# *A Perspective on Interaction Tests in Genetic Association Studies*

Hugues Aschard,$_1$

[1]*Harvard School of Public Health, Department of Epidemiology, Boston, MA 02115, USA*

Corresponding author:

Hugues Aschard

haschard@hsph.harvard.edu

## Abstract

The identification of gene-gene and gene-environment interaction in human traits and diseases is an active area of research that generates high expectation, and most often lead to high disappointment. This is partly explained by a misunderstanding of some of the inherent characteristics of interaction effects. Here, I untangle several theoretical aspects of standard regression-based interaction tests in genetic association studies. In particular, I discuss variables coding scheme, interpretation of effect estimate, power, and estimation of variance explained in regard of various hypothetical interaction patterns. I show first that the simplest biological interaction models—in which the magnitude of a genetic effect depends on a common exposure—are among the most difficult to identify. Then, I demonstrate the demerits of the current strategy to evaluate the contribution of interaction effects to the variance of quantitative outcomes and argue for the use of new approaches to overcome these issues. Finally I explore the advantages and limitations of multivariate models when testing for interaction between multiple SNPs and/or multiple exposures, using either a joint test, or a test of interaction based on risk score. Theoretical and simulated examples presented along the manuscript demonstrate that the application of these methods can provide a new perspective on the role of interaction in multifactorial traits.

## Introduction

Hundreds of studies have searched for gene-gene and gene-environment interaction effects in human data with the underlying motivation of identifying or at least accounting for potential biological interaction. So far, this quest has been quite unsuccessful and the large number of methods that have been developed to improve detection[1-5] have not qualitatively changed this situation. This lack of discovery in the face of a large research investment has been discussed in several review papers that have pointed out a number of issues specific to interaction tests, including exposure assessment, time-dependent effect, confounding effect and multiple comparisons[2; 6; 7]. While these factors are obvious barriers to the identification of interaction effects, it appears that some of the limitations of standard regression-based interaction tests that pertain to the nature of interaction effects are greatly underestimated. Previous work showed the detection of some biologically meaningful interaction effects requires larger sample sizes than marginal effects for a similar effect size[8; 9], however it is not an absolute rule. Understanding the theoretical basis of this lack of power can help us optimizing study design to improve detection of interaction effect in human traits and diseases, and open the path for new methods development. Moreover the interpretation of effect estimates from interaction models often suffer from various imprecisions. Compared to marginal models, the coding scheme for interacting variables can impact effect estimates and association signals for the main effects[9]. Also, the current strategy to derive the contribution of interaction effects to the variance of an outcome greatly disadvantages interaction effects and are inappropriate when the goal of a study is not prediction but to assess the relative importance of an interaction term from a biological perspective. While alternative approaches exist, they have not so far been considered in genetic association studies. Finally, the development of new pairwise gene-gene and gene-environment interaction tests is reaching some limits, because the number of assumption that can be leveraged to improve power is limited when only two predictors are considered. With the exponential increase of available genetic and non-genetic data, the development and application of multivariate interaction tests offer new opportunities to building powerful approaches and moving the field forward.

## Methods and Results

### *Coding scheme and effect estimates*

Consider an interaction effect between a single nucleotide polymorphism (SNP) $G$ and an exposure $E$ (which can be an environmental exposure or another genetic variant) on a quantitative outcome $Y$. For simplicity I assume in all further derivation that $E$ is normally distributed with variance 1, and $G$ and $E$ are independents. The simplest and most commonly assumed underlying model for $Y$ (i.e. the model used to generate the values of $Y$) when testing for an interaction effect between $G$ and $E$ is defined as follows:

$$Y = \beta_G \times G + \beta_E \times E + \beta_{GE} \times G \times E + \varepsilon$$

where $\beta_G$ is the main effect of $G$, $\beta_E$ is the main effect of $E$, $\beta_{\mathrm{GE}}$ is a linear interaction between $G$ and $E$, and $\varepsilon$, the residual, is normally distributed with mean 0 and variance $\sigma^2$ sets so that the variance of Y equals 1. One can then evaluate the impact of applying linear transformation of the genotype and/or the exposure when testing for main and interaction effects. For example, assuming $E$ has a mean > 0 and $G$ is defined as the number of coded allele in the generative model, $Y$ can be rewritten as a function of $G_{std}$ and $E_{std}$, the standardized $G$ and $E$:

$$Y = \beta'_G \times G_{std} + \beta'_E \times E_{std} + \beta'_{GE} \times G_{std} \times E_{std} + \varepsilon'$$

where $\beta'_G$ ,$\beta'_E$ and $\beta'_{GE}$ are the main effects of $G_{std}$ and $E_{std}$ and their interaction. Relating the standardized and unstandardized equations, we obtain (**Appendix A**):

$$\beta'_G = (\beta_G + \beta_{GE} \times \mu_E) \times \sigma_G$$

$$\beta'_E = (\beta_E + \beta_{GE} \times \mu_G) \times \sigma_E$$

$$\beta'_{GE} = \beta_{GE} \times \sigma_E \times \sigma_G$$

where $\mu_G$, $\sigma_G$, $\mu_E$ and $\sigma_E$ are the mean and variance of $G$ and $E$, respectively. Hence, the estimated main effects of $G_{std}$ and $E_{std}$ not only scale with the variance of $G$ and $E$ but can also change qualitatively if there is an interaction effect. In comparison, the interaction effect $\beta'_{GE}$ only scales with the predictors variance, however, because $\beta'_{GE}$ does not depend on $\sigma_{GE}$ the variance of the interaction term but on the variance of $G$ and $E$, the magnitude of the interaction effect can change.

Which coding scheme for $G$ and $E$ has the most biological sense can only be discussed on a case by case basis and is therefore out of the scope of this paper. The important point here, is that coding scheme should be chosen carefully when testing an interaction as it can correspond to profoundly different patterns on the outcome. This is illustrated in **Figure 1**, which shows the contribution of a pure interaction effect ($\beta_G = \beta_E = 0$ and $\beta_{GE} \neq 0$) to $Y$. When $G$ and $E$ are centered, the joint effect of $G$ and $E$ is similar across the most extreme sub-groups (low exposure and homozygote for the protective allele *vs* high exposure and homozygote for the risk allele) and opposite effect otherwise (**Figure 1a**). Conversely, when $G$ and $E$ are positive or null, the interaction term simply corresponds to an increase (or decrease if the interaction term is negative) of the magnitude of a genetic effect when the exposure increases (**Figure 1b**). Hence, assuming $G$ and/or $E$ have a negative range in the generative model – besides it might have limited biological meaning – implies interaction effects of different nature as compared to model using positive predictors only. Furthermore, when the mean of the exposure increases while its variance is fixed, a realistic interaction effect for genetic data (i.e. explaining a small amount of the outcome variance) will appear more and more as a sole genetic effect (see **Figure S1**).

While estimates for a specific coding scheme can be derived from estimates obtained from another coding scheme, questions arise on which final coding to choose and how to interpret estimates when modeling an interaction. This point, and the motivation for adding non-linear terms in general have been already debated and several general guidelines have been proposed (see for example the review by Robert J. Friedrich[10]). The consensus was that, if the range of the independent variables do naturally includes zero (e.g. smoking status, genetic variants) there is no problem in interpreting the estimated main and interaction effect. For an interaction effect between A and B, the main effect of A corresponds to the effect of A when B is null and conversely. Conversely, if the range of the variables do not naturally encompass zero, then the observed estimates "*will be an extrapolations beyond the observed range of experience*"[10]. Centering the variables can be an option to address this concern. In that case, the main effect of A and B would represent the effect of A among individuals having the mean value of B and conversely. However, as mentioned previously, using centered variables induces a less interpretable interaction term. A reasonable alternative consists in shifting the exposure values so that it has a minimum value close to 0, or alternatively to use ordinal categories of the exposure (e.g. high versus low BMI as done to define obesity), so that the main effect of A would correspond to the effect among the lowest observed value of B and conversely.

### *Power considerations*

The power of the tests from the interaction model and from a marginal genetic model defined as $Y = \beta_{mG} \times G + \varepsilon_m$, can be compared when deriving the non-centrality parameters (*ncp*) of the predictors of interest. Assuming all effects are small, so that $\sigma^2$ the residual variance is close to 1, these *ncp* can be approximated by (see **Appendix B**):

$$ncp_G \approx N \times \sigma_G^2 \times \beta_G^2 \times \frac{\sigma_E^2}{\mu_E^2 + \sigma_E^2}$$

$$ncp_E \approx N \times \sigma_E^2 \times \beta_E^2 \times \frac{\sigma_G^2}{\mu_G^2 + \sigma_G^2}$$

$$ncp_{GE} \approx N \times \sigma_E^2 \times \sigma_G^2 \times \beta_{GE}^2 = N \times \beta_{GE}'^2$$

$$ncp_{mG} \approx N \times \sigma_G^2 \times (\beta_G + \beta_{GE} \times \mu_E)^2 = N \times \beta_G'^2$$

Note that in such scenario adjusting for the effect of $E$ in the marginal genetic model has a minor impact on $ncp_{mG}$. It would only be important in the presence of a strong exposure effect, such an effect would reduce $\sigma^2$, the residual variance in the interaction model, and increase the *ncp*s from the interaction model but not $ncp_{mG}$.

The above equations indicate first that the significance of the marginal test of $G$ and the interaction test are invariant with the coding used in the model tested, while the significance of the test of the main genetic and exposure effects can change dramatically when shifting the mean of $G$ and $E$. Second, as illustrated in **Figure 2,** depending on the parameters of the distribution of the exposure and the genetic variants in the generative model, the relative power of each test can be dramatically different. For example if the genetic variant has only a main linear effect but is not interacting with the exposure, we obtain $ncp_G = ncp_{mG} \times \sigma_E^2/(\mu_E^2 + \sigma_E^2)$, so that testing for $\beta_{mG}$ will be much more powerful that testing for $\beta_G$ if the mean of $E$ is large, although there is no interaction effect here. When the generative model includes an interaction effect only ($\beta_G = \beta_E = 0$ and $\beta_{GE} \neq 0$), we obtain $ncp_{mG} = ncp_{GE} \times \mu_E^2/\sigma_E^2$. Again, the marginal test of the genetic effect can be dramatically more powerful than the test of interaction effect although the underlying model includes only an interaction term but no main effect.

More generally it follows that the power to detect an interaction effect explaining for example 1% of the variance of $Y$ but inducing no marginal genetic effect (i.e. when $E$ is centered as in **Figure 1a**) is much higher than for an interaction explaining the same amount of variance but whose effect can be capture by a marginal term (i.e. when $E$ is not centered as in **Figure 1b-d**). This result is a direct consequence of the covariance between $\beta_G$ and $\beta_{GE}$ that arise when having non-centered exposure in the generative model (**Figure 2e**). This covariance equals $\mu_E \times \sigma_G^2$ (**Appendix C**). It induces uncertainty on the estimation of the predictor effects, which decreases the significance of the estimates in the interaction model. With increasing inter-correlations between predictors it becomes impossible to disentangle the effects of one predictor from another, the standard errors of the effect estimates becoming infinitely large and the power decreases to the null[11]. As showed in the simulation study from **Figure S2**-**S3** these results are consistent for both linear and logistic regression and when assuming non-normal distribution of the exposure.

This lead to the non-intuitive situation where the power to detect a relatively simple and parsimonious interaction effect from a biological perspective – defined as the product of a genetic variant and an exposure both coded to be positive or null – is very small; and in most scenarios where the main genetic and interaction effects do not canceled each other (*see* e.g. [12]) the marginal association test of $G$ would be more powerful. In comparison a more exotic interaction effect as defined in **Figure 1a** and **Figure S1e**, would be both much easier to detect in a genome-wide interaction screening and not captured in a GWAS of marginal genetic effect.

### *Proportion of variance explained*

In genetic association studies the proportion of variance explained by an interaction term is commonly evaluated as the amount of variance of the outcome it can explain on top of the marginal linear effect of the interacting factors[13]. Following the aforementioned principle, one can derive the contribution of $G$ ($r_G^2$), $E$ ($r_E^2$) and $G \times E$ ($r_{GE}^2$) to the variance of the outcome using the estimates from the standardize model, in which the interaction term is independent from $G$ and $E$:

$$r_G^2 = \beta_G'^2 = (\beta_G \times \sigma_G + \beta_{GE} \times \mu_E \times \sigma_G)^2$$

$$r_E^2 = \beta_E'^2 = (\beta_E \times \sigma_E + \beta_{GE} \times \mu_G \times \sigma_E)^2$$

$$r_{GE}^2 = \beta_{GE}'^2 = (\beta_{GE} \times \sigma_E \times \sigma_G)^2$$

The total variance explained by the predictors in the interaction model equals $r_{model}^2 = r_G^2 + r_E^2 + r_{GE}^2$ (**Appendix D**). It follows that one can draw various scenarios where the estimated main effect of $E$ and $G$ can be equal to zero but have a non-zero contribution to the variance of $Y$ because of the interaction effect. Indeed, the **Figure 3** shows that depending on the frequencies of the causal allele and the distribution of the exposure in the generative model, the vast majority of the contribution of the interaction term to the variance of $Y$ will be attributed to either the genetic variant or the exposure. This is in agreement with recent work showing that even if a large proportion of the genetic effect on a given trait is induced by interaction effects, the observed contribution of interaction terms to the heritability can still be very small[13]. Because such interaction effects have small contribution to $r_{model}^2$ on top of the marginal effects of $E$ and $G$, they have a very limited utility for prediction purposes in the general population[14; 15].

This is a strong limitation when the goal is not prediction but to understand the underlying architecture of the trait under study and to evaluate the relative importance of main and interaction effects from a public health perspective. Lewontin[16] highlighted this issue a few decades ago, showing that the analysis of causes and the analysis of variance are not necessarily overlapping concepts. His work presents various scenarios where "*the analysis of variance will give a completely erroneous picture of the causative relations between genotype, environment, and phenotype because the particular distribution of genotypes and environments in a given population*". Since then, a number of theoretical studies have explored the issue of assigning importance to correlated predictors[17-20] and several alternatives measures have been proposed. To my knowledge, none of these measures has been considered so far in human genetic association studies. The advantages and limitation of these alternatives have been debated for years and no clear consensus arose, however Pratt axiomatic justification[21] for one of these measures – further presented in the literature as the Product Measure[22], Pratt index or Pratt's measure[23] – has various interesting properties that makes it a relevant substitute. For a predictor $X_i$, the Pratt's index that we refer further as $r^{2*}$, is defined as the product of $\beta_{X_i}$, the standardized coefficient from the multivariate model (where all predictors are scaled to have mean 0 and variance 1, including the interaction term), times its marginal (or zero-order) correlation with the outcome $cor(Y, X_i)$, i.e. $r^{2*}_{X_i} = \beta_{X_i} \times cor(Y, X_i)$.

By definition, $r^{2*}_{X_i}$ attributes a predictor's importance as a direct function of its estimated effect and therefore addresses the concern previously raised. Among other relevant properties, it depends only on regression coefficients, multiple correlation and residual variance but not higher moments, and it does not change with (non-constant) linear transformation of predictors other than $X_i$. It also has convenient additivity properties as it satisfies the condition $r^{2*}_G + r^{2*}_E + r^{2*}_{GE} = r^2_{model}$ (**Appendix D**), so that the overall contribution of the predictors is the sum of their individual contribution, and for example the cumulated contribution of multiple interaction effects can easily be evaluated by summing $r^{2*}_{X_i}$. The Pratt's index also received criticisms[20; 22], in particular for allowing $r^{2*}_X$ being negative[23]. Pratt's answer to this concern is that $r^{2*}_{X_i}$ only describes the average contribution of a predictor to the outcome variance in one dimension and is therefore, as any one-dimension measure, a sub-optimal representation of the complexity of the underlying model. For example, a negative $r^{2*}_{X_i}$ means that if we were able to remove the effect of $X_i$, the variance of the outcome would increase because of the correlation of $X_i$ with other predictors.

From a practical perspective, $r^{2*}_{X_i}$ can be expressed as a function of the estimated effects, the means and the variances of $E$ and $G$ (**Appendix D**), and can therefore be derived from summary statistics of standard GWAS:

$$r^{2*}_G = \beta^2_G \times \sigma^2_G + \beta_G \times \beta_{GE} \times \sigma^2_G \times \mu_E$$

$$r_E^{2*} = \beta_E^2 \times \sigma_E^2 + \beta_E \times \beta_{GE} \times \sigma_E^2 \times \mu_G$$

$$r_{GE}^{2*} = \beta_{GE}^2 \times \sigma_{GE}^2 + \beta_{GE} \times (\beta_G \times \mu_E \times \sigma_G^2 + \beta_E \times \mu_G \times \sigma_E^2)$$

As showed in **Figure 4** and **Figure S4**, the Pratt index can recover the pattern of the causal model in situations where the standard approach would dramatically underestimate the contribution of the interaction effects. It can therefore be of great use in future studies to evaluate the importance of potentially modifiable exposures that influence the genetic component of multifactorial traits.

### *Improving detection through multivariate interaction tests*

Using statistical technics such as the Pratt index can provide clues on the importance of interaction effects; however it does not help in mapping interaction. Increasing power mostly relies on two principles: increasing sample size, and leveraging assumptions on the underlying model. The case-only test, which assumes independence between the genetic variant and the exposure, and a two steps strategy which select candidate variants for interaction test based on their marginal linear effects or other parameters, are good examples of the later principle[4; 24; 25]. However, only a limited number of assumptions can be made for a single variant by a single exposure interaction test. With the overwhelming wave of genomic and environmental data, I suggest that a major path to move the field forward is to extend this principle while considering jointly more parameters.

This principle has first been applied over the past few years with the joint test of main genetic and interaction effects[26]. The *ncp* of such a joint test can be expressed as a function of main and interaction estimates ($\beta_G$ and $\beta_{GE}$), their variances ($\sigma_{\beta_G}^2$ and $\sigma_{\beta_{GE}}^2$) and their covariance $\gamma$ (**Appendix E**). By accounting for $\gamma$ the joint test recovers most of the power lost by the univariate test of the main genetic and interaction effect (so the situation where neither the interaction effect nor the main genetic effect are significant, while the joint test is, e.g. SNP rs11654749 in[27]). More importantly, in the presence of both main and interaction effects, it can outperform the marginal test of $G$. Although this is at the cost of decreased precision, i.e. if the test is significant, one cannot conclude whether association signal is driven by the main or the interaction effect. Moreover this would be true only if the contribution of the interaction effect on top of the marginal effect is large enough so that it balanced the increase in number of degree of freedom[28; 29] (**Figure 2**).

Application of the joint test of main genetic effect and a single gene by exposure interaction term is now relatively common in GWAS setting[27; 30; 31]. However, exploring further multivariate interactions with multiple exposures is limited by practical considerations. Existing software to perform the joint test in a meta-analysis context[29; 32] only allow the analysis of a single interaction term mostly because it requires the variance-covariance matrix between estimates, which is not provided by popular GWAS software. Leveraging the results from the previous sections on can show that the *ncp* of the joint test of main genetic effect and interactions with $l$ independent exposures can be expressed as the sum of *ncp* from the test of $G$ and the $G \times E_{cent.i}$ where $E_{cent.i}$ is the centered exposure $i$ (**Appendix E**):

$$ncp_{G+GE} = N \times \sigma_G^2 \times \beta_G''^2 + \sum_{i=1\ldots l} \left[N \times \sigma_G^2 \times \sigma_E^2 \times \beta_{GE_{cent.i}}''^2\right]$$

where $\beta_G''$ and $\beta_{GE_{cent.i}}''$ are the effects of $G$ and $G \times E_{cent.i}$. Such a test is robust to non-normal distribution of the exposure, and realistic correlation (<0.1) between the genetic variant and the exposures, but sensitive to the relatively high correlation (>0.1) that could be observed between exposures (**Figure S5-S6**). Hence, one can perform meta-analysis of a joint test including multiple interaction effects using existing software simply by centering exposures. In brief one would have to perform first a standard inverse-variance meta-analysis to derive chi-squares for the $l + 1$ terms from the model considered, and then to sum all chi-squares to form a chi-square with $l + 1$ *df*. Importantly, centering the exposures will be of interest only when testing jointly multiple interactions and the main genetic effect. In comparison, the combined test of multiple interaction effects can be simply performed by summing chi-squares from each independent interaction test or from interaction test derived in a joint model. As previously, the validity of this approach relies on independence between the genetic variant and the exposures, and between the exposures. Finally, a more general solutions that should be explored in future studies would consists, as proposed for the analysis of multiple phenotypes (e.g.[33]), in estimating the correlation between all tests considered (main genetic effect and/or multiple interaction effects) using genome-wide summary statistics in order to form a multivariate test.

A second major direction for the development of multivariate test is to assume the effects of multiple genetic variants depend on a single “scaling” variable $E$. Various powerful tests can be built under such an assumption. A rising approach consists in testing for interaction between the scaling variable and a genetic risk score (GRS), derived as the weighted sum of the risk alleles. Several interaction effects have been identified using this strategy[34-39], some being replicated in independent studies[36; 37]. This relative success, as compared to other univariate analysis, has generated discussion regarding potential underlying mechanisms[15; 40-43]. Overall, testing for an interaction effect between a GRS and a single exposure consists in expanding the principle of a joint test of multiple interactions while leveraging the assumption that, for a given choice of coded alleles, most interaction effects are going in the same direction. It is similar in essence to the burden test that has been widely used for rare variant analysis[44]. In its simplest form it can be expressed as the sum of all interaction effects and it captures therefore deviation of the mean of interaction effects from 0. When interaction effects are null on average, the joint test of all interaction tests (as previously described) will likely be the most powerful approach as it allows interaction effects to be heterogeneous. Conversely, if interactions tend to go in the same direction, the GRS-based test can outperform other approaches (**Figure 5)**. Of course, in a realistic scenario, a number of non-interacting SNPs would be included in the GRS, diluting the overall interaction signal and therefore decreasing power. However, the gain in power for the multivariate approaches can remain substantial even when a large proportion of the SNPs tested (e.g. 95% in the example from **Figure 5**) is not interacting with the exposure.

As showed in **Appendix F**, assuming the SNPs in the GRS are independents, the GRS by $E$ interaction test can be derived from individual interaction effect estimates. More precisely, consider testing the effect of a weighted GRS on $Y$:

$$Y \sim \gamma_{GRS} \times GRS + \gamma_E \times E + \gamma_{INT} \times GRS \times E$$

where $\gamma_{GRSm}$, $\gamma_E$ and $\gamma_{INT}$ are the main effect of the weighted GRS, the main effect of $E$ and the interaction effect between $E$ and the GRS, respectively. The test of $\gamma_{INT}$ is asymptotically equivalent to the meta-analysis of $\gamma_{G_i \times E}$, the interaction effects between $G_i$ and $E$, using an inverse-variance weighted sum to derive a 1 *df* chi-square, i.e. (see **Appendix F**, **Figure S7-S8,** and [45]):

$$\left(\frac{\hat{\gamma}_{INT}}{\hat{\sigma}_{\gamma_{INT}}}\right)^2 = \frac{\left(\sum_m \frac{w_i \times \hat{\gamma}_{G_i \times E}}{\hat{\sigma}^2_{\gamma_{G_i \times E}}}\right)^2}{\sum_m \frac{w_i^2}{\hat{\sigma}^2_{\gamma_{G_i \times E}}}}$$

where $w_i$ is the weight given to SNP $i$. A number of strategies can be used for the weighting scheme. In an agnostic search, assuming the interaction effects are independents of the SNPs characteristics, one should weight each SNP by the inverse of their standard deviation ($w_i = 1/\sigma_{G_i}$). Alternatively, others have use weights proportional to the marginal genetic effect of the SNPs, assuming the magnitude of the marginal and interaction effects are correlated. The relative power of each of these weighting schemes depends on their relevance in regard to the true underlying model. Finally, applying GRS-based interaction tests implicitly supposed a set of candidate genetic variants have been identified. The current rationale consists in assuming that most interacting variants also display a marginal linear effect and therefore have focused on GWAS hits, however other screening methods can be used[46; 47]. Moreover existing knowledge, such as functional annotation[48] or existing pathway database[49] can be leverage to refine the sets of SNPs to be aggregated into a GRS.

## Discussion

Advancing knowledge of how genetic and environmental factors combine to influence human traits and diseases remains a key objective of research in human genetics. Ironically, the simplest and most parsimonious biological interaction models—those in which the effect of a genetic variant is either enhanced or decreased depending on a common exposure—are among the most difficult to identify. Furthermore, the contribution of such interaction effects can be dramatically underestimated when measured as the drop in $r^2$ if the interaction term was removed from the model. Here, I argue for the use of new approaches and analytical strategies to address these concerns. This includes first using methods such as the Pratt index to evaluate the relative importance of interaction effects in genetic association studies. These methods can highlight important modifiable exposures influencing genetic mechanisms, which could be missed with the existing approach. Second, besides increasing sample size, increasing power to detect interaction effects in future studies will mostly rely on leveraging additional assumptions on the underlying model. In the big data era, where millions of genetic variants are measured on behalf of multiple environmental exposures and endo-phenotypes, this means using multivariate models. A variety of powerful statistical tests can be devised assuming multiple environmental exposures interact with multiple genetic variants. As showed in this study, the application of such approaches can dramatically improve power to detect interaction than standard univariate tests. While these methods comes at the cost of decreased precision—i.e. a significant signal would point out multiple potential culprit—they can identify interaction effects that would potentially be of greater clinical relevance that univariate pairwise interaction[14; 15]. The application of these methods in genetic association studies offers great opportunities for moving the field forward and providing a new perspective on interaction effects in human traits and diseases.

## Appendices

### *Appendix A: Effect estimates from standardized and unstandardized predictors*

Following the notation defined from the main text, the outcome $Y$ can be expressed as:

$$
\begin{aligned}
Y &= \beta_G \times G + \beta_E \times E + \beta_{GE} \times G \times E + \varepsilon \\
&= \beta_G \times (G_{std}\sigma_G + \mu_G) + \beta_E \times (E_{std}\sigma_E + \mu_E) + \beta_{GE} \times (G_{std}\sigma_G + \mu_G) \times (E_{std}\sigma_E + \mu_E) + \varepsilon \\
&= (\beta_G\sigma_G + \beta_{GE}\mu_E\sigma_G) \times G_{std} + (\beta_E\sigma_E + \beta_{GE}\mu_G\sigma_E) \times E_{std} + (\beta_{GE}\sigma_G\sigma_E) \times G_{std} \times E_{std} + \varepsilon'
\end{aligned}
$$

where $\varepsilon'$ is a term that depends neither on $G_{std}$ nor $E_{std}$. This leads to the following relationship between the standardized and unstandardized estimates:

$$\beta'_G = \beta_G \times \sigma_G + \beta_{GE} \times \mu_E \times \sigma_G \quad \Leftrightarrow \quad \beta_G = \frac{\beta'_G}{\sigma_G} - \frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}$$

$$\beta'_E = \beta_E \times \sigma_E + \beta_{GE} \times \mu_G \times \sigma_E \quad \Leftrightarrow \quad \beta_E = \frac{\beta'_E}{\sigma_E} - \frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}$$

$$\beta'_{GE} = \beta_{GE} \times \sigma_E \times \sigma_G \quad \Leftrightarrow \quad \beta_{GE} = \frac{\beta'_{GE}}{\sigma_E \times \sigma_G}$$

The variances of the unstandardized estimates equal:

$$\sigma^2_{\beta_G} = var\left(\frac{\beta'_G}{\sigma_G} - \frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}\right) = \frac{\sigma^2_{\beta'_G}}{\sigma^2_G} + \frac{\sigma^2_{\beta'_{GE}} \times \mu^2_E}{\sigma^2_E \times \sigma^2_G} + C_1 * cov(\beta'_G, \beta'_{GE})$$

$$\sigma^2_{\beta_E} = var\left(\frac{\beta'_E}{\sigma_E} - \frac{\beta'_{GE} \times \mu_G}{\sigma_E \times \sigma_G}\right) = \frac{\sigma^2_{\beta'_E}}{\sigma^2_E} + \frac{\sigma^2_{\beta'_{GE}} \times \mu^2_G}{\sigma^2_E \times \sigma^2_G} + C_2 * cov(\beta'_E, \beta'_{GE})$$

$$\sigma^2_{\beta_{GE}} = var\left(\frac{\beta'_{GE}}{\sigma_E \times \sigma_G}\right) = \frac{\sigma^2_{\beta'_{GE}}}{\sigma^2_E \times \sigma^2_G}$$

Where $C_1$ and $C_2$ are constants that depend on the mean and variance of $G$ and $E$. **Appendix C** shows that $cov(\beta'_G, \beta'_{GE}) = cov(\beta'_E, \beta'_{GE}) = 0$ when $G$ and $E$ are independents. Moreover, when $G$ and $E$ are standardized, $\sigma^2_{\beta'_G} = \sigma^2_{\beta'_E} = \sigma^2_{\beta'_{GE}} = \frac{\sigma^2}{N}$, where $\sigma^2$ is the residual variance of $Y$. Assuming the main effects of $G$ and $E$ and their interaction is small, so that $\sigma^2 \approx 1$, the variance of the estimates simplify:

$$\sigma_{\beta_G}^2 = \frac{\mu_E^2 + \sigma_E^2}{N \times \sigma_E^2 \times \sigma_G^2}$$

$$\sigma_{\beta_E}^2 = \frac{\mu_G^2 + \sigma_G^2}{N \times \sigma_E^2 \times \sigma_G^2}$$

$$\sigma_{\beta_{GE}}^2 = \frac{1}{N \times \sigma_E^2 \times \sigma_G^2}$$

### *Appendix B: Non-centrality parameters for marginal and interaction models*

Using the estimates and variances from **Appendix A** one can derive $ncp_G$, $ncp_E$, and $ncp_{GE}$, the non-centrality parameters (ncp) of the genetic main effect, the exposure main effect and the interaction effect under the assumptions of small effect sizes and $G - E$ independence:

$$ncp_G = \frac{\beta_G^2}{\sigma_{\beta_G}^2} = \frac{\beta_G^2}{\frac{\mu_E^2 + \sigma_E^2}{N \times \sigma_E^2 \times \sigma_G^2}} = N \times \sigma_G^2 \times \beta_G^2 \times \frac{\sigma_E^2}{\mu_E^2 + \sigma_E^2}$$

$$ncp_E = \frac{\beta_E^2}{\sigma_{\beta_E}^2} = \frac{\beta_E^2}{\frac{\mu_G^2 + \sigma_G^2}{N \times \sigma_G^2 \times \sigma_E^2}} = N \times \sigma_E^2 \times \beta_E^2 \times \frac{\sigma_G^2}{\mu_G^2 + \sigma_G^2}$$

$$ncp_{GE} = \frac{\beta_{GE}^2}{\sigma_{\beta_{GE}}^2} = \frac{\beta_{GE}^2}{\frac{1}{N \times \sigma_E^2 \times \sigma_G^2}} = N \times \sigma_E^2 \times \sigma_G^2 \times \beta_{GE}^2$$

These $ncp$ can be compared with $ncp_{mG}$, the non-centrality parameter from the test of $G$ in a marginal model. The marginal effect of $G$, $\beta_{mG}$ is by definition the sum of the main effect of $G$ plus the marginal contribution from interaction terms involving $G$. It can be approximate by:

$$\beta_{mG} = \frac{cov(Y, G)}{\sigma_G^2} = \frac{\beta_G' \times \sigma_G}{\sigma_G^2} = \beta_G + \beta_{GE} \times \mu_E$$

The marginal estimated effect of $E$ can be derived similarly and equals:

$$\beta_{mE} = \beta_E + \beta_{GE} \times \mu_G$$

so that $ncp_{mG}$ and $ncp_{mE}$ (the *ncp* of the marginal test of $E$) can be expressed as follows:

$$ncp_{mG} = N \times \sigma_G^2 \times (\beta_G + \beta_{GE} \times \mu_E)^2$$

$$ncp_{mE} = N \times \sigma_E^2 \times (\beta_E + \beta_{GE} \times \mu_G)^2$$

### *Appendix C: Variance-covariance for the GxE term and its estimated effect*

To derive the covariance and the correlation parameter between $G$ and $G \times E$ we first calculate $\sigma_{GE}^2$, the variance of the interaction term $G \times E$ under the assumption of independence between $G$ and $E$. Assuming the standard coding for $G$ [0,1,2], and the frequency of the coded allele is $p$, and $E$ is normally distributed so that $E^2$ follows a non-central chi-square distribution with one degree of freedom, it can be express as:

$$\sigma_{GE}^2 = \mathrm{E}[G^2] \times \mathrm{E}[E^2] - \mathrm{E}[G]^2 \times \mathrm{E}[E]^2 = \sum_{G \in Range(G)} \left(G^2 \times pr(G)\right) \times \int_{\mathbb{R}} E^2 dP - (2 \times p)^2 \times \mu_E^2$$

$$= ((2 \times p \times (1-p)) + p^2 \times 4) \times \sigma_E^2 \times \left(1 + \frac{\mu_E^2}{\sigma_E^2}\right) - (2 \times p)^2 \times \mu_E^2$$

$$= 2 \times (p + p^2) \times (\sigma_E^2 + \mu_E^2) - 4 \times p^2 \times \mu_E^2$$

$$= \sigma_G^2 \times \sigma_E^2 + \mu_G^2 \times \sigma_E^2 + \mu_E^2 \times \sigma_G^2$$

Under the same assumption, one can derive $cov(G, G \times E)$, the covariance between $G$ and $G \times E$:

$$cov(G, G \times E) = \mathrm{E}[G^2] \times \mathrm{E}[E] - \mathrm{E}[G] \times \mathrm{E}[G] \times \mathrm{E}[E]$$

$$= (2 \times p \times (1-p) + p^2 \times 4) \times \mu_E - (2 \times p)^2 \times \mu_E$$

$$= \mu_E \times \sigma_G^2$$

Similarly, one can derive the covariance between the exposure and the interaction and show that $cov(E, G \times E) = \mu_G \times \sigma_E^2$. From this it appears that $cov(G, G \times E_{std}) = cov(E, G_{std} \times E) = cov(G_{std}, G_{std} \times E_{std}) = 0$.

The correlation between $G$ and $G \times E$ equals then:

$$cor(G, G \times E) = \frac{cov(G, G \times E)}{\sigma_G \times \sigma_{GE}} = \frac{\mu_E \times \sigma_G^2}{\sigma_G \times \sigma_{GE}} = \frac{\mu_E}{\sqrt{\sigma_E^2 \times \left(1 + \frac{\mu_G^2}{\sigma_G^2}\right) + \mu_E^2}}$$

We derive then the covariance and correlation between the estimated effect of $G$ and $G \times E$. In its general form the variance-covariance matrix of estimates from the interaction model can be obtained using its matrix formulation: $\sigma_{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\sigma^2$, where $\mathbf{X}$, the matrix of predictor variables, equals $[1, G, E, G \times E]$ and $\sigma^2$ is the variance of the residual of $Y$.

$$\sigma_{\boldsymbol{\beta}} = \left( N \times \begin{bmatrix} \mathrm{E}[1^2] & \mathrm{E}[G] & \mathrm{E}[E] & \mathrm{E}[G \times E] \\ \mathrm{E}[G] & \mathrm{E}[G^2] & \mathrm{E}[G \times E] & \mathrm{E}[G^2 \times \mathrm{E}] \\ \mathrm{E}[E] & \mathrm{E}[G \times E] & \mathrm{E}[E^2] & \mathrm{E}[G \times E^2] \\ \mathrm{E}[G \times E] & \mathrm{E}[G^2 \times \mathrm{E}] & \mathrm{E}[G \times E^2] & \mathrm{E}[G^2 \times E^2] \end{bmatrix} \right)^{-1} \times \sigma^2$$

This is a relatively complex form, however when the predictors are standardized $\mathrm{E}[E] = \mathrm{E}[G] = 0$, and assuming $G$ and $E$ are independents, the formulation of $\sigma_{\boldsymbol{\beta}}$ greatly simplify, as all the off-diagonal elements of the matrix are null, so that:

$$\sigma_{\boldsymbol{\beta}} = \frac{\sigma^2}{N} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \dfrac{1}{\mathrm{E}[G^2]} & 0 & 0 \\ 0 & 0 & \dfrac{1}{\mathrm{E}[E^2]} & 0 \\ 0 & 0 & 0 & \dfrac{1}{\mathrm{E}[G^2] \times \mathrm{E}[E^2]} \end{bmatrix}$$

which implies that $cov(\beta'_G, \beta'_{GE}) = 0$. Building on this, and using the equations from **Appendix A**, we can derive $\gamma$, the covariance between $\beta_G$ and $\beta_{GE}$:

$$\gamma = cov(\beta_G, \beta_{GE}) = cov\left(\frac{\beta'_G}{\sigma_G} - \beta_{GE} \times \mu_E, \beta_{GE}\right) = \frac{cov(\beta'_G, \beta'_{GE})}{\sigma_E \times \sigma_G^2} - \mu_E \times \sigma^2_{\beta_{GE}} = \frac{-\mu_E}{N \times \sigma_E^2 \times \sigma_G^2}$$

The correlation follows:

$$cor(\beta_G, \beta_{GE}) = \frac{cov(\beta_G, \beta_{GE})}{\sigma_{\beta_G} \times \sigma_{\beta_{GE}}} = \frac{-\mu_E \times \sqrt{\dfrac{1}{N \times \sigma_E^2 \times \sigma_G^2}}}{\sqrt{\dfrac{\mu_E^2 + \sigma_E^2}{N \times \sigma_E^2 \times \sigma_G^2}}} = \frac{-\mu_E}{\sqrt{\mu_E^2 + \sigma_E^2}}$$

### *Appendix D: Derivation of the Pratt index*

To estimate the variance explained by predictors or other related measures, we first derive the expected variance of the outcome for a given generative model. For a single interaction term and assuming $G - E$ independence, it equals:

$$\begin{aligned}var(Y) &= var(\beta_E \times E + \beta_G \times G + \beta_{GE} \times G \times E + \varepsilon)\\ &= \beta_E^2 \times \sigma_E^2 + \beta_G^2 \times \sigma_G^2 + \beta_{GE}^2 \times \sigma_{GE}^2 + 2 \times \beta_{GE} \times \big(\beta_G \times cov(G, GE) + \ \beta_E \times cov(E, GE)\big) + \sigma^2\\ &= (\beta_G \times \sigma_G + \beta_{GE} \times \mu_E \times \sigma_G)^2 + (\beta_E \times \sigma_E + \beta_{GE} \times \mu_G \times \sigma_E)^2 + (\beta_{GE} \times \sigma_E \times \sigma_G)^2 + \sigma^2\\ &= \beta_G'^2 + \beta_E'^2 + \beta_{GE}'^2 + \sigma^2\end{aligned}$$

When more interaction terms are included, the outcome variance becomes a little more complex as additional covariance terms are added. For example assuming $k$ interactions between $E$ and $G_i$, $i = 1 \dots k$, the variance of $Y$ becomes:

$$\begin{aligned}var(Y) = \beta_E^2 \times \sigma_E^2 + \sum_i \left[\beta_{G_i}^2 \times \sigma_{G_i}^2\right] + \sum_i \left[\beta_{G_iE}^2 \times \sigma_{G_iE}^2\right] + 2 \times \sum_i \left[\beta_{G_iE} \times \beta_{G_i} \times \mu_E \times \sigma_{G_i}^2\right] + 2\\ \times \sum_i \beta_{G_iE} \times \beta_E \times \mu_{G_i} \times \sigma_E^2 + \sum_i \sum_{j \neq i} \left[cov\big(G_i \times E, G_j \times E\big)\right] + \sigma^2\end{aligned}$$

$$\begin{aligned}= \beta_E^2 \times \sigma_E^2 + \sum_i \left[\beta_{G_i}^2 \times \sigma_{G_i}^2\right] + \sum_i \left[\beta_{G_iE}^2 \times \sigma_{G_iE}^2\right] + 2 \times \sum_i \left[\beta_{G_iE} \times \beta_{G_i} \times \mu_E \times \sigma_{G_i}^2\right] + 2\\ \times \sum_i \beta_{G_iE} \times \beta_E \times \mu_{G_i} \times \sigma_E^2 + \sum_i \sum_{j \neq i} \left[\beta_{G_iE} \times \beta_{G_jE} \times \mu_{G_i} \times \mu_{G_j} \times \sigma_E^2\right] + \sigma^2\end{aligned}$$

For simplicity let us assume $\sigma^2$ is set so that $var(Y) = 1$ in all further derivation. When testing a single interaction term and using the equivalences from **Appendix A-B** one can show that the Pratt index can be expressed as a function of the estimates from the interaction model and the mean and variance of the genetic variant and the exposures considered:

$$r_G^{2*} = (\beta_G \times \sigma_G) \times cor(Y, G) = (\beta_G \times \sigma_G) \times \beta_{mG} \ \times \sigma_G = \beta_G^2 \times \sigma_G^2 + \beta_G \times \beta_{GE} \times \sigma_G^2 \times \mu_E$$

$$r_E^{2*} = (\beta_E \times \sigma_E) \times cor(Y, E) = (\beta_E \times \sigma_E) \times (\beta_E + \beta_{GE} \times \mu_G) \times \sigma_E = \beta_E^2 \times \sigma_E^2 + \beta_E \times \beta_{GE} \times \sigma_E^2 \times \mu_G$$

$$\begin{aligned}r_{GE}^{2*} &= (\beta_{GE} \times \sigma_{GE}) \times cor(Y, G \times E)\\ &= (\beta_{GE} \times \sigma_{GE}) \times \frac{cov(\beta_G \times G + \beta_E \times E + \beta_{GE} \times G \times E + \varepsilon, G \times E)}{\sigma_{GE}}\end{aligned}$$

$$= \beta_{GE} \times (\beta_G \times \mu_E \times \sigma_G^2 + \beta_E \times \mu_G \times \sigma_E^2 + \beta_{GE} \times \sigma_{GE}^2)$$

$$= \beta_{GE}^2 \times (\mu_G^2 \times \sigma_E^2 + \mu_E^2 \times \sigma_G^2 + \sigma_E^2 \times \sigma_G^2) + \beta_{GE} \times (\beta_G \times \mu_E \times \sigma_G^2 + \beta_E \times \mu_G \times \sigma_E^2)$$

When summing the above Pratt index we obtain:

$$r_G^{2*} + r_E^{2*} + r_{GE}^{2*} = \beta_G^2 \times \sigma_G^2 + \beta_G \times \beta_{GE} \times \sigma_G^2 \times \mu_E + \beta_E^2 \times \sigma_E^2 + \beta_E \times \beta_{GE} \times \sigma_E^2 \times \mu_G + \beta_{GE}^2 \times \mu_G^2 \times \sigma_E^2$$
$$+ \beta_{GE}^2 \times \mu_E^2 \times \sigma_G^2 + \beta_{GE}^2 \times \sigma_E^2 \times \sigma_G^2 + \beta_{GE} \times (\beta_G \times \mu_E \times \sigma_G^2 + \beta_E \times \mu_G \times \sigma_E^2)$$

$$= \beta_G^2 \times \sigma_G^2 + 2 \times (\beta_G \times \sigma_G) \times (\beta_{GE} \times \mu_E \times \sigma_G) + (\beta_{GE} \times \mu_E \times \sigma_G)^2 + \beta_E^2 \times \sigma_E^2 + 2$$
$$\times (\beta_E \times \sigma_E) \times (\beta_{GE} \times \mu_G \times \sigma_E) + (\beta_{GE} \times \mu_G \times \sigma_E)^2 + (\beta_{GE} \times \sigma_E \times \sigma_G)^2$$

$$= (\beta_G \times \sigma_G + \beta_{GE} \times \mu_E \times \sigma_G)^2 \ + (\beta_E \times \sigma_E + \beta_{GE} \times \mu_G \times \sigma_E)^2 + (\beta_{GE} \times \sigma_E \times \sigma_G)^2$$

$$= r_G^2 + r_E^2 + r_{GE}^2$$

The cumulative contribution of multiple interactions involving independent SNPS can also be derived from summary statistics, although the derivation is a little less friendly because of additional covariance terms. For example assuming $k$ interactions between $E$ and $G_i$, $i = 1 \ldots k$, we obtain (**Figure S4**) :

$$r_G^{2*} = \sum_i \left[\beta_{G_i}^2 \times \sigma_{G_i}^2 + \beta_{G_i} \times \beta_{G_iE} \times \sigma_{G_i}^2 \times \mu_E\right]$$

$$r_E^{2*} = \beta_E^2 \times \sigma_E^2 + \sum_i \left[\beta_E \times \beta_{G_iE} \times \sigma_E^2 \times \mu_{G_i}\right]$$

$$r_{GE}^{2*} = \sum_i \left[\beta_{G_iE}^2 \times \sigma_{G_iE}^2\right] + \sum_i \left[\beta_{G_iE} \times \beta_{G_i} \times \mu_E \times \sigma_{G_i}^2\right] + \sum_i \left[\beta_{G_iE} \times \beta_E \times \mu_{G_i} \times \sigma_E^2\right]$$
$$+ \sum_i \left[\beta_{G_iE} \times \sum_{j \neq i} \left[\beta_{G_jE} \times \mu_{G_i} \times \mu_{G_j} \times \sigma_E^2\right]\right]$$

On should note that estimating the Pratt index for the exposure can be difficult in practice when the number of interaction is large, as it would require the estimated main exposure effect from a joint model including all SNPs main effect and all interactions term with the exposure. Also, because of the correlation between main and interaction terms, $r_{X_i}^{2*}$, as the standard $r_{X_i}^2$, only approximate the amount the variance will change if $X_i$ was held constant. For the latter measure, one can refer to[21].

### *Appendix E: Joint test of main and interaction effects*

The multiple regression least square provides the estimated effect of the genetic main effect and interaction effects $\boldsymbol{\beta} = (\beta_G, \beta_{GE})$ and their variance-covariance matrix $\boldsymbol{\Sigma}$**.** The multivariate Wald test of the two parameters, which follow a 2 *df* chi-square can be expressed as:

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = \begin{bmatrix}\beta_G & \beta_{GE}\end{bmatrix}\begin{bmatrix}\sigma^2_{\beta_G} & \gamma \\ \gamma & \sigma^2_{\beta_{GE}}\end{bmatrix}^{-1}\begin{bmatrix}\beta_G \\ \beta_{GE}\end{bmatrix},$$

where $\gamma$ is the covariance between $\beta_G$ and $\beta_{GE}$. It can be further developed as:

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = \begin{bmatrix}\beta_G & \beta_{GE}\end{bmatrix} \times \frac{1}{\sigma^2_{\beta_G} \times \sigma^2_{\beta_{GE}} - \gamma^2}\begin{bmatrix}\sigma^2_{\beta_{GE}} & -\gamma \\ -\gamma & \sigma^2_{\beta_G}\end{bmatrix}\begin{bmatrix}\beta_G \\ \beta_{GE}\end{bmatrix}$$

$$= \frac{1}{\sigma^2_{\beta_G} \times \sigma^2_{\beta_{GE}} - \gamma^2} \times \begin{bmatrix}\sigma^2_{\beta_{GE}} \times \beta_G - \gamma \times \beta_{GE} & -\gamma \times \beta_G + \sigma^2_{\beta_G} \times \beta_{GE}\end{bmatrix}\begin{bmatrix}\beta_G \\ \beta_{GE}\end{bmatrix}$$

$$= \frac{\left(\sigma^2_{\beta_{GE}} \times \beta_G - \gamma \times \beta_{GE}\right) \times \beta_G + \left(\sigma^2_{\beta_G} \times \beta_{GE} - \gamma \times \beta_G\right) \times \beta_{GE}}{\sigma^2_{\beta_G} \times \sigma^2_{\beta_{GE}} - \gamma^2}$$

$$= \frac{\sigma^2_{\beta_{GE}} \times \beta_G^2 + \sigma^2_{\beta_G} \times \beta_{GE}^2 - 2 \times \gamma \times \beta_G \times \beta_{GE}}{\sigma^2_{\beta_G} \times \sigma^2_{\beta_{GE}} - \gamma^2}$$

For clarity we derived the nominator and the denominator separately, so that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = A/B$

$$A = \hat{\sigma}^2_{\beta_{GE}} \times \left(\frac{\beta'_G}{\sigma_G} - \frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}\right)^2 + \hat{\sigma}^2_{\beta_G} \times \left(\frac{\beta'_{GE}}{\sigma_E \times \sigma_G}\right)^2 + 2 \times \mu_E \sigma^2_{\beta_{GE}} \times \left(\frac{\beta'_G}{\sigma_G} - \frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}\right) \times \left(\frac{\beta'_{GE}}{\sigma_E \times \sigma_G}\right)$$

$$= \hat{\sigma}^2_{\beta_{GE}} \times \left(\left(\frac{\beta'_G}{\sigma_G}\right)^2 - \left(\frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}\right)^2\right) + \hat{\sigma}^2_{\beta_G} \times \left(\frac{\beta'_{GE}}{\sigma_E \times \sigma_G}\right)^2$$

$$= \frac{\left(\frac{\beta'_G}{\sigma_G}\right)^2 - \left(\frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}\right)^2 + \left(\frac{\beta'_{GE} \times \mu_E}{\sigma_E \times \sigma_G}\right)^2 + \left(\frac{\beta'_{GE} \times \sigma_E}{\sigma_E \times \sigma_G}\right)^2}{N \times \sigma_E^2 \times \sigma_G^2}$$

$$= \frac{\left(\frac{\beta'_G}{\sigma_G}\right)^2 + \left(\frac{\beta'_{GE}}{\sigma_G}\right)^2}{N \times \sigma_E^2 \times \sigma_G^2}$$

The denominator $B$ equals:

$$B = \hat{\sigma}^2_{\beta_G} \times \hat{\sigma}^2_{\beta_{GE}} - \left(-\mu_E \times \sigma^2_{\beta_{GE}}\right)^2 = \frac{\sigma_E^2}{(N \times \sigma_E^2 \times \sigma_G^2)^2}$$

So that the joint test of $G$ and $G \times E$ effects equals:

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = N \times \sigma_E^2 \times \frac{\left(\frac{\beta'_G}{\sigma_G}\right)^2 + \left(\frac{\beta'_{GE}}{\sigma_G}\right)^2}{\sigma_E^2 \times \sigma_G^2} = N \times \beta'^2_G + N \times \beta'^2_{GE}$$

which is the sum of the individuals Wald test for the main effect and the interaction effect when $G$ and $E$ are standardized. Moreover, leveraging previous equivalences, we can express the joint test as a function of $\beta''_G$ and $\beta''^2_{GE}$, the estimated main and interaction effects from the model where $E$ has been centered, so the test can be further expressed as:

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = N \times \beta''^2_G \times \sigma_G^2 + N \times \beta''^2_{GE} \times \sigma_G^2 \times \sigma_E^2$$

### *Appendix F: GRS-based test, joint test and univariate test of multiple interaction effects*

We denote $\boldsymbol{\beta} = (\beta_{G_1}, \beta_{G_2}, \dots \beta_{G_m})$ a vector of effects from $m$ independent SNP, and $\sigma^2_{\beta_{G_i}}$ and $w_i$ are the variance of each estimate and weight of each SNP $i$ in the genetic risk score (GRS), respectively. The effect of the weighted GRS on the outcome, $\gamma_{GRS}$, equals:

$$\gamma_{GRS} = \frac{cov(Y, \mathrm{GRS})}{\sigma^2_{GRS}} = \frac{cov(Y, \sum_m [w_i \times G_i])}{\sigma^2_{\sum_m [w_i \times G_i]}} = \frac{\sum_m \left(w_i \times \beta_{G_i} \times \sigma^2_{G_i}\right)}{\sum_m \left[w_i^2 \times \sigma^2_{G_i}\right]} = \frac{\sum_m \frac{w_i \times \beta_{G_i}}{\sigma^2_{\beta_{G_i}}}}{\sum_m \frac{w_i^2}{\sigma^2_{\beta_{G_i}}}}$$

Consecutively, $\sigma_{\gamma_{GRS}}$ the variance of $\gamma_{GRS}$ can be derived as follows:

$$\sigma^2_{\gamma_{GRS}} = var\left(\frac{\sum_m \frac{w_i \times \beta_{G_i}}{\sigma^2_{\beta_{G_i}}}}{\sum_m \frac{w_i^2}{\sigma^2_{\beta_{G_i}}}}\right) = \sum\nolimits_m \left(\left(\frac{\frac{w_i}{\sigma^2_{\beta_{G_i}}}}{\sum_m \frac{w_i^2}{\sigma^2_{\beta_{G_i}}}}\right)^2 \times var(\beta_{G_i})\right) = \frac{\sum_m \left(\frac{w_i^2}{\sigma^2_{\beta_{G_i}}}\right)}{\left(\sum_m \frac{w_i^2}{\sigma^2_{\beta_{G_i}}}\right)^2} = \frac{1}{\sum_m \frac{w_i^2}{\sigma^2_{\beta_{G_i}}}}$$

So that the chi-square of the marginal effect of the test of GRS equals:

$$\left(\frac{\gamma_{GRS}}{\sigma_{\gamma_{GRS}}}\right)^2 = \frac{\left(\sum_m \frac{w_i \times \beta_{G_i}}{\sigma^2_{\beta_{G_i}}}\right)^2}{\sum_m \frac{w_i^2}{\sigma^2_{\beta_{G_i}}}}$$

which corresponds to the inverse-variance weighted sum meta-analysis of each individual genetic variant. Similarly, one can derive the expected chi-square of the GRS by exposure interaction effect using $\gamma_{G_i \times E}$, the interaction effect between each SNP $i$ and the exposure. Under the assumption of independence of the *m* interaction terms we obtain:

$$\left(\frac{\gamma_{INT}}{\sigma_{\gamma_{INT}}}\right)^2 = \frac{\left(\sum_m \frac{\gamma_{G_i \times E}}{\sigma^2_{\gamma_{G_i \times E}}}\right)^2}{\sum_m \frac{1}{\sigma^2_{\gamma_{G_i \times E}}}}$$

Hence for standardized $G$ and $E$ the *ncp* of the $GRS$ by $E$ interaction test equals $\mathrm{ncp}_{\mathrm{GRS \times E}} = N \times \left(\sum_{\mathrm{i=1 \ldots m}} \gamma'_{G_i \times E}\right)^2 / m$, where *N* is the sample size and $\gamma'_{G_i \times E}$ is the interaction effect from the standardized model, and follows a chi-square with one degree of freedom. In comparison, the *ncp* for the test of the strongest pairwise interaction, i.e. the interaction that explained the largest amount of variance, equals $\mathrm{ncp}_{\mathrm{pairwise}} = \max\left(N \times \gamma'_{G_i \times E}\right)$.

## Supplemental Data

Supplemental Data include eight figures.

## Acknowledgements

I am grateful to Peter Kraft, Noah Zaitlen, Ami Joshi, John Pratt and Donald Halstead for helpful discussions and comments. This research was funded by NIH grants R03 HG006720.

## Web Resources

All simulations were conducted using the R sofware. http://www.r-project.org/

## Reference

1. Cordell, H.J. (2009). Detecting gene-gene interactions that underlie human diseases. Nature reviews Genetics 10, 392-404.
2. Aschard, H., Lutz, S., Maus, B., Duell, E.J., Fingerlin, T.E., Chatterjee, N., Kraft, P., and Van Steen, K. (2012). Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Human genetics 131, 1591-1613.
3. Thomas, D. (2010). Gene--environment-wide association studies: emerging approaches. Nature reviews Genetics 11, 259-272.
4. Gauderman, W.J., Zhang, P., Morrison, J.L., and Lewinger, J.P. (2013). Finding novel genes by testing G x E interactions in a genome-wide association study. Genetic epidemiology 37, 603-613.
5. Hutter, C.M., Mechanic, L.E., Chatterjee, N., Kraft, P., Gillanders, E.M., and Tank, N.C.I.G.-E.T. (2013). Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. Genetic epidemiology 37, 643-657.
6. Thomas, D. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annual review of public health 31, 21-36.
7. Bookman, E.B., McAllister, K., Gillanders, E., Wanke, K., Balshaw, D., Rutter, J., Reedy, J., Shaughnessy, D., Agurs-Collins, T., Paltoo, D., et al. (2011). Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop. Genetic epidemiology.

8. Greenland, S. (1983). Tests for interaction in epidemiologic studies: a review and a study of power. Statistics in medicine 2, 243-251.

9. Aiken, L.S., West, S.G., and Reno, R.R. (1991). Multiple regression: Testing and interpreting interactions.(Newbury Park, CA: Sage).

10. Friedrich, J.R. (1982). In Defense of Multiplicative Terms in Multiple Regression Equations. American Journal of Political Science 26797-833.

11. Farrar, D.E., and Glauber, R.R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics 49, 92-107.

12. Weiss, N.S. (2008). Subgroup-specific associations in the face of overall null results: should we rush in or fear to tread? Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 17, 1297-1299.

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

14. Aschard, H., Chen, J., Cornelis, M.C., Chibnik, L.B., Karlson, E.W., and Kraft, P. (2012). Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. American journal of human genetics 90, 962-972.

15. Aschard, H., Zaitlen, N., Lindstrom, S., and Kraft, P. (2015). Variation in Predictive Ability of Common Genetic Variants by Established Strata: The Example of Breast Cancer and Age. Epidemiology 26, 51-58.

16. Lewontin, R.C. (1974). Annotation: the analysis of variance and the analysis of causes. American journal of human genetics 26, 400-411.

17. Darlington, R.B. (1968). Multiple regression in psychological research and practice. Psychological bulletin 69, 161-182.

18. Green, P.E., D., C.J., and DeSarbo, W.S. (1978). A New Measure of Predictor Variable Importance in Multiple Regression. Journal of Marketing Research 15, 356-360.

19. Budescu, D.V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. Psychological bulletin 114, 542-551.

20. Chao, Y.C., Zhao, Y., Kupper, L.L., and Nylander-French, L.A. (2008). Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. Journal of occupational and environmental hygiene 5, 519-529.

21. Pratt, J.W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In Proceedings of the Second International Conference in Statistics, T.P.a.S. Puntanen, ed. (Tampere, Finland), pp 245-260.

22. Bring, J. (1996). A Geometric Approach to Compare Variables in a Regression Model. The American Statistician 50, 57-62.

23. Thomas, D.R., Hughes, E., and Zumbo, B.D. (1998). On Variable Importance in Linear Regression. Social Indicators Research 45, 253-275

24. Mukherjee, B., Ahn, J., Gruber, S.B., and Chatterjee, N. (2012). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. American journal of epidemiology 175, 177-190.

25. Dai, J.Y., Kooperberg, C., Leblanc, M., and Prentice, R.L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. Biometrika 99, 929-944.

26. Kraft, P., Yen, Y.C., Stram, D.O., Morrison, J., and Gauderman, W.J. (2007). Exploiting gene-environment interaction to detect genetic associations. Human heredity 63, 111-119.

27. Hancock, D.B., Artigas, M.S., Gharib, S.A., Henry, A., Manichaikul, A., Ramasamy, A., Loth, D.W., Imboden, M., Koch, B., McArdle, W.L., et al. (2012). Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. PLoS genetics 8, e1003098.

28. Clayton, D., and McKeigue, P.M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358, 1356-1360.

29. Aschard, H., Hancock, D.B., London, S.J., and Kraft, P. (2010). Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. Human heredity 70, 292-300.

30. Manning, A.K., Hivert, M.F., Scott, R.A., Grimsby, J.L., Bouatia-Naji, N., Chen, H., Rybin, D., Liu, C.T., Bielak, L.F., Prokopenko, I., et al. (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nature genetics 44, 659-669.

31. Hamza, T.H., Chen, H., Hill-Burns, E.M., Rhodes, S.L., Montimurro, J., Kay, D.M., Tenesa, A., Kusel, V.I., Sheehan, P., Eaaswarkhanth, M., et al. (2011). Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS genetics 7, e1002237.

32. Manning, A.K., LaValley, M., Liu, C.T., Rice, K., An, P., Liu, Y., Miljkovic, I., Rasmussen-Torvik, L., Harris, T.B., Province, M.A., et al. (2011). Meta-analysis of gene-environment interaction: joint

estimation of SNP and SNP x environment regression coefficients. Genetic epidemiology 35, 11-18.

33. Zhu, X., Feng, T., Tayo, B.O., Liang, J., Young, J.H., Franceschini, N., Smith, J.A., Yanek, L.R., Sun, Y.V., Edwards, T.L., et al. (2015). Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. American journal of human genetics 96, 21-36.
34. Pollin, T.I., Isakova, T., Jablonski, K.A., de Bakker, P.I., Taylor, A., McAteer, J., Pan, Q., Horton, E.S., Delahanty, L.M., Altshuler, D., et al. (2012). Genetic modulation of lipid profiles following lifestyle modification or metformin treatment: the Diabetes Prevention Program. PLoS genetics 8, e1002895.
35. Qi, L., Cornelis, M.C., Zhang, C., van Dam, R.M., and Hu, F.B. (2009). Genetic predisposition, Western dietary pattern, and the risk of type 2 diabetes in men. The American journal of clinical nutrition 89, 1453-1458.
36. Ahmad, S., Rukh, G., Varga, T.V., Ali, A., Kurbasic, A., Shungin, D., Ericson, U., Koivula, R.W., Chu, A.Y., Rose, L.M., et al. (2013). Gene x physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry. PLoS genetics 9, e1003607.
37. Qi, Q., Chu, A.Y., Kang, J.H., Jensen, M.K., Curhan, G.C., Pasquale, L.R., Ridker, P.M., Hunter, D.J., Willett, W.C., Rimm, E.B., et al. (2012). Sugar-sweetened beverages and genetic risk of obesity. The New England journal of medicine 367, 1387-1396.
38. Fu, Z., Shrubsole, M.J., Li, G., Smalley, W.E., Hein, D.W., Cai, Q., Ness, R.M., and Zheng, W. (2013). Interaction of cigarette smoking and carcinogen-metabolizing polymorphisms in the risk of colorectal polyps. Carcinogenesis 34, 779-786.
39. Langenberg, C., Sharp, S.J., Franks, P.W., Scott, R.A., Deloukas, P., Forouhi, N.G., Froguel, P., Groop, L.C., Hansen, T., Palla, L., et al. (2014). Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. PLoS medicine 11, e1001647.
40. Ebbeling, C.B., and Ludwig, D.S. (2013). Sugar-sweetened beverages, genetic risk, and obesity. The New England journal of medicine 368, 287.
41. Malavazos, A.E., Briganti, S., and Morricone, L. (2013). Sugar-sweetened beverages, genetic risk, and obesity. The New England journal of medicine 368, 286.
42. Goran, M.I. (2013). Sugar-sweetened beverages, genetic risk, and obesity. The New England journal of medicine 368, 285-286.
43. Greenfield, J.R., Samaras, K., and Campbell, L.V. (2013). Sugar-sweetened beverages, genetic risk, and obesity. The New England journal of medicine 368, 285.

44. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. American journal of human genetics 95, 5-23.

45. Dastani, Z., Hivert, M.F., Timpson, N., Perry, J.R., Yuan, X., Scott, R.A., Henneman, P., Heid, I.M., Kizer, J.R., Lyytikainen, L.P., et al. (2012). Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS genetics 8, e1002607.

46. Aschard, H., Zaitlen, N., Tamimi, R.M., Lindstrom, S., and Kraft, P. (2013). A nonparametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes. Genetic epidemiology 37, 323-333.

47. Pare, G., Cook, N.R., Ridker, P.M., and Chasman, D.I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS genetics 6, e1000981.

48. Consortium, E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636-640.

49. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research 42, D199-205.

**Figure 1. Examples of interaction patterns for a gene by exposure effect on height**

Pattern of contribution of an interaction term to human height when shifting the location of the genetic variant and the exposure. In a) the interaction is defined as the product of centered genetic variant and exposure, while in b) genetic variant and exposure are positive or null.

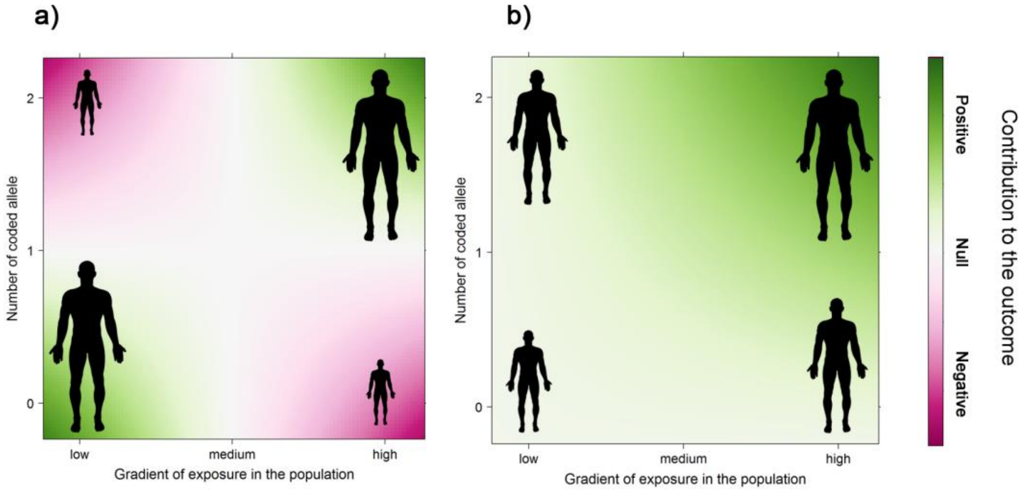


**Figure 2. Relative power of the joint test of main genetic and interaction effects.**

Power comparison for the tests of the main genetic effect (*main.G*), the interaction effect (*int.GxE*) and the joint effect (*Joint G.GxE*) from the interaction model, and the test of the marginal genetic effect (*mar.G*). The outcome $Y$ is define as a function of a genetic variant $G$ coded as [0,1,2] with a minor allele frequency of 0.3, and the interaction of $G$ with an exposure $E$ normally distributed with variance 1 and mean $\bar{E}$. The genetic and interaction effects vary so that they explain 0% and 0.04% (a), 0.1% and 0.1% (b), 0.6% and 0.1% with effect in opposite direction, and 0.4% and 0% (d) of the variance of $Y$, respectively. Power and $\rho_{G,G\times E}$, the correlation between $G$ and the $G \times E$ interaction term (e) were plotted for a sample size of 10,000 individuals and increasing $\bar{E}$ from 0 to 5.

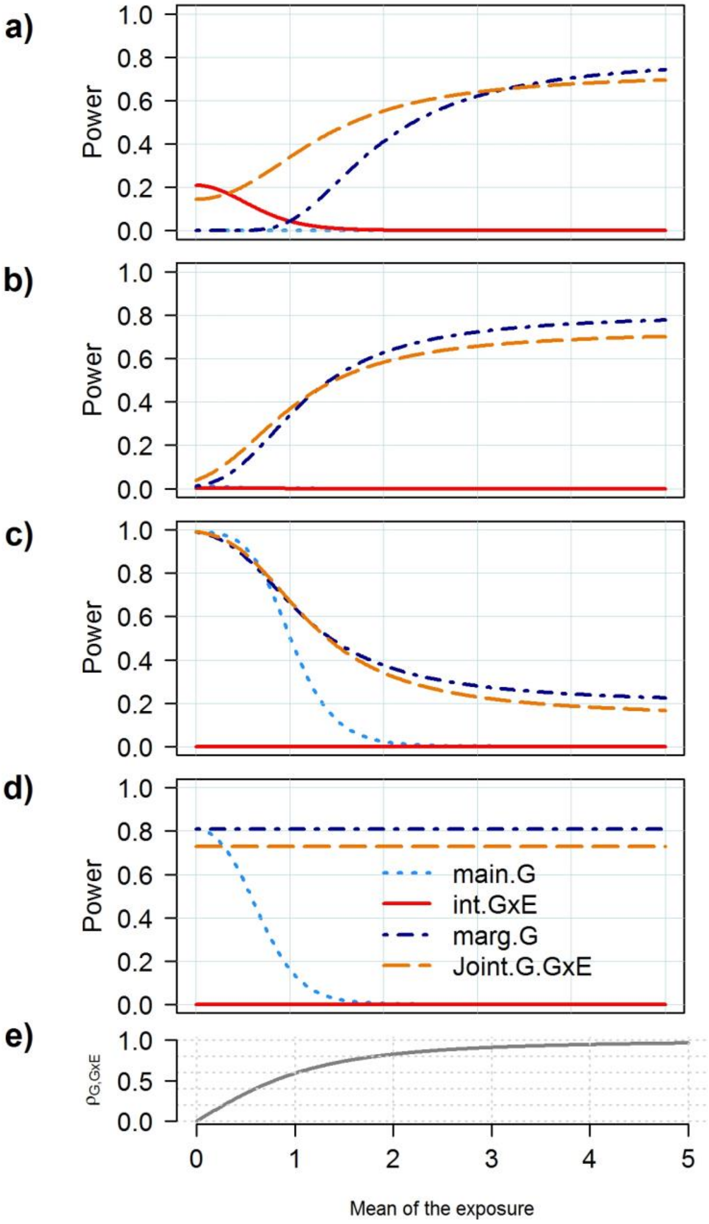


**Figure 3: Examples of attribution of phenotypic variance explained by an interaction effect.**

Proportion of variance of an outcome $Y$ explained by a genetic variant $G$, an exposure $E$ and their interaction $G$x$E$ in a model harboring a pure interaction effect only ($Y = \beta_{GE} \times G \times E + \varepsilon$). The exposure $E$ follows a normal distribution with a standard deviation of 1 and mean of 0 (a), 2 (b) and 4 (c). The genetic variant is biallelic with a risk allele frequency increasing from 0.01 to 0.99. The interaction effect is set so that the maximum of the variance explained by the model equals 1%.

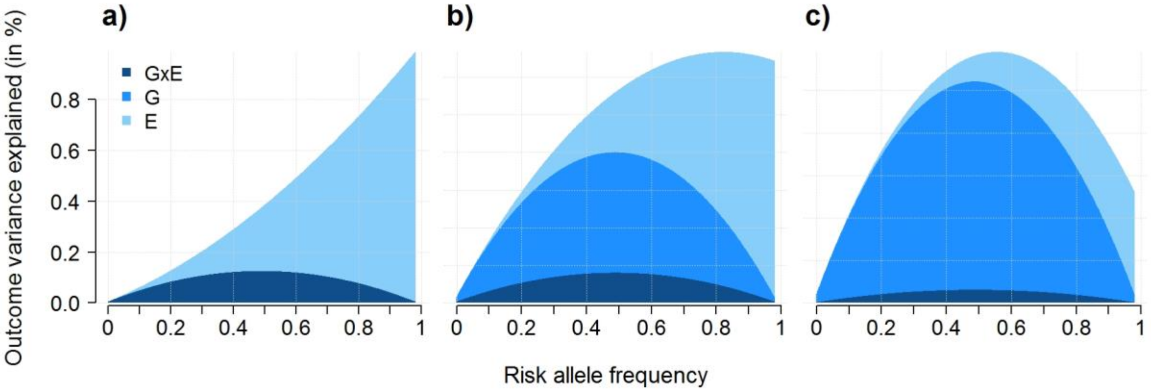


**Figure 4. Importance of an interaction term as defined by the Pratt index.**

Contribution of a genetic variant *G* with minor allele frequency of 0.5, a normally distributed exposure *E* with mean of 4 and variance of 1 and their interaction *G*x*E,* to the variance of a normally distributed outcome *Y,* based on the standard approach –the marginal contribution of *E* and *G* and the increase in $r^2$ when adding the interaction term– (grey boxes), and based on the Pratt index (blue boxes), across 10,000 replicates of 5,000 subjects. For illustration purposes the predictors explain jointly 10% of the variance of *Y*. In scenario a) all G, E and GxE have equals contribution, while in scenario b), c) and d) there was no interaction effect, no exposure effect, and no genetic effect, respectively.

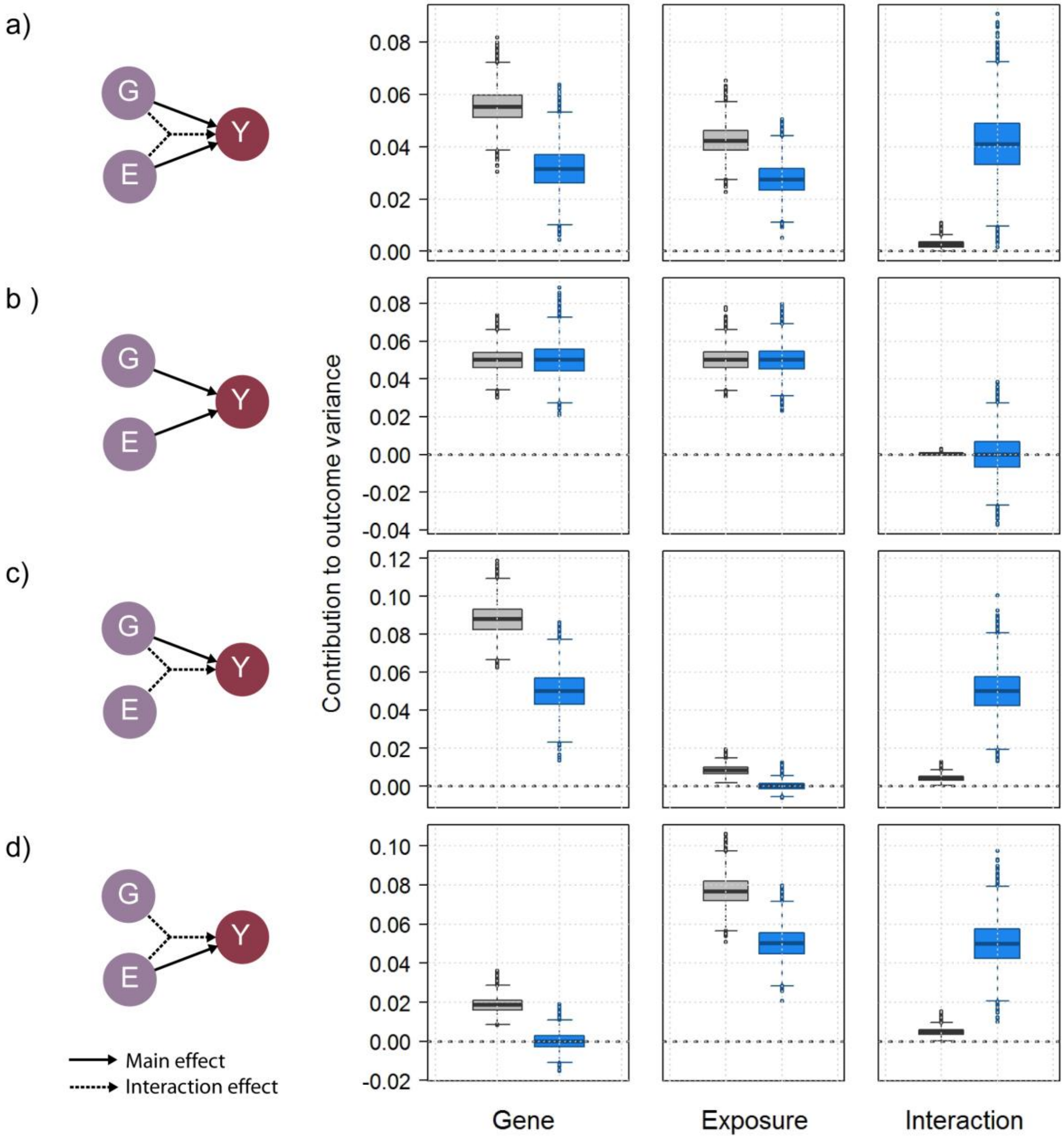
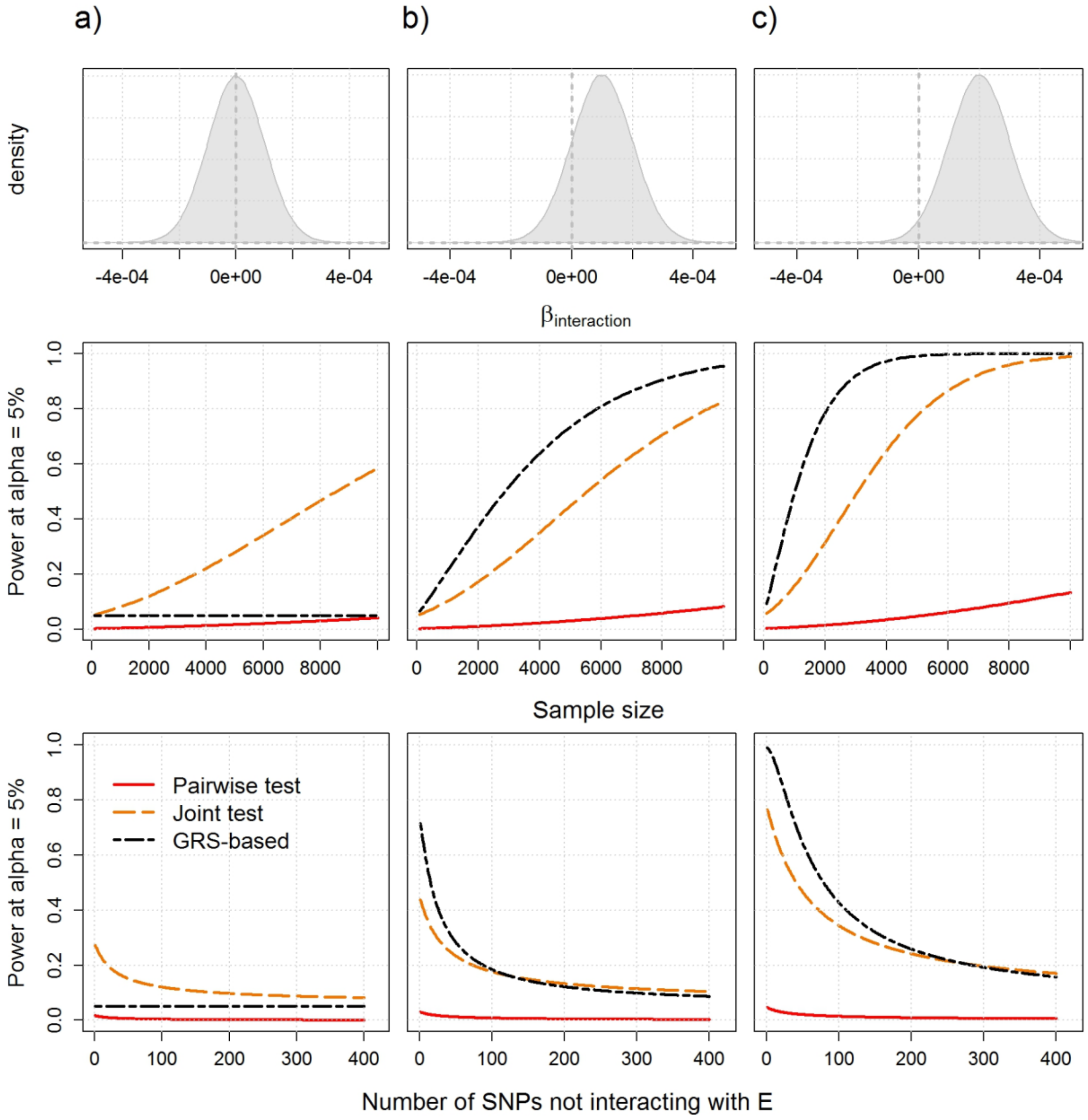


**Figure 5. Advantages and limitation of testing interaction effect with a genetic risk score.**

Examples of power comparison for the combined analysis of interaction effects between 20 SNPs and a single exposure. Power was derived for three scenarios: the interaction effects are normally distributed (upper panels) and (a) centered, (b) slightly positive so that 25% of the interactions are negative, and (c) positive only. Three tests are compared while increasing sample size from 0 to 10,000: the joint test of all interaction terms, the genetic risk score by exposure interaction test, and the test of the strongest interaction effect after correction for the 20 tests performed (middle panels). The lower panels show power of the three tests for a sample size of 5,000, when including 1 to 400 non-interacting SNPs on top of the 20 causal SNPs in the analysis.


a)
b)
c)
density
-4e-04
0e+00
4e-04
$\beta_{interaction}$
Power at alpha = 5%
Sample size
Pairwise test
Joint test
GRS-based
Number of SNPs not interacting with E