Tradeoffs in Modeling Context Dependency in Complex Trait Genetics

Eric Weine[1,2,3], Samuel Pattillo Smith[1,2], Rebecca Kathryn Knowlton[4], Arbel Harpak[1,2,+]

[1]Department of Integrative Biology, The University of Texas at Austin

[2]Department of Population Health, The University of Texas at Austin

[3]Department of Human Genetics, University of Chicago

[4]Department of Statistics and Data Sciences, The University of Texas at Austin

[+] Correspondence should be addressed to: arbelharpak@utexas.edu

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 derive governing principles for when modeling GxE in GWAS is useful. We derive a decision rule for choosing between competing models for the estimation of allelic effects. The rule weighs the increased noise of context-specific effect estimation against the potential bias when context dependency is ignored. In human GWAS, the increased noise of context-specific estimation often outweighs the bias reduction, rendering GxE models less useful when variants are considered independently. However, for complex traits, considering patterns of context dependency across many variants jointly can substantially improve estimation and trait prediction. Finally, we demonstrate via example that considering polygenic trends of GxE may also be important for interpretation, as analyses based on independently ascertained "top hits" alone can be misleading.

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 phenotype. The estimand here can be thought of as average effect of the allele over a distribution of multidimensional contexts. 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 [17–20]. 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 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 or genome-wide.

Results and Discussion

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

The problem setup. We consider a sample of 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 phenotype 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 phenotype. We assume a generative model of the form

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

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, conferring to the mean phenotype 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 distributions in the two models.

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 assumes that for $i = 1, \ldots, n+m$,

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

where $\alpha_{A \cup B}$ is the mean phenotype 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 ([21, 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. 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

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

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

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

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-parmaterization 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. $\frac{1+\omega_A}{1-\omega_A}$ 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.

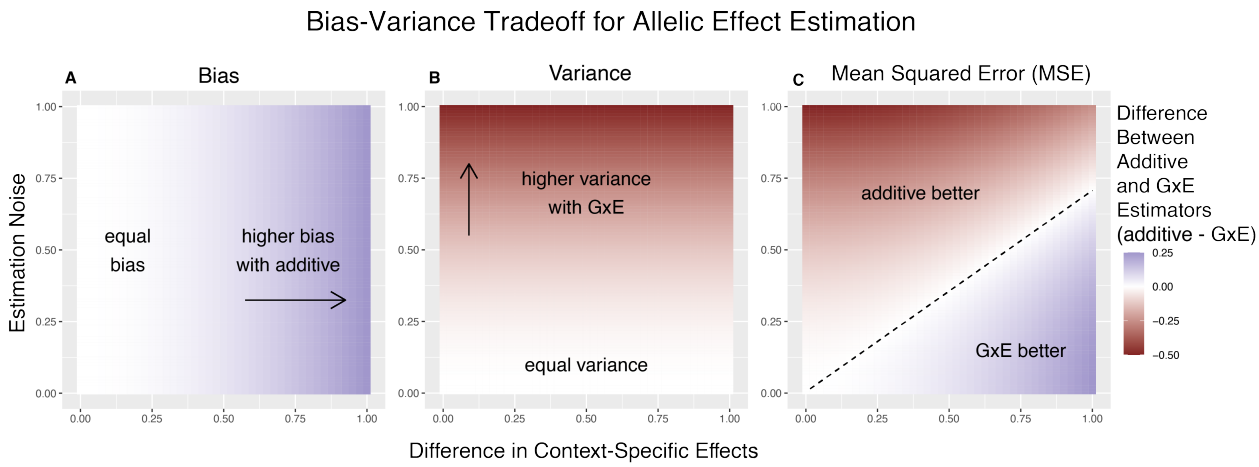

Figure 1: Bias-Variance tradeoff for single-site estimation with equal estimation noise and equal heterozygozity across contexts. The x-axis shows the difference in context specific effects, while the y-axis shows the standard deviation of the context-specific estimators—both in raw measurement units. The color on the plot indicates the difference between the additive and GxE estimators in bias (A), variance (B) or MSE (C). ***(A)*** *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.*

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, 22]. 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**; **Methods**). 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

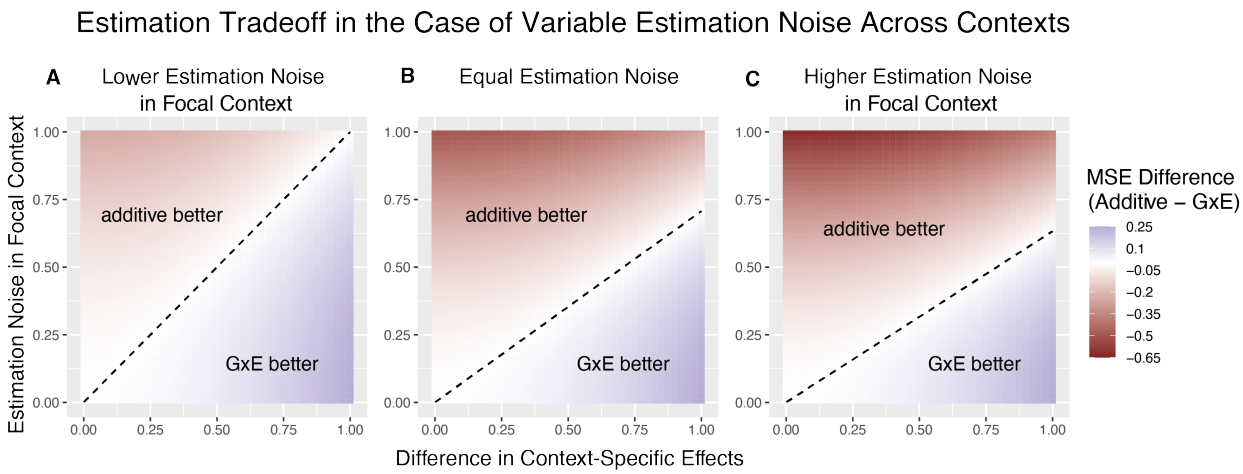

Figure 2: The decision boundary with different ratios of context-specific estimation noises. In all panels, the heterozygozity of the variant is assumed to be equal across contexts. The x and y axes are the same as in Fig. 1. Estimation noise in the focal context, A, is half that of the other context, B. ***(B)*** *Estimation noise is equal in both contexts.* ***(C)*** *Estimation noise in focal context is double that of the other context.*

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**).

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, 19, 23–27]. It stands to reason that to evaluate context-dependence in complex traits, we would also want to jointly consider highly polygenic patterns, rather than just the patterns at the loci most strongly associated with a trait [3, 13, 28–31].

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, 32, 33]. In what follows, we will therefore

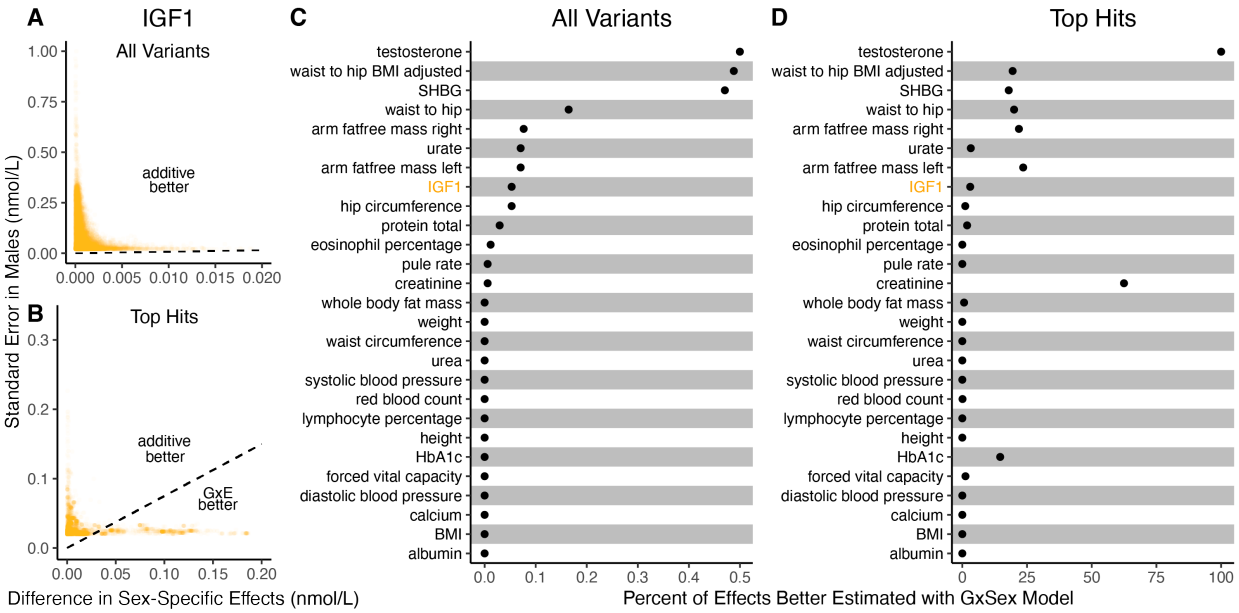

Figure 3: Applying the decision rule to sex-dependent effects on human physiological traits. ***(A-B)*** *The x-axis shows the estimated absolute difference between the effect of variants in males and females. The y-axis shows the measured standard error for each variant in males, the focal context here. 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 conutinuous physiological traits. To estimate these percentages, one single nucleotide variant is sampled from each of 1,700 approximately independent auotosomal linkage blocks, and this procedure is repeated 10 times. Shown are average percentages across the 10 iterations.*

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* [34], GxSex effects on various traits in humans [35, 36] and GxDietxAge effects on body weight in mice [37].

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 [23, 25], 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* [38]. 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 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.*

The choice of categories a variant may fall into is motivated by the longstanding search for "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, 32, 39]. 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 that there is 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.

A key example is the case where true effects are concordant in sign but differ in magnitude. The strong positive covariance of estimated effects observed genome-wide (**Fig. 4A**) suggests this case merits consideration. Such variants may fall into each of the existing categories, depending on the magnitude of effects and statistical power in each of the contexts. Three of the four possible classifications are clearly wrong, but what about the "shared" category?

The class "shared" may be interpreted as suggesting lack of context dependency (Fig. 1C in [38]). However, it will tend to include variants having strong effects under both diets, regardless of whether or not the diet-specific effects are similar. As Pallares et al. also note, there are marked differences in the magnitude of diest-specific estimated effects of variants in the "shared" category. Amongst the approximately $1,500$ variants labeled as shared, the estimated effect under the high-sugar diet is on average about $1.3\times$ that of the estimated effect in the control diet.

Notably, the classification as diet-specific does not imply that a variant has an unusually large effect under this diet. On average, variants classified as shared actually have a slightly larger estimated effect under the high-sugar diet than variants classified as high-sugar specific (t-test p=1.6×10^{-5}). Thus, instead of suggesting little-to-no effect under the control diet and a large effect under the high-sugar diet (as predicted by the hypothesis that cryptic genetic variation is pervasive), the classification as high-sugar specific may commonly just point us to intermediate size effects—large enough to be significant in the systematically-larger effects

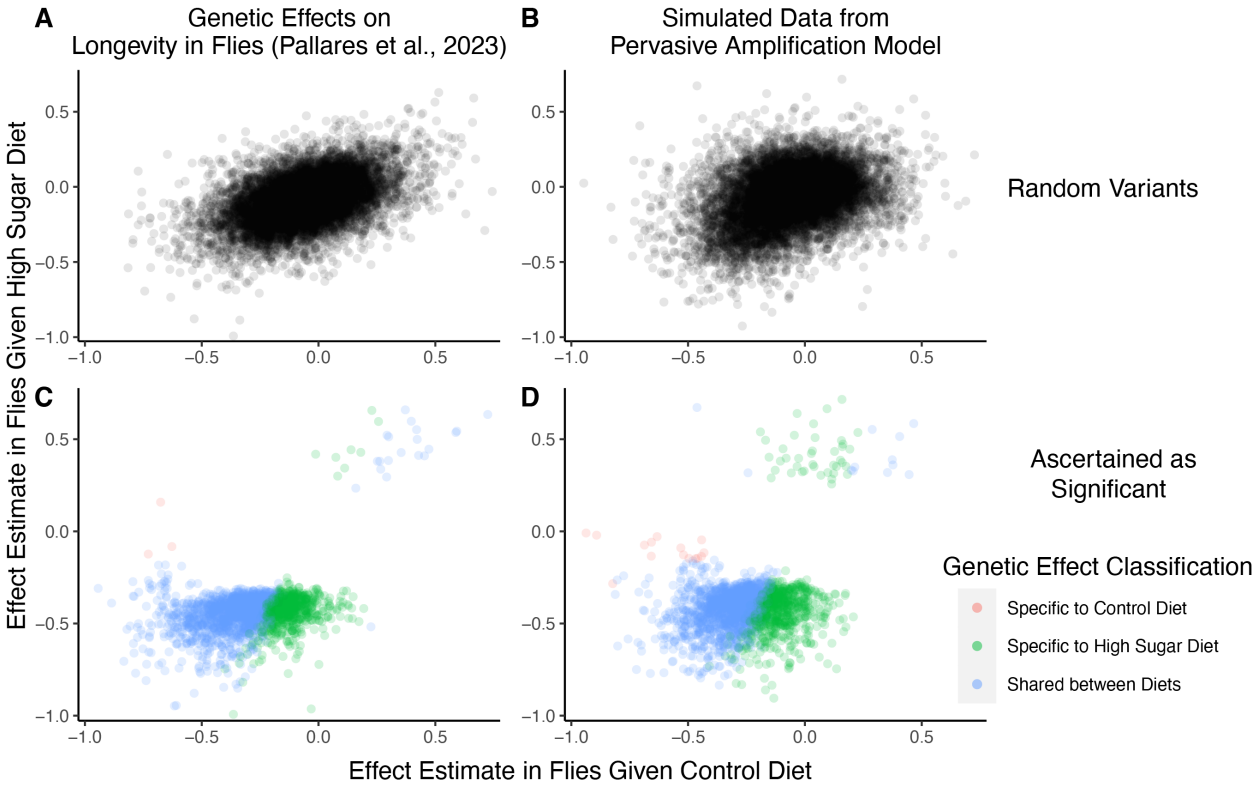

Figure 4: A focus on top hits may be lead to mischaracterization of polygenic GxE. ***(A)*** *Data from an experiment measuring genetic effects on longevity in caged flies given one of two diets. Shown are allelic effect estimates under each diet for a random sample of approximately 12K variants.* ***(B)*** *Simulated data where all true allelic effects are exactly* 1.4 *times larger in the high-sugar diet context. 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., and is at odds with their conclusions about evidence for pervasive cryptic genetic variation, 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).*

context (where power is higher) yet too small to be significant in the systematically-smaller effects context (where power is lower).

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. In particular, we sampled from a mixture of 60% of variants having no effect under either diet and 40% 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 the ap-

proach: 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 conclusion, the classification of "top hits" alone may not be representative of the extent of GxE nor of the most pervasive modes of GxE.

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 [33, 40–43]. 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 [17, 18, 44, 45]. 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 the 833 most significantly associated variants (corresponding to one-third of the causal variants).

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, conferring to more GxE (**Fig. S4B**).

Finally, we considered a polygenic GxE approach, as implemented in "multivariate adaptive shrinkage" (*mash*) [28], 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

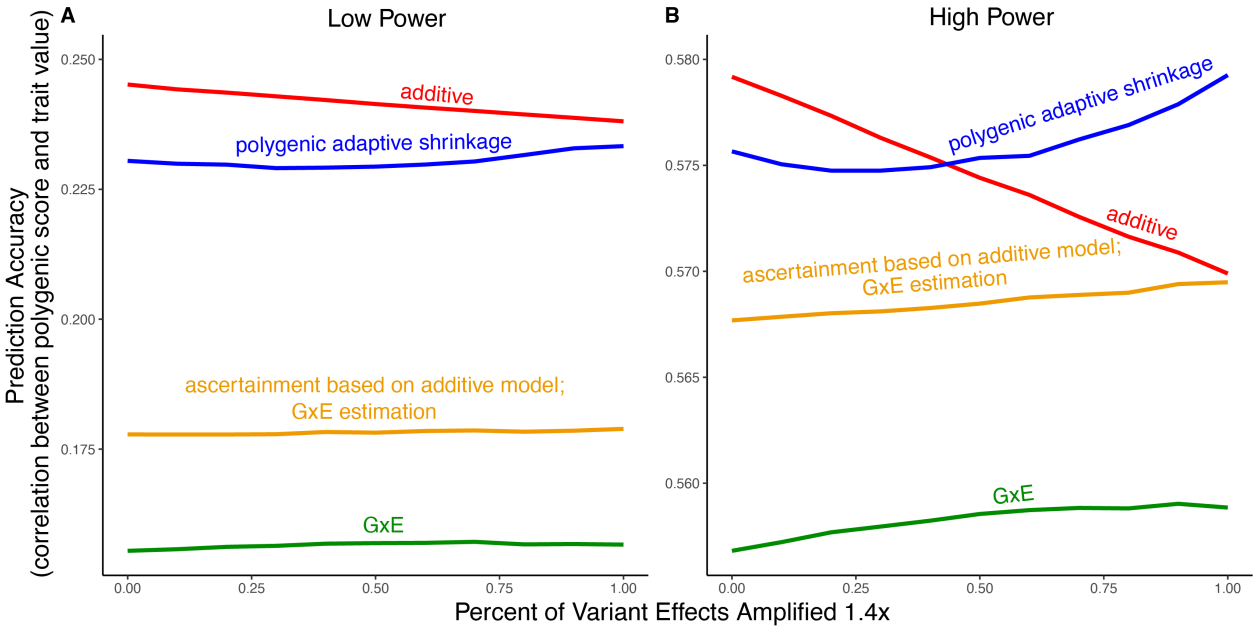

Figure 5: ***Polygenic score performance for context-dependent prediction models.*** *In each simulation, a GWAS is performed on* $5,000$ *biallelic variants, half of which have no effect in either context. Of the other half, some percent of the variants (indicated on the x-axis) had effects* $1.4\times$ *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.*

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.

Conclusion. In this work, we argue through examples that in order to understand context dependency in complex traits, and also to improve prediction, the consideration of polygenic GxE trends is key. 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 a similarly important missing piece in our understanding of context-dependent effects.

Acknowledgements

We thank Peter Andolfatto, Doc Edge, 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. S.P. Smith was supported by NIH RF1AG073593.

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:

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

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

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 phenotypes, 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

$$\begin{aligned}
\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 phenotypes)} \\
&= \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.
\end{aligned}$$

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

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

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

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

$$V_A < \omega_A^2 V_A + \omega_B^2 V_B + \big((\omega_A-1)\beta_A + \omega_B\beta_B\big)^2$$

$$\iff 1 < \omega_A^2 + \omega_B^2 \frac{V_B}{V_A} + \frac{\big((\omega_A - 1)\beta_A + \omega_B \beta_B\big)^2}{V_A}$$
$$\iff 1 - \omega_A^2 < \omega_B^2 \frac{V_B}{V_A} + \frac{\big((\omega_A - 1)\beta_A + \omega_B \beta_B\big)^2}{V_A}.$$

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

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

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

Validating and applying the decision rule to real data

To validate the decision rule above, we considered the example of sex as a context for physiological traits using UK Biobank data [22]. For each of 27 physiological quantitative traits, we first split the sample by sex chromosome karyotypes into XX individuals (females) and XY individuals (famles). For each sex, we randomly split the sample into a training set with 80% of the individuals and a test set with 20% of the individuals. We estimated within-sex GWAS in the training and test sets separately. We denote by $\hat{\beta}_{Z_i}^{(t)}$ the marginal effect estimate of the ith SNP in sex Z in set t, and by $\hat{s}_{Z_i}^{(t)}$ the corresponding standard error. We denote by $\hat{\beta}_{M \cup F_i}^{(train)}$ the estimated effect of the ith SNP in the union of training sets (with both sexes included). In our validation procedure, we treated $\hat{\beta}_{M_i}^{(test)}$ as the "ground-truth," i.e., as the true effect. While $\hat{\beta}_{M_i}^{(test)}$ is of course a very noisy estimate of β_{M_i}, it is an unbiased estimate, so we expect that estimates that are closer to $\hat{\beta}_{M_i}^{(test)}$ in the aggregate will also be closer to β_{M_i}. Using this assumption, we define two empirical quantities,

$$\begin{aligned}
error(\hat{\beta}_{M_i}^{(train)}) &= (\hat{\beta}_{M_i}^{(train)} - \hat{\beta}_{M_i}^{(test)})^2 \\
error(\hat{\beta}_{M \cup F_i}^{(train)}) &= (\hat{\beta}_{M \cup F_i}^{(train)} - \hat{\beta}_{M_i}^{(test)})^2.
\end{aligned}$$

Given $\sqrt{V_{M_i}}$ and $|\beta_{M_i} - \beta_{F_i}|$, **Eq. 4** allows us to calculate the expected difference between $error(\hat{\beta}_{M_i}^{(train)})$ and $error(\hat{\beta}_{M \cup F_i}^{(train)})$. However, we do not observe $\sqrt{V_{M_i}}$ or $|\beta_{M_i} - \beta_{F_i}|$. Instead, we estimated these quantities and examined the agreement of the estimates in the two sets. We estimated $\sqrt{V_{M_i}}$ as $\hat{s}_{M_i}^{(train)}$. Estimating $|\beta_{M_i} - \beta_{F_i}|$ is more difficult, as the intuitive $|\hat{\beta}_{M_i}^{(train)} - \hat{\beta}_{F_i}^{(train)}|$ is an upwardly biased estimator. Instead, we

used an empirical shrinkage estimator (*ash* [46]) to estimate $|\beta_{M_i} - \beta_{F_i}|$. In essence, *ash* takes in a vector of effect estimates $\hat{\boldsymbol{\beta}}$ and a vector of corresponding standard errors $\hat{\boldsymbol{s}}$, estimates a prior distribution of the true effects $p(\boldsymbol{\beta})$, and then uses this prior to obtain a posterior distribution of true effects.

Specifically, for a large random sample of sites, $i = 1, \ldots, p$, we ran *ash* on the vectors

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}^{(train)}_{M_1} - \hat{\beta}^{(train)}_{F_1}, \ldots, \hat{\beta}^{(train)}_{M_p} - \hat{\beta}^{(train)}_{F_p}\right)$$
$$\hat{\boldsymbol{s}} = \left(\sqrt{(\hat{s}^{(train)}_{M_1})^2 + (\hat{s}^{(train)}_{F_1})^2}, \ldots, \sqrt{(\hat{s}^{(train)}_{M_p})^2 + (\hat{s}^{(train)}_{F_p})^2}\right).$$

and took the absolute value of *ash*'s output posterior mean estimates of $\beta_{M_i} - \beta_{F_i}$ as our estimate of $|\beta_{M_i} - \beta_{F_i}|$ for each site. With our estimates of $|\beta_{M_i} - \beta_{F_i}|$ and $\sqrt{V_{M_i}}$ in hand, we calculated the expected difference in squared errors using **Eq. 4** (x-axis in **Fig. 3A**) and compared it with the actual difference between $error(\hat{\beta}^{(train)}_{M_i})$ and $error(\hat{\beta}^{(train)}_{M \cup F_i})$.

For **Fig. 3A,B**, we used the same procedure as above to estimate the x and y axes, except we did not employ any data splitting.

Simulation of GxDiet effects on longevity in Drosophila

In **Fig. 4**, we compare the empirical allelic effectestimates derived empirically by Pallares et al. to ones we estimated from simulations of pervasive amplification. Here, we detail the simulation approach.

Our goal was to emulate the joint distribution of genetic effects on longevity observed in Pallares et al. We first designed a generative model for the true effects under each diet. For variants $j = 1, \ldots, 50,000$, we simulated a true effect under the high-sugar diet (β_{h_j}) and for the control environment (β_{c_j}). 60% of variants were randomly assigned to have no effect under either diet, with the remaining effects sampled as

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

which corresponds to perfect correlation of effects between diets where effects in the high-sugar diet are systematically amplified 1.4× (and systematically smaller by −.025 units) compared to the control diet. We selected these parameters based on inspection of the resulting distribution of effects and its 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. This process yielded a vector of estimated effects in the high-sugar group and control group, $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\beta}}_c$, respectively, and vectors of standard errors in the high-sugar group and control group, $\hat{\boldsymbol{s}}_h$ and $\hat{\boldsymbol{s}}_c$, respectively. We computed a z-test for each variant in 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 [47], 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 $j = 1, \ldots, 50,000$:

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 $j = 1, \ldots, p$ 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 $1-\alpha$ is the proportion of non-null SNPs which are generated as perfectly correlated but with $1.5\times$ 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 $\hat{\boldsymbol{s}}_{A\cup B}$ 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 $n \times 2$ matrices $\begin{bmatrix}\hat{\boldsymbol{\beta}}_A & \hat{\boldsymbol{\beta}}_B\end{bmatrix}$ (of effects) and $\begin{bmatrix}\hat{\boldsymbol{s}}_A & \hat{\boldsymbol{s}}_B\end{bmatrix}$ (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 $\Omega_A \subset \{1, \ldots, p\}$ consisting of the variants with the 833 smallest p-values and $\Omega_B \subset \{1, \ldots, p\}$ 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.

References

1. El-Soda, M., Malosetti, M., Zwaan, B. J., Koornneef, M. & Aarts, M. G. Genotype× environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends in plant science* **19,** 390–398 (2014).
2. Vieira, C. *et al.* Genotype-environment interaction for quantitative trait loci affecting life span in Drosophila melanogaster. *Genetics* **154,** 213–227 (2000).
3. Des Marais, D. L., Hernandez, K. M. & Juenger, T. E. 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 (2013).
4. Smith, E. N. & Kruglyak, L. Gene–environment interaction in yeast gene expression. *PLoS biology* **6,** e83 (2008).
5. Paaby, A. B. & Rockman, M. V. Cryptic genetic variation: evolution's hidden substrate. *Nature Reviews Genetics* **15,** 247–258 (2014).
6. Munafò, M. R., Zammit, S. & Flint, J. 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 (2014).
7. Kraft, P. & Aschard, H. Finding the missing gene–environment interactions. *European Journal of Epidemiology* **30,** 353–355 (2015).
8. Sella, G. & Barton, N. H. 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 (2019).
9. Zhu, C. *et al.* Amplification is the primary mode of gene-by-sex interaction in complex human traits. *Cell Genomics,* 100297 (2023).
10. Schwaba, T. *et al.* Comparison of the Multivariate Genetic Architecture of Eight Major Psychiatric Disorders Across Sex. *medRxiv,* 2023–05 (2023).
11. Elgart, M. *et al.* Correlations between complex human phenotypes vary by genetic background, gender, and environment. *Cell Reports Medicine* **3,** 100844 (2022).
12. Duncan, L. E. & Keller, M. C. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry* **168,** 1041–1049 (2011).
13. Gibson, G. & Lacek, K. A. Canalization and robustness in human genetics and disease. *Annual Review of Genetics* **54,** 189–211 (2020).
14. Brown, B. C., Ye, C. J., Price, A. L., Zaitlen, N., Consortium, A. G. E. N. T. 2. D., *et al.* Transethnic genetic-correlation estimates from summary statistics. *The American Journal of Human Genetics* **99,** 76–88 (2016).
15. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genetics* **13,** e1006711 (2017).
16. Balliu, B. *et al.* An integrated approach to identify environmental modulators of genetic risk factors for complex traits. *The American Journal of Human Genetics* **108,** 1866–1879 (2021).
17. Fisher, R. A. *The genetical theory of natural selection* (Clarendon Press, 1930).
18. Falconer, D. S. & Mackay, T. F. *Introduction to quantitative genetics* (Longman, 1996).
19. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610,** 704–712 (2022).

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

21. Casella, G. & Berger, R. L. *Statistical Inference* (Cengage Learning, 2021).

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

23. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *Elife* **10,** e58615 (2021).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

39. 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).

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

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

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

43. Spence, J. P., Sinnott-Armstrong, N., Assimes, T. & Pritchard, J. K. A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics. *bioRxiv,* 2022–04 (2022).

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

45. Young, A. I. Solving the missing heritability problem. *PLoS Genetics* **15,** e1008222 (2019).

46. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18,** 275–294 (2017).

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