.

This authors' choice of four categories a variant may fall into may be motivated by the wish to test for the presence of "cryptic genetic variation"—genetic variation that is maintained in a context where it is functionally neutral but carries large effects in a new or stressful context [3, 5, 33, 40]. Indeed, of the variants Pallares et al. classified as having an effect (one hundredth of variants tested), approximately 31% were high-sugar specific, while the remaining 69% of the variants were shared. Fewer than 1% were labelled as having control specific effects. They concluded that high-sugar specific effects on longevity are pervasive, compatible with the hypothesis of widespread cryptic genetic variation for longevity.

This characterization of GxE, based on "top hits", places an emphasis on the context(s) in which trait associations are statistically significant, rather than on estimating how the context-specific effects covary. In addition, this particular classification system also does not cover all possible ways in which context-specific effects may differ. In the Supplementary Materials, we discuss these interpretation difficulties further.

We next show that a generative model that differs qualitatively from the cryptic genetic variation model yields results that are highly similar to those observed by Pallares et al. We simulated data under pervasive amplification. Specifically, we sampled from a mixture of 40% of variants having no effect under either diet and 60% of variants having an effect under both diets—but exactly 1.4× larger under a high-sugar diet. We then simulated the noisy estimation of these effects, and employed the classification approach of Pallares et al. to the simulated data (**Methods**).

The patterns of allelic effects in the control compared to high-sugar contexts were qualitatively similar in the experimental data and our pervasive amplification simulation. This is true both genome-wide (**Fig. 4A** compared to **Fig. 4B**) and for the set of variants classified as significant with their classification approach (**Fig. 4C** compared to **Fig. 4D**). The similarity of ascertained variants further highlights caveats of interpretation based on the classification of "top hits": despite the fact that we did not simulate any variants that only have an effect under the high-sugar diet, approximately 36% of significant variants were classified as specific to the high-sugar diet (green points in **Fig. 4D**), comparable to the 31% of variants classified as high-sugar specific in the experimental data (**Fig. 4C**). These variants simply have sub-significant associations in the control group and significant associations in the high-sugar group. In addition, every variant in the shared category (blue points in **Fig. 4D**) in fact has a larger effect in the high-sugar diet than in the control diet, which cannot be captured by the classification system itself but represents the only mode of GxE in our simulation.

To recap, we simulated a mode of GxE that is not considered in Pallares et al. (i.e., pervasive amplification) and that is at odds with their conclusions about evidence for a large discrete class of SNPs with diet-specific effects (i.e., cryptic genetic variation). The close match of our simulation to the empirical results of Pallares et al. therefore illustrates that the characterization of GxE via hypothesis testing and classification at each variant independently may lead to erroneous interpretation when applied to empirical complex trait data as well. In the **Supplementary Materials**, we show that a re-analysis of the Pallares et al. data that is based on estimating the covariance of allelic effects directly is consistent with pervasive amplification as well (**Fig. S4**). In conclusion, the classification of "top hits" alone may not be representative of the extent of GxE nor of the most pervasive modes of GxE.

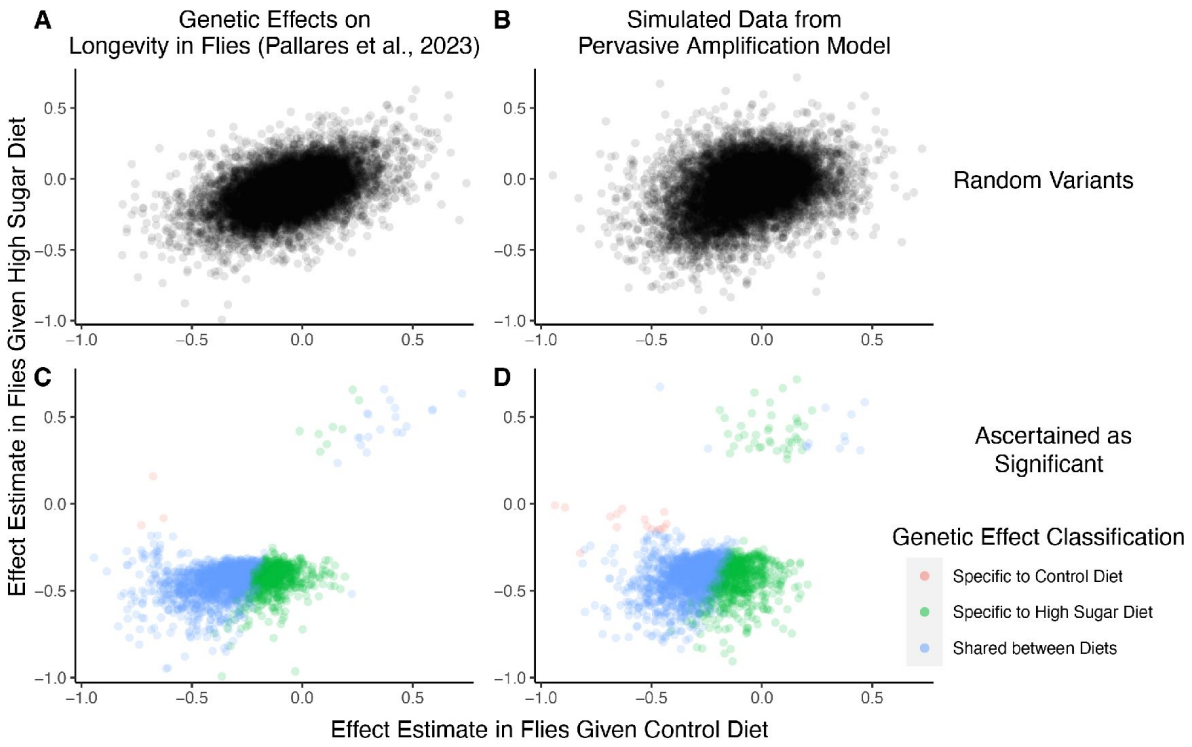


Figure 4

A focus on top hits may be lead to mischaracterization of polygenic GxE. **(A)** Data from an experiment measuring allelic effects on longevity in caged flies given one of two diets, "control" and "high sugar". Shown are allelic effect estimates under each diet for a random sample of approximately 12K variants. Simulated data where all true allelic effects are exactly 1.4 times larger under a high-sugar diet. The effects are estimated with sampling noise mimicking the Pallares et al. data. **(C)** Allelic effect estimates of variants ascertained as significant and classified as "diet-specific" or "shared" by Pallares et al. **(D)** Simulated effects ascertained as significant and classified using a similar procedure to that applied in (C). While the generative mode of GxE we used in our simulations was not considered by Pallares et al., the simulation results (left panels) closely match the patterns observed in their data (right panels) across all effects (top panels) and as reflected via their classification approach (bottom panels).

The utility of modeling GxE for complex trait prediction

Modeling context dependency of genetic effects may hold the potential for constructing polygenic scores that are more accurate, or improve their portability across contexts [34, 41–45]. Evidence for the utility of GxE models in polygenic score prediction, however, has been underwhelming and GxE models are still rarely applied [9, 10]. A key reason behind this apparent discrepancy is the bias–variance tradeoff for individual variants discussed above. If context-specific effects are similar—a likely possibility for highly polygenic traits with the majority of heritability owing to small causal effects—then additive models will tend to outperform [18, 19, 46, 47]. This is because the unbiasedness of GxE estimation does not make up for the cost of additional estimator variance, resulting from sample stratification by context or the addition of explicit interaction terms [10].

We exemplify the relative importance of variance compared to bias in polygenic scoring using simulations. We continue with the generative model of pervasive amplification as an example. Namely, we simulated a GWAS of a continuous trait with independent effects in 2, 500 variants (50% of variants included in the GWAS). Effects were either the same in two contexts, A and B, or 1.4 times larger in context B. The GWAS is conducted with either a small sample size or a large sample size, conferring low or high statistical power, respectively. We then constructed polygenic scores using 833 variants (corresponding to one-third of the causal variants), which were ascertained as most significantly associated with the trait according to either the additive model (orange and red in **Fig. 5**) in or context-specific hypothesis tests (green and blue in **Fig. 5**).

Even in settings with pervasive GxE, additive polygenic scores (red lines in **Fig. 5**) outperformed context-specific scores (green lines in **Fig. 5**). The advantage of the additive model is manifested in two ways: more accurate estimation, as discussed above, but also better identification of true associations with the trait. We considered the two advantages separately. It is sometimes better to ascertain variants using the lower variance approach and estimate effects using the lower-bias approach. In our simulations, this strategy (orange lines in **Fig. 5**) was preferable to using the GxE model for both ascertainment and estimation (green line). It was not preferable to using the additive model (red line) for both approaches; but it was the preferable strategy under a slightly different parametric regime, corresponding to more GxE (**Fig. S4B**).

Finally, we considered a polygenic GxE approach, as implemented in “multivariate adaptive shrinkage” (*mash*) [29], a method to estimate context-specific effects by leveraging common patterns of effect covariance between contexts observed across the genome. *mash* models the underlying distribution of effects in all contexts as a mixture of zero-centered Multivariate Normal distributions with different covariance structures (as well as the null matrix, to induce additional shrinkage). After estimating this distribution via maximum likelihood, *mash* uses it as a prior to obtain posterior effect estimates for each variant in each context. As a result, posterior effect estimates across contexts regress towards commonly observed patterns of covariance of allelic effects across contexts.

In our simulations, in the presence of substantial amplification, the polygenic adaptive shrinkage approach outperformed all other methods as long as the study was adequately powered (**Fig. 5B**). This is thanks to the unique ability (compared to the three other approaches) to leverage the sharing of signals across variants, including the extent and nature of context dependency. With low power, however, the additive model performed best (**Fig. 5A**). We attribute this to the variance cost associated with the polygenic adaptive shrinkage approach—driven by the estimation of additional parameters for capturing the genome-wide covariance relationships.

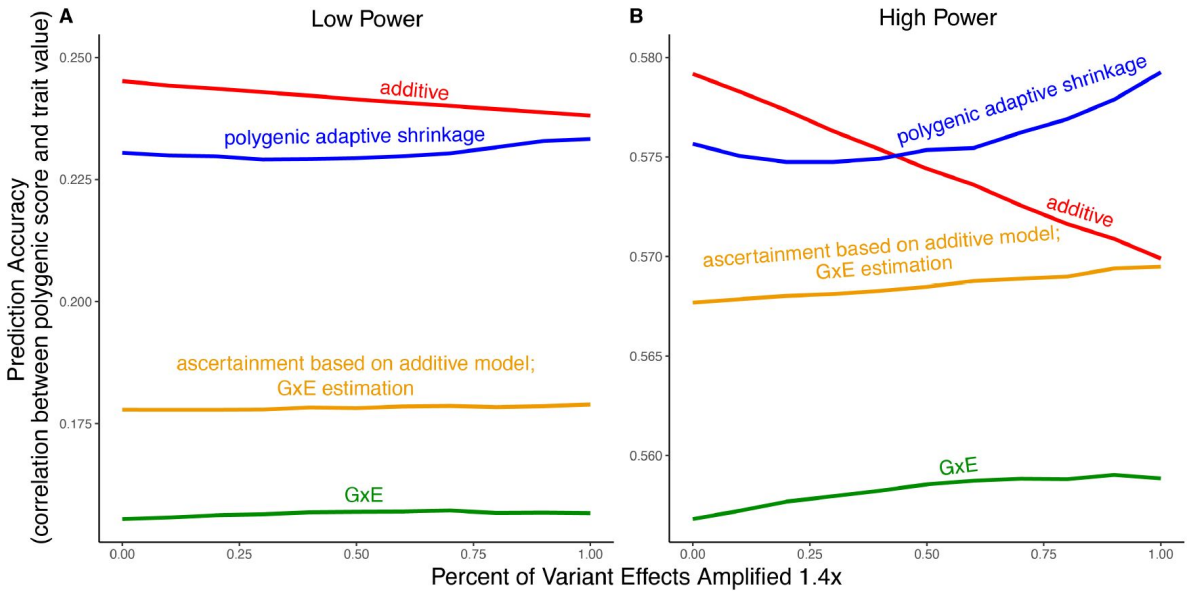


Figure 5

Polygenic score performance for context-dependent prediction models.

In each simulation, a GWAS is performed on 5, 000 biallelic variants, half of which have no effect in either context. Of the other half, some percent of the variants (indicated on the x-axis) had effects 1.4× larger in one of contexts and the remaining SNPs had equal effects in both contexts. The broad sense heritability was set to 0.4 in all simulations. The y-axis shows the average, over 11, 000 simulations, of the out-of-sample Pearson correlation between polygenic score and trait value. **(A)** Results with a GWAS sample size of 1, 000 individuals. **(B)** Results with a GWAS size of 50, 000 individuals.

Conclusion

When genetic variants are considered independently, the estimation of their effects in different contexts can be boiled down to a bias-variance tradeoff. For complex traits, we show through example that further considering polygenic patterns of GxE can be key for understanding context-dependent genetic architecture and to aid in prediction. The notion that complex trait analyses should combine observations at top associated loci alongside polygenic trends has gained traction with additive models of trait variation; it may be similarly important in our understanding of context-dependency.

Methods

Expressing the Additive Estimator as a Linear Combination of GxE Estimators

In this section, we prove the result of **Eq. 3**, stating that

$$\hat{\beta}_{A\cup B} = \omega_A \hat{\beta}_A + \omega_B \hat{\beta}_B$$

for some non-negative weights ω_A and ω_B. To do this, we will need some additional notation. Let $\bar{g}_A$ denote the average number of effect alleles in individuals in context A, and let $\bar{g}_{A\cup B}$ denote the average effect allele count across all individuals. Similarly, let $\bar{y}_A$ denote the average trait value in context A, and let $\bar{y}_{A\cup B}$ denote the average trait value across all individuals.

As an OLS estimator, the context-specific estimator is defined as

$$\begin{aligned}
\hat{\beta}_A &= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)(y_i - \bar{y}_A)}{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2} \\
&= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i - \sum_{i=1}^{n}(g_i - \bar{g}_A)\bar{y}_A}{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2} \\
&= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i - \bar{y}_A \sum_{i=1}^{n}(g_i - \bar{g}_A)}{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2} \\
&= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i}{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2}, \text{ since } \sum_{i=1}^{n}(g_i - \bar{g}_A) = 0.
\end{aligned}$$

Similarly, the additive estimator can be written as:

$$\begin{aligned}\hat{\beta}_{A\cup B} &= \frac{\sum_{i=1}^{n+m}(g_i-\bar{g}_{A\cup B})(y_i-\bar{y}_{A\cup B})}{\sum_{i=1}^{n+m}(g_i-\bar{g}_{A\cup B})^2}\\ &= \frac{\sum_{i=1}^{n+m}(g_i-\bar{g}_{A\cup B})y_i}{\sum_{i=1}^{n+m}(g_i-\bar{g}_{A\cup B})^2} \text{ (by the same logic as above)}\\ &= \frac{\sum_{i=1}^{n}(g_i-\bar{g}_{A\cup B})y_i+\sum_{i=n+1}^{n+m}(g_i-\bar{g}_{A\cup B})y_i}{\sum_{i=1}^{n+m}(g_i-\bar{g}_{A\cup B})^2}.\end{aligned}$$

We will show that the weights in **Eq. 3** depend on the effect allele frequency in the two contexts, f_A and f_B. We will assume mean-centered traits, such that $\sum_{i=1}^{n} y_i = 0$ and $\sum_{i=n+1}^{n+m} y_i = 0$ We note that mean-centering is inconsequential for effect estimation. We can then write

$$\hat{\beta}_{A\cup B} = \frac{\sum_{i=1}^{n}(g_i - \bar{g}_{A\cup B})y_i + \sum_{i=n+1}^{n+m}(g_i - \bar{g}_{A\cup B})y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_{A\cup B})y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_{A\cup B})y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A + (\bar{g}_A - \bar{g}_{A\cup B}))y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B + (\bar{g}_B - \bar{g}_{A\cup B}))y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i + \sum_{i=1}^{n}(\bar{g}_A - \bar{g}_{A\cup B})y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)y_i + \sum_{i=n+1}^{n+m}(\bar{g}_B - \bar{g}_{A\cup B})y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i + (\bar{g}_A - \bar{g}_{A\cup B})\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)y_i + (\bar{g}_B - \bar{g}_{A\cup B})\sum_{i=n+1}^{n+m} y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} \text{ (by our assumption of mean centered traits)}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} \cdot \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2}{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2} + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)y_i}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} \cdot \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)^2}{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)^2}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)y_i}{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2} \cdot \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2} + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)y_i}{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)^2} \cdot \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)^2}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$$

$$= \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}\hat{\beta}_A + \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)^2}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}\hat{\beta}_B.$$

Thus, $\omega_A = \frac{\sum_{i=1}^{n}(g_i - \bar{g}_A)^2}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$ and $\omega_B = \frac{\sum_{i=n+1}^{n+m}(g_i - \bar{g}_B)^2}{\sum_{i=1}^{n+m}(g_i - \bar{g}_{A\cup B})^2}$ in **Eq. 3**. We note that the numerator of ω_A is n times the sample heterozygozity in context A, and the numerator of ω_B is m times the sample heterozygozity in context B. Thus, we have shown that

$$\omega_A \propto nH_A \text{ and } \omega_B \propto nH_B, \tag{7}$$

where H_A and H_B are the sample heterozygozities in context A and B, respectively. And, in the special case where $f_A = f_B$, because this implies that the sample heterozygozities will be approximately equal across contexts, we have that

$$\omega_A = \frac{n}{n+m} \text{ and } \omega_B = \frac{m}{m+n}. \tag{8}$$

Linearity of the decision rule

In **Eq. 5**, under the assumption that $V_A = V_B$ and $H_A = H_B$, the decision boundary is expressed as a linear function of $|\beta_A - \beta_B|$ and $\sqrt{V_A}$ as

$$\sqrt{\frac{m}{2n}}|\beta_A - \beta_B| > \sqrt{V_A}.$$

Here, we prove that the linearity of the decision rule holds in the more general case where $\frac{V_A}{V_B} = r$ for some fixed value of r. **Eq. 5** then follows as a special case of this fact when $r = 1$.

Starting from **Eq. 4**, we prefer the GxE estimator to the additive estimator when estimating β_A if

$$\begin{aligned}
&V_A < \omega_A^2 V_A + \omega_B^2 V_B + ((\omega_A - 1)\beta_A + \omega_B\beta_B)^2 \\
\iff &V_A < \omega_A^2 V_A + \frac{\omega_B^2}{r} V_A + ((\omega_A - 1)\beta_A + \omega_B\beta_B)^2 \\
\iff &V_A - \omega_A^2 V_A - \frac{\omega_B^2}{r} V_A < ((\omega_A - 1)\beta_A + \omega_B\beta_B)^2 \\
\iff &(1 - \omega_A^2 - \frac{\omega_B^2}{r})V_A < ((\omega_A - 1)\beta_A + \omega_B\beta_B)^2 \\
\iff &V_A < \frac{((\omega_A - 1)\beta_A + \omega_B\beta_B)^2}{1 - \omega_A^2 - \frac{\omega_B^2}{r}} \text{ (assuming } 1 - \omega_A^2 - \frac{\omega_B^2}{r} > 0) \\
\iff &\sqrt{V_A} < \frac{|(\omega_A - 1)\beta_A + \omega_B\beta_B|}{\sqrt{1 - \omega_A^2 - \frac{\omega_B^2}{r}}} \text{ (again assuming } 1 - \omega_A^2 - \frac{\omega_B^2}{r} > 0)
\end{aligned}$$

If our assumption that $1 - \omega_A^2 - \frac{\omega_B^2}{r} > 0$ does not hold, we note that the GxE model is always preferable and technically speaking there exists no decision rule between the two models. Now, when heterozygozities (and thus minor allele frequencies) are equal across contexts, then **Eq. 8** implies $\omega_A + \omega_B = 1$. Therefore, we may write the decision rule as

$$\begin{aligned}
&\sqrt{V_A} < \frac{|(1 - \omega_B - 1)\beta_A + \omega_B\beta_B|}{\sqrt{1 - \omega_A^2 - \frac{\omega_B^2}{r}}} \\
\iff &\sqrt{V_A} < \frac{|\omega_B(\beta_B - \beta_A)|}{\sqrt{1 - \omega_A^2 - \frac{\omega_B^2}{r}}} \\
\iff &\sqrt{V_A} < \frac{\omega_B}{\sqrt{1 - \omega_A^2 - \frac{\omega_B^2}{r}}}|\beta_A - \beta_B| \text{ (by properties of the absolute value)} \\
\iff &\sqrt{V_A} < \frac{1 - \omega_A}{\sqrt{1 - \omega_A^2 - \frac{(1-\omega_A)^2}{r}}}|\beta_A - \beta_B|.
\end{aligned}$$

Here, we see that for any fixed r the decision rule is linear with a slope determined by r (**Fig. 2**). Now, in the special case where $r = 1$, we have

$$\begin{aligned}
&\sqrt{V_A} < \frac{1-\omega_A}{\sqrt{1-\omega_A^2-(1-\omega_A)^2}}|\beta_A-\beta_B| \\
\iff &\sqrt{V_A} < \frac{1-\omega_A}{\sqrt{1-\omega_A^2-1-\omega_A^2+2\omega_A}}|\beta_A-\beta_B| \\
\iff &\sqrt{V_A} < \frac{1-\omega_A}{\sqrt{2\omega_A(1-\omega_A)}}|\beta_A-\beta_B| \\
\iff &\sqrt{V_A} < \sqrt{\frac{1-\omega_A}{2\omega_A}}|\beta_A-\beta_B|
\end{aligned}$$

Now, substituting the definitions of ω_A and ω_B in the case of equal minor allele frequencies given in **Eq. 8**, we can write

$$\begin{aligned}
&\sqrt{V_A} < \sqrt{\frac{1}{2}}\sqrt{\frac{1-\frac{n}{n+m}}{\frac{n}{n+m}}}|\beta_A-\beta_B| \\
\iff &\sqrt{V_A} < \sqrt{\frac{1}{2}}\sqrt{\frac{\frac{m}{n+m}}{\frac{n}{n+m}}}|\beta_A-\beta_B| \\
\iff &\sqrt{V_A} < \sqrt{\frac{m}{2n}}|\beta_A-\beta_B|.
\end{aligned}$$

This inequality is instead an equality under the conditions stated in **Eq. 5**. Finally, again using the definition of ω_A and ω_B given in **Eq. 8**, we note that our assumption that $1-\omega_A^2-\frac{\omega_B^2}{r} > 0$ will always hold in the case of equal minor allele frequencies and $r = 1$, as

$$\begin{aligned}
1-\omega_A^2-\frac{\omega_B^2}{r} &= 1-\frac{n^2}{(n+m)^2}-\frac{m^2}{(n+m)^2} \\
&= \frac{(n+m)^2-n^2-m^2}{(n+m)^2} \\
&= \frac{2nm}{(n+m)^2},
\end{aligned}$$

which is strictly positive.

Re-parameterized decision rule in terms of unitless quantities

In **Eq. 6**, under the assumption that $H_A = H_B$, we re-state the decision rule in terms of the signal-to-noise ratio. Here, we prove this result.

From **Eq. 4**, we have that we should select the GxE model to estimate β_A if and only if

$$\begin{aligned}
&V_A < \omega_A^2 V_A + \omega_B^2 V_B + \left((\omega_A-1)\beta_A+\omega_B\beta_B\right)^2 \\
\iff &1 < \omega_A^2 + \omega_B^2\frac{V_B}{V_A} + \frac{\left((\omega_A-1)\beta_A+\omega_B\beta_B\right)^2}{V_A} \\
\iff &1-\omega_A^2 < \omega_B^2\frac{V_B}{V_A} + \frac{\left((\omega_A-1)\beta_A+\omega_B\beta_B\right)^2}{V_A}.
\end{aligned}$$

Now, because $H_A = H_B$, we know by **Eq. 8** that $\omega_A + \omega_B = 1$. Then, we may write the decision rule as

$$
\begin{aligned}
& 1-\omega_A^2 < \omega_B^2 \frac{V_B}{V_A} + \frac{\left((1-\omega_B-1)\beta_A + \omega_B\beta_B\right)^2}{V_A} \\
\iff & 1-\omega_A^2 < \omega_B^2 \frac{V_B}{V_A} + \frac{\left(\omega_B(\beta_B-\beta_A)\right)^2}{V_A} \\
\iff & 1-\omega_A^2 < \omega_B^2 \frac{V_B}{V_A} + \omega_B^2 \frac{(\beta_A-\beta_B)^2}{V_A} \\
\iff & \frac{1-\omega_A^2}{\omega_B^2} < \frac{V_B}{V_A} + \frac{(\beta_A-\beta_B)^2}{V_A} \\
\iff & \frac{1-\omega_A^2}{(1-\omega_A)^2} < \frac{V_B}{V_A} + \frac{(\beta_A-\beta_B)^2}{V_A} \\
\iff & \frac{1-\omega_A^2}{(1-\omega_A)^2} - \frac{V_B}{V_A} < \frac{(\beta_A-\beta_B)^2}{V_A} \\
\iff & \frac{(1-\omega_A)(1+\omega_A)}{(1-\omega_A)^2} - \frac{V_B}{V_A} < \frac{(\beta_A-\beta_B)^2}{V_A} \\
\iff & \frac{1+\omega_A}{1-\omega_A} - \frac{V_B}{V_A} < \frac{(\beta_A-\beta_B)^2}{V_A}
\end{aligned}
$$

as is stated in **Eq. 6**.

Simulation of GxDiet effects on longevity in Drosophila

In **Fig. 4**, we compare the effect estimates of Pallares et al. to ones we got in simulations of pervasive amplification. Here, we detail the simulation approach.

We first generated true effects under each diet. For variants j = 1, ..., 50, 000, we sampled a true effect under the high-sugar diet (β_{h_j}) and under the control diet (β_{c_j}). A random 60% of variants were set to have no effect under either diet, with the effects of the remaining 40% of variants sampled as

$$
\begin{bmatrix} \beta_{c_j} \\ \beta_{h_j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} -0.125 \\ -0.15 \end{bmatrix}, 0.01 \cdot \begin{bmatrix} 1 & 1.4 \\ 1.4 & 1.96 \end{bmatrix} \right).
$$

This corresponds to a systematic amplification of 1.4× in the high-sugar compared to the control diet. We selected these parameters based on inspection of the resulting distribution of effects and their correspondence to the Pallares et al. data.

We then simulated the effect estimation. For each variant, the effect estimate was simulated as Normally distributed with mean equal to the true effect and standard deviation equal to a randomly sampled (with replacement) standard error from the effect estimates of Pallares et al. That is, given the simulated values of the true effect estimates β_{c_j} and β_{h_j}, we simulated effect estimates as

$$
\begin{bmatrix} \hat{\beta}_{c_j} \\ \hat{\beta}_{h_j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta_{c_j} \\ \beta_{h_j} \end{bmatrix}, \begin{bmatrix} \hat{s}_{c_k}^2 & 0 \\ 0 & \hat{s}_{h_k}^2 \end{bmatrix} \right),
$$

where k represents the index of a randomly selected variant from the empirical data of Pallares et al. and $\hat{s}_{c_k}$ and $\hat{s}_{h_k}$ are the corresponding estimated standard errors for the effect estimates in the control and high-sugar groups, respectively. This process yielded vectors of estimated effects in the

high-sugar group and control group, $\hat{\beta}_h$ and $\hat{\beta}_c$ respectively, and vectors of estimated standard errors in the high-sugar group and control group, $\hat{\boldsymbol{s}}_h$ and $\hat{\boldsymbol{s}}_c$ respectively. We then performed a Z-test for each variant under each diet, yielding two vectors of p-values, $\boldsymbol{p}_h$ and $\boldsymbol{p}_c$, corresponding to the high-sugar and control diets, respectively.

Using these p-values, we followed a similar approach to Pallares et al. to classify the variants (**Fig. 4D**). First, as in Pallares et al., we computed q-values separately for each diet [48], yielding $\boldsymbol{q}_h$ and $\boldsymbol{q}_c$, corresponding to the q-values of non-zero effects in the high-sugar and control diets, respectively. Then, we employed the following classification scheme for each variant j = 1, ..., 50, 000:

1. if $q_{h_j} \geq .01$ and $q_{c_j} \geq .01 \rightarrow$ classify as *no effect*.
2. if $q_{h_j} < .01$ and $p_{c_j} \geq .1 \rightarrow$ classify as *high-sugar specific effect*.
3. if $q_{c_j} < .01$ and $p_{h_j} \geq .1 \rightarrow$ classify as *control specific effect*.
4. if $q_{c_j} < .01$ and $q_{h_j} < .01 \rightarrow$ classify as *shared effect*.

We note that p-value and q-value cutoffs used are nominally different than those used in the Pallares et al. study.

Polygenic Score Simulations

In **Fig. 5**, we show the results of multiple simulations where we compute polygenic scores in each of two contexts under amplification. Here, we detail the generation of data in the simulations and the methods for constructing polygenic scores.

As in **Results and Discussion**, we assumed that we have n + m observations of a continuous trait, where the first n individuals are observed in context A and the final m are observed in context B. For convenience, in this case we assumed n = m. Now, for variants j = 1, ..., p we generated true effects in contexts A and B independently from the mixture model

$$\begin{bmatrix}\beta_{A_j}\\ \beta_{B_j}\end{bmatrix} \sim \pi_0\delta_0 + (1-\pi_0)\left(\alpha\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\right) + (1-\alpha)\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\frac{3}{2} & 1\\ 1 & \frac{2}{3}\end{bmatrix}\right)\right),$$

where π_0 (which we set to 0.5) represents the proportion of SNPs with null effects in both contexts, α represents the proportion of non-null SNPs which have exactly equal effects in both contexts, and 1 − α is the proportion of non-null SNPs which are generated as perfectly correlated but with 1.5× the standard deviation in context A. Let $\vec{\beta}_A$ and $\vec{\beta}_B$ represent the resulting p-vectors of true effects for contexts *A* and *B*, respectively.

Next, we generated genotype counts for each of the n + m individuals at all p variants. Specifically, we independently generated genotypes as

$$f_j \sim \frac{1}{2}Beta(s_1, s_2) \text{ for } j = 1, \ldots, p$$
$$g_{ij} \sim Binomial(2, f_j) \text{ for } i = 1, \ldots, n+m,$$

where f_j is the minor allele frequency at variant j in the population, s_1 and s_2 are parameters controlling the distribution of minor allele freqeuncies in the population, and g_{ij} is the observed genotype for individual i at variant j. Here, we set s_1 = 1 and s_2 = 5. Let $\boldsymbol{G}_A$ and $\boldsymbol{G}_B$ represent the generated $n \times p$ matrices of genotypes in contexts A and B, respectively.

Finally, we generated the observed continuous traits for context A ($\vec{y}_A$) and context B ($\vec{y}_B$) as

$$\vec{y}_A \sim \mathcal{N}(\boldsymbol{G}_A\vec{\beta}_A, \sigma_A^2 I_n)$$
$$\vec{y}_B \sim \mathcal{N}(\boldsymbol{G}_B\vec{\beta}_B, \sigma_B^2 I_m),$$

where σ_A^2 and σ_B^2 are the observation variances in contexts A and B, respectively, and I_w is the $w \times w$ identity matrix. In our simulations, we set σ_A^2 and σ_B^2 such that the narrow sense heritability is 40% in each context. So that we may later test the accuracy of our polygenic scores, we generated both a training set (consisting of n individuals in each context, where n = 1, 000 in the low power simulation and n = 50, 000 in the high power simulation) for effect estimation and a test set (consisting of 3, 000 individuals in each context) using the above distributions.

Fig. 5 compares four distinct approaches for constructing polygenic scores, derived from three allelic effect estimation approaches: additive estimation with shrinkage, GxE estimation with shrinkage and *mash*. First, the additive and GxE estimates are derived independently for each variant as described in **Results and Discussion**. Let $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ be the p-vectors of GxE estimates of effects in context A and B, respectively. Similarly, let $\hat{\boldsymbol{s}}_A$ and $\hat{\boldsymbol{s}}_B$ be the p-vectors of the standard errors of GxE estimates of effects in context A and B, respectively. Finally, let $\hat{\boldsymbol{\beta}}_{A\cup B}$ be the p-vector of estimated effects from the additive model and be the p-vector of standard errors of estimated effects from the additive model. Using the GxE estimates, we also constructed estimates of the effects in each context using *mash*. Specifically, we ran mash on the n × 2 matrices $\left[\hat{\boldsymbol{\beta}}_A \quad \hat{\boldsymbol{\beta}}_B\right]$ (of effects) and $\left[\hat{\boldsymbol{s}}_A \quad \hat{\boldsymbol{s}}_B\right]$ (of standard errors). *mash* then yields $p(\vec{\beta}_A|\hat{\boldsymbol{\beta}}_A, \hat{\boldsymbol{s}}_A)$ and $p(\vec{\beta}_B|\hat{\boldsymbol{\beta}}_B, \hat{\boldsymbol{s}}_B)$, the posterior distributions of the effects in contexts A and B, respectively.

To construct each polygenic score, we made two choices. First, a choice between the three sets of p-values (or pseudo p-values, see below) for thresholding—we include the 833 (corresponding to one-third of the causal variants) most significant variants in the polygenic score. The second choice was between the three sets of effect estimates to be used as weights in the polygenic score (**Fig. 5**). For instance, when the GxE model was used for ascertainment, we selected the set of variants $\Omega_A \subset \{1, ..., p\}$ consisting of the variants with the 833 smallest p-values and $\Omega_B \subset \{1, ..., p\}$ consisting of the variants with the 833 smallest p-values (derived from $\hat{\boldsymbol{\beta}}_B$ and $\hat{\boldsymbol{s}}_B$). Then, we predicted trait values (out of sample) by multiplying the effect estimates of our chosen "estimation method" (for *mash* we use the posterior mean) by the effect allele count at each of the selected variants for the individual in question.

Acknowledgements

We thank Doc Edge, Marc Feldman, Mark Kirkpatrick, Molly Przeworski, Anil Raj, Elliot Tucker-Drob and members of the Harpak Lab for comments on the manuscript. We thank Peter Andolfatto, Julien Ayroles and Tom Juenger for helpful discussions. All authors were supported by NIH R35GM151108 to A.H. S.P. Smith was also supported by NIH RF1AG073593. This study was conducted using the UK Biobank resource under application 61666, as approved by the University of Texas at Austin institutional review board (protocol 2019-02-0125).

Additional files

Supplementary Materials

References

1. El-Soda M., Malosetti M., Zwaan B. J., Koornneef M., Aarts M. G. 2014) **Genotype× environment interaction QTL mapping in plants: lessons from Arabidopsis** *Trends in plant science* **19**:390–398

2. Vieira C., et al. 2000) **Genotype-environment interaction for quantitative trait loci affecting life span in Drosophila melanogaster** *Genetics* **154**:213–227

3. Des Marais D. L., Hernandez K. M., Juenger T. E. 2013) **Genotype-by-environment interaction and plasticity: exploring genomic responses of plants to the abiotic environment** *Annual Review of Ecology, Evolution, and Systematics* **44**:5–29

4. Smith E. N., Kruglyak L. 2008) **Gene–environment interaction in yeast gene expression** *PLoS biology* **6**:e83

5. Paaby A. B., Rockman M. V. 2014) **Cryptic genetic variation: evolution's hidden substrate** *Nature Reviews Genetics* **15**:247–258

6. Munafó M. R., Zammit S., Flint J. 2014) **Practitioner Review: A critical perspective on gene–environment interaction models–what impact should they have on clinical perceptions and practice?** *Journal of child Psychology and Psychiatry* **55**:1092–1101

7. Kraft P., Aschard H. 2015) **Finding the missing gene–environment interactions** *European Journal of Epidemiology* **30**:353–355

8. Sella G., Barton N. H. 2019) **Thinking about the evolution of complex traits in the era of genome-wide association studies** *Annual Review of Genomics and Human Genetics* **20**:461–493

9. Zhu C., et al. 2023) **Amplification is the primary mode of gene-by-sex interaction in complex human traits** *Cell Genomics*

10. Schwaba T., et al. 2023) **Comparison of the Multivariate Genetic Architecture of Eight Major Psychiatric Disorders Across Sex** *medRxiv*

11. Elgart M., et al. 2022) **Correlations between complex human phenotypes vary by genetic background, gender, and environment** *Cell Reports Medicine* **3**:100844

12. Duncan L. E., Keller M. C. 2011) **A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry** *American Journal of Psychiatry* **168**:1041–1049

13. Gibson G., Lacek K. A. 2020) **Canalization and robustness in human genetics and disease** *Annual Review of Genetics* **54**:189–211

14. Brown B. C., Ye C. J., Price A. L., Zaitlen N., Consortium, A. G. E. N. T. 2. D., et al. 2016) **Transethnic genetic-correlation estimates from summary statistics** *The American Journal of Human Genetics* **99**:76–88

15. Ge T., Chen C.-Y., Neale B. M., Sabuncu M. R., Smoller J. W. 2017) **Phenome-wide heritability analysis of the UK Biobank** *PLoS Genetics* **13**:e1006711

16. Balliu B., et al. 2021) **An integrated approach to identify environmental modulators of genetic risk factors for complex traits** *The American Journal of Human Genetics* **108**:1866–1879

17. Veller C., Przeworski M., Coop G. 2023) **Causal interpretations of family GWAS in the presence of heterogeneous effects** *bioRxiv* :2023–11

18. Fisher R. A. 1930) **The genetical theory of natural selection** Clarendon Press

19. Falconer D. S., Mackay T. F. 1996) **Introduction to quantitative genetics**

20. Yengo L., et al. 2022) **A saturated map of common genetic variants associated with human height** *Nature* **610**:704–712

21. Zwick M. E., Cutler D. J., Chakravarti A. 2000) **Patterns of genetic variation in Mendelian and complex traits** *Annual Review of Genomics and Human Genetics* **1**:387–407

22. Casella G., Berger R. L. 2021) **Statistical Inference**

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

24. Sinnott-Armstrong N., Naqvi S., Rivas M., Pritchard J. K. 2021) **GWAS of three molecular traits high-lights core genes and pathways alongside a highly polygenic background** *eLife* **10**:e58615

25. Shi H., Kichaev G., Pasaniuc B. 2016) **Contrasting the genetic architecture of 30 complex traits from summary association data** *The American Journal of Human Genetics* **99**:139–153

26. Boyle E. A., Li Y. I., Pritchard J. K. 2017) **An expanded view of complex traits: from polygenic to omnigenic** *Cell* **169**:1177–1186

27. Liu X., Li Y. I., Pritchard J. K. 2019) **Trans effects on gene expression can drive omnigenic inheritance** *Cell* **177**:1022–1034

28. Wray N. R., Wijmenga C., Sullivan P. F., Yang J., Visscher P. M. 2018) **Common disease is more complex than implied by the core gene omnigenic model** *Cell* **173**:1573–1580

29. Urbut S. M., Wang G., Carbonetto P., Stephens M. 2019) **Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions** *Nature Genetics* **51**:187–195

30. Zhang L., et al. 2021) **QTL× environment interactions underlie ionome divergence in switchgrass** *G3* **11**:jkab144

31. Paaby A. B., Gibson G. 2016) **Cryptic genetic variation in evolutionary developmental genetics** *Biology* **5**:28

32. Aschard H., et al. 2017) **Evidence for large-scale gene-by-smoking interaction effects on pulmonary function** *International journal of epidemiology* **46**:894–904

33. Gibson G., Dworkin I. 2004) **Uncovering cryptic genetic variation** *Nature Reviews Genetics* **5**:681–690

34. Miao J., et al. 2022) **Reimagining gene-environment interaction analysis for human complex traits** *bioRxiv* :2022–12

35. Spierer A. N., et al. 2021) **Natural variation in the regulation of neurodevelopmental genes modifies flight performance in Drosophila** *PLoS Genetics* **17**:e1008887

36. Traglia M., Bout M., Weiss L. A. 2022) **Sex-heterogeneous SNPs disproportionately influence gene expression and health** *PLoS Genetics* **18**:e1010147

37. Bernabeu E., et al. 2021) **Sex differences in genetic architecture in the UK Biobank** *Nature Genetics* **53**:1283–1289

38. Wright K. M., et al. 2022) **Age and diet shape the genetic architecture of body weight in diversity outbred mice** *eLife* **11**:e64329

39. Pallares L. F., et al. 2022) **Dietary stress remodels the genetic architecture of lifespan variation in outbred Drosophila** *Nature Genetics* :1–7

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

41. Mostafavi H., et al. 2020) **Variable prediction accuracy of polygenic scores within an ancestry group** *eLife* **9**:e48376

42. Patel R. A., et al. 2022) **Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits** *The American Journal of Human Genetics* **109**:1286–1297

43. Turley P., et al. 2018) **Multi-trait analysis of genome-wide association summary statistics using MTAG** *Nature Genetics* **50**:229–237

44. Spence J. P., Sinnott-Armstrong N., Assimes T., Pritchard J. K. 2022) **A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics** *bioRxiv* :2022–4

45. Wang J. Y., et al. 2024) **Three Open Questions in Polygenic Score Portability** *bioRxiv*

46. Hill W. G., Goddard M. E., Visscher P. M. 2008) **Data and theory point to mainly additive genetic variance for complex traits** *PLoS Genetics* **4**:e1000008

47. Young A. I. 2019) **Solving the missing heritability problem** *PLoS Genetics* **15**:e1008222

48. Storey J. D. 2003) **The positive false discovery rate: a Bayesian interpretation and the q-value** *The Annals of Statistics* **31**:2013–2035

Author information

Eric Weine

Department of Integrative Biology, The University of Texas at Austin, Austin, United States, Department of Population Health, The University of Texas at Austin, Austin, United States, Department of Human Genetics, University of Chicago, Chicago, United States

Samuel Pattillo Smith
Department of Integrative Biology, The University of Texas at Austin, Austin, United States, Department of Population Health, The University of Texas at Austin, Austin, United States
ORCID iD: 0000-0002-6269-0276

Rebecca Kathryn Knowlton
Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, United States

Arbel Harpak
Department of Integrative Biology, The University of Texas at Austin, Austin, United States, Department of Population Health, The University of Texas at Austin, Austin, United States
ORCID iD: 0000-0002-3655-748X

For correspondence: arbelharpak@utexas.edu

Editors

Reviewing Editor
George Perry
Pennsylvania State University, University Park, United States of America

Senior Editor
George Perry
Pennsylvania State University, University Park, United States of America

Reviewer #1 (Public review):

Experiments in model organisms have revealed that the effects of genes on heritable traits are often mediated by environmental factors -- so-called gene-by-environment (or GxE) interactions. In human genetics, however, where indirect statistical approaches must be taken to detect GxE, limited evidence has been found for pervasive GxE interactions. The present manuscript argues that the failure of statistical methods to detect GxE may be due to how GxE is modelled (or not modelled) by these methods.

The authors show, via re-analysis of an existing dataset in Drosophila, that a polygenic 'amplification' model can parsimoniously explain patterns of differential genetic effects across environments. (Work from the same lab had previously shown that the amplification model is consistent with differential genetic effects across the sexes for a number of traits in humans.) The parsimony of the amplification model allows for powerful detection of GxE in scenarios in which it pertains, as the authors show via simulation.

Before the authors consider polygenic models of GxE, however, they present a very clear analysis of a related question around GxE: When one wants to estimate the effect of an individual allele in a particular environment, when is it better to stratify one's sample by environment (reducing sample size, and therefore increasing the variance of the estimator) versus using the entire sample (including individuals not in the environment of interest, and therefore biasing the estimator away from the true effect specific to the environment of interest)? Intuitively, the sample-size cost of stratification is worth paying if true allelic effects differ substantially between the environment of interest and other environments (i.e., GxE interactions are large), but not worth paying if effects are similar across environments. The authors quantify this trade-off in a way that is both mathematically precise and conveys the

above intuition very clearly. They argue on its basis that, when allelic effects are small (as in highly polygenic traits), single-locus tests for GxE may be substantially underpowered.

The paper is an important further demonstration of the plausibility of the amplification model of GxE, which, given its parsimony, holds substantial promise for the detection and characterization of GxE in genomic datasets. However, the empirical and simulation examples considered in the paper (and previous work from the same lab) are somewhat "best-case" scenarios for the amplification model, with only two environments and with these environments amplifying equally the effects of only a single set of genes. It would be an important step forward to demonstrate the possibility of detecting amplification in more complex scenarios, with multiple environments each differentially modulating the effects of multiple sets of genes. This could be achieved via simulations similar to those presented in the current manuscript.

Comments on revisions:

The authors have (with reasonable justification) said that my main recommendations for strengthening the conclusions of the paper are beyond its scope, and they have thoughtfully responded to my (and the other reviewer's) other comments. The paper is now more clearly written---in particular, the connection between the single-locus bias-variance tradeoff calculations and the polygenic results is much more transparent than before. Given that the authors have (again, with fair justification) chosen not to address my major comment, my broad assessment of the paper is unchanged---I think it is an important contribution to a critical topic---and I have no further comments for its improvement (though I note an issue with figure referencing in the captions of Supplementary Figs S2 and S3).

https://doi.org/10.7554/eLife.99210.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Experiments in model organisms have revealed that the effects of genes on heritable traits are often mediated by environmental factors---so-called gene-by-environment (or GxE) interactions. In human genetics, however, where indirect statistical approaches must be taken to detect GxE, limited evidence has been found for pervasive GxE interactions. The present manuscript argues that the failure of statistical methods to detect GxE may be due to how GxE is modelled (or not modelled) by these methods.*
>
> *The authors show, via re-analysis of an existing dataset in Drosophila, that a polygenic 'amplification' model can parsimoniously explain patterns of differential genetic effects across environments. (Work from the same lab had previously shown that the amplification model is consistent with differential genetic effects across the sexes for several traits in humans.) The parsimony of the amplification model allows for powerful detection of GxE in scenarios in which it pertains, as the authors show via simulation.*
>
> *Before the authors consider polygenic models of GxE, however, they present a very clear analysis of a related question around GxE: When one wants to estimate the effect of an individual allele in a particular environment, when is it better to stratify one's sample by environment (reducing sample size, and therefore increasing the variance of the estimator) versus using the entire sample (including individuals not in the environment of interest, and therefore biasing the estimator away from the true effect specific to the*

environment of interest)? Intuitively, the sample-size cost of stratification is worth paying if true allelic effects differ substantially between the environment of interest and other environments (i.e., GxE interactions are large), but not worth paying if effects are similar across environments. The authors quantify this trade-off in a way that is both mathematically precise and conveys the above intuition very clearly. They argue on its basis that, when allelic effects are small (as in highly polygenic traits), single-locus tests for GxE may be substantially underpowered.

The paper is an important further demonstration of the plausibility of the amplification model of GxE, which, given its parsimony, holds substantial promise for the detection and characterization of GxE in genomic datasets. However, the empirical and simulation examples considered in the paper (and previous work from the same lab) are somewhat "best-case" scenarios for the amplification model, with only two environments, and with these environments amplifying equally the effects of only a single set of genes. It would be an important step forward to demonstrate the possibility of detecting amplification in more complex scenarios, with multiple environments each differentially modulating the effects of multiple sets of genes. This could be achieved via simulations similar to those presented in the current manuscript.

Reviewer #2 (Public Review):

Summary:

Wine et al. describe a framework to view the estimation of gene-context interaction analysis through the lens of bias-variance tradeoff. They show that, depending on trait variance and context-specific effect sizes, effect estimates may be estimated more accurately in context-combined analysis rather than in context-specific analysis. They proceed by investigating, primarily via simulations, implications for the study or utilization of gene-context interaction, for testing and prediction, in traits with polygenic architecture. First, the authors describe an assessment of the identification of context-specificity (or context differences) focusing on "top hits" from association analyses. Next, they describe an assessment of polygenic scores (PGSs) that account for context-specific effect sizes, showing, in simulations, that often the PGSs that do not attempt to estimate context-specific effect sizes have superior prediction performance. An exception is a PGS approach that utilizes information across contexts. Strengths:

The bias-variance tradeoff framing of GxE is useful, interesting, and rigorous. The PGS analysis under pervasive amplification is also interesting and demonstrates the bias-variance tradeoff.

Weaknesses:

The weakness of this paper is that the first part -- the bias-variance tradeoff analysis -- is not tightly connected to, i.e. not sufficiently informing, the later parts, that focus on polygenic architecture. For example, the analysis of "top hits" focuses on the question of testing, rather than estimation, and testing was not discussed within the bias-variance tradeoff framework. Similarly, while the PGS analysis does demonstrate (well) the bias-variance tradeoff, the reader is left to wonder whether a bias-variance deviation rule (discussed in the first part of the manuscript) should or could be utilized for PGS construction.

We thank the editors and the reviewers for their thoughtful critique and helpful suggestions throughout. In our revision, we focused on tightening the relationship between the analytical single variant bias-variance tradeoff derivation and the various empirical analyses that follow.

We improved discussion of our scope and what is beyond our scope. For example, our language was insufficiently clear if it suggested to the editor and reviewers that we are developing a method to characterize polygenic GxE. Developing a new method that does so (let alone evaluating performance across various scenarios) is beyond the scope of this manuscript.

Similarly, we clarify that we use amplification only as an example of a mode of GxE that is not adequately characterized by current approaches. We do not wish to argue it is an omnibus explanation for all GxE in complex traits. In many cases, a mixture of polygenic GxE relationships seems most fitting (as observed, for example, in Zhu et al., 2023, for GxSex in human physiology).

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *MAJOR COMMENT*
>
> *The amplification model is based on an understanding of gene networks in which environmental variables concertedly alter the effects of clusters of genes, or modules, in the network (e.g., if an environmental variable alters the effect of some gene, it indirectly and proportionately alters the effects of genes downstream of that gene in the network---or upstream if the gene acts as a bottleneck in some pathway). It is clear in this model that (i) multiple environmental variables could amplify distinct modules, and (ii) a single environmental variable could itself amplify multiple separate modules, with a separate amplification factor for each module.*
>
> *However, perhaps inspired by their previous work on GxSex interactions in humans, the authors' focus in the present manuscript is on cases where there are only two environments ("control" and "high-sugar diet" in the Drosophila dataset that they reanalyze, and "A" and "B" in their simulations [and single-locus mathematical analysis]), and they consider models where these environments amplify only a single set of genes, i.e., with a single amplification factor. While it is of course interesting that a single-amplification-factor model can generate data that resemble those in the Drosophila dataset that the authors re-analyze, most scenarios of amplification GxE will presumably be more complex. It seems that detecting amplification in these more complex scenarios using methods such as the authors do in their final section will be correspondingly more difficult. Indeed, in the limit of sufficiently many environmental variables amplifying sufficiently many modules, the scenario would resemble one of idiosyncratic single-locus GxE which, as the authors argue, is very difficult to detect. That more complex scenarios of amplification, with multiple environments separately amplifying multiple modules each, might be difficult to detect statistically is potentially an important limitation to the authors' approach, and should be tested in their simulations.*

We agree that characterizing GxE when there is a mixture of drivers of context-dependency is difficult. Developing a method that does so across multiple (and perhaps not pre-defined) contexts is of high interest to us but beyond the scope of the current manuscript

We note that for GxSex, modeling this mixture does generally improve phenotypic prediction, and more so in traits where we infer amplification as a major mode of GxE.

> *MINOR COMMENTS*
>
> *Lines 88-90: "This estimation model is equivalent to a linear model with a term for the interaction between context and reference allele count, in the sense that context-specific allelic effect estimators have the same distributions in the two models."*

> *Does this equivalence require the model with the interaction term also to have an interaction term for the intercept, i.e., the slope on a binary variable for context (since the generative model in Eq. 1 allows for context-specific intercepts)?*

It does require an interaction term for the intercept. This is e_i (and its effect beta_E) in Eq. S2 (line 70 of the supplement).

> *Lines 94-96: Perhaps just a language thing, but in what sense does the estimation model described in lines 92-94 "assume" a particular distribution of trait values in the combined sample? It's just an OLS regression, and one can analyze its expected coefficients with reference to the generative model in Eq. 1, or any other model. To say that it "assumes" something presupposes its purpose, which is not clear from its description in lines 92-94.*

We corrected "assume" to "posit".

> *Lines 115-116: It should perhaps be noted that the weights wA and wB need not sum to 1.*

Indeed; it is now explicitly stated.

> *Lines 154-160: I think the role of r could be made even clearer by also discussing why, when VA>>VB, it is better to use the whole-sample estimate of betaA than the sample-A-specific estimate (since this is a more counterintuitive case than the case of VA<<VB discussed by the authors).*

This is addressed in lines 153-154, stating: "Typically, this (VA<<VB) will also imply that the additive estimator is greatly preferable for estimating β_B , as β_B will be extremely noisy"

> *Line 243 and Figure 4 caption: The text states that the simulated effects in the high-sugar environment are 1.1x greater than those in the control environment, while the caption states that they are 1.4x greater.*

We have corrected the text to be consistent with our simulations.

> *TYPOS/WORDING*
>
> *Line 14: "harder to interpret" --> "harder-to-interpret"*
>
> *Line 22: We --> we*
>
> *Line 40: "as average effect" -> "as the average effect"?*
>
> *Line 57: "context specific" --> "context-specific"*
>
> *Line 139: "re-parmaterization" --> "re-parameterization"*
>
> *Lines 140, 158, 412: "signal to noise" --> "signal-to-noise"*
>
> *Figure 3C,D: "pule rate" --> "pulse rate"*
>
> *The caption of Figure 3: "conutinous" --> "continuous"*
>
> *Line 227: "a variant may fall" --> "a variant may fall into"*
>
> *Line 295: "conferring to more GxE" --> "conferring more GxE" or "corresponding to more GxE"? This is very pedantic, but I think "bias-variance" should be "bias--variance"*

> *throughout, i.e., with an en-dash rather than a hyphen.*

We have corrected all of the above typos.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *(This section repeats some of what I wrote earlier).*
>
> *- First polygenic architecture part: the manuscript focuses on "top hits" in trying to identify sets of variants that are context-specific. This "top hits" approach seems somewhat esoteric and, as written, not connected tightly enough to the bias-variance tradeoff issue. The first section of the paper which focuses on bias-variance trade-off mostly deals with estimation. The "top hits" section deals with testing, which introduces additional issues that are due to thresholding. Perhaps the authors can think of ways to make the connection stronger between the bias-variance tradeoff part to the "top hits" part, e.g., by introducing testing earlier on and/or discussion estimation in addition to testing in the "top hits" part of the manuscript. The second polygenic architecture part: polygenic scores that account for interaction terms. Here the authors focused (well, also here) on pervasive amplification in simulations. This part combines estimation and testing (both the choice of variants and their estimated effects are important). In pervasive amplification the idea is that causal variants are shared, the results may be different than in a model with context-specific effects and variant selection may have a large impact. Still, I think that these simulations demonstrate the idea developed in the bias-variance tradeoff part of the paper, though the reader is left to wonder whether a bias-variance decision rule should or could be utilized for PGS construction.*

In both of these sections we discuss how the consideration of polygenic GxE patterns alters the conclusions based on the single-variant tradeoff. In the "top hits" section, we show that single-variant classification itself, based on a series of marginal hypothesis tests alone, can be misleading. The PGS prediction accuracy analysis shows that both approaches are beaten by the polygenic GxE estimation approach. Intuitively, this is because the consideration of polygenic GxE can mitigate both the bias and variance, as it leverages signals from many variants.

We agree that the links between these sections of the paper were not sufficiently clear, and have added signposting to help clarify them (lines 176-180; lines 275-277; lines 316-321).

> *- Simulation of GxDiet effects on longevity: the methods of the simulation are strange, or communicated unclearly. The authors' report (page 17) poses a joint distribution of genetic effects (line 439), but then, they simulated effect estimates standard errors by sampling from summary statistics (line 445) rather than simulated data and then estimating effect and effect SE. Why pose a true underlying multivariate distribution if it isn't used?*

We rewrote the Methods section "Simulation of GxDiet effects on longevity in Drosophila to make our simulation approach clearer (lines 427-449). We are indeed simulating the true effects from the joint distribution proposed. However, in order to mimic the noisiness of the experiment in our simulations, we sample estimated effects from the true simulated effects, with estimation noise conferring to that estimated in the Pallares et al. dataset (i.e., sampling estimation variances from the squares of empirical SEs).

> *- How were the "most significantly associated variants" selected into the PGS in the polygenic prediction part? Based on a context-specific test? A combined-context test of effect size estimates?*

For the “Additive” and “Additive ascertainment, GxE estimation” models (red and orange in Fig. 5, respectively), we ascertain the combined-context set. For the “GxE” and “polygenic GxE” (green and blue in Fig. 5, respectively) models, we ascertain in a context-specific test. We now state this explicitly in lines 280-288 and lines 507-526.

> *- As stated, I find the conclusion statement not specific enough in light of the rest of the manuscript. “the consideration of polygenic GxE trends is key” - this is very vague. What does it mean “to consider polygenic GxE trends” in the context of this paper? I can’t tell. “The notion that complex trait analyses should combine observations at top associated loci” - I don’t think the authors really refer to combining “observations”, rather perhaps combine information from top associated loci. But this does not represent the “top hits” approach that merely counts loci by their testing patterns. “It may be a similarly important missing piece...” What does “it” refer to? The top loci? What makes it an important missing piece?*

We rewrote the conclusion paragraph to address these concerns (lines 316-321).

https://doi.org/10.7554/eLife.99210.2.sa0