Phantom epistasis through the lens of genealogies

Anastasia Ignatieva[1,*] and Lino A. F. Ferreira[1]

[1]Department of Statistics, University of Oxford
[*]Correspondence: anastasia.ignatieva@stats.ox.ac.uk

December 3, 2024

Abstract

Phantom epistasis arises when, in the course of testing for gene-by-gene interactions, the omission of a causal variant (with a purely additive effect on the phenotype) causes the spurious inference of a significant interaction between two SNPs. This is more likely to arise when the two SNPs are in relatively close proximity, so while true epistasis between nearby variants could be commonplace, in practice there is no reliable way of telling apart true epistatic signals from false positives. By considering the causes of phantom epistasis from a genealogy-based perspective, we leverage the rich information contained within reconstructed genealogies (in the form of ancestral recombination graphs) to address this problem. We propose a novel method for explicitly quantifying how likely it is that a given pairwise interaction is the result of phantom epistasis, which can be applied to pairs of SNPs regardless of the genetic distance between them. Our method uses only publicly-available data and so does not require access to the phenotypes and genotypes used for detecting interactions. Using simulations, we show that the method has good sensitivity at even low genetic distances (around 0.5cM), and demonstrate its power to detect phantom epistasis using real data from previous studies. This opens up the exciting possibility of distinguishing spurious interactions in *cis* from those reflecting real biological effects.

1 Introduction

The study of the genetic basis of human traits and diseases has progressed rapidly in recent years. Fuelled by large biobanks, genome-wide association studies (GWASs) have led to the discovery of thousands of genetic variants associated with a wide array of phenotypes (Sollis et al., 2023; Tam et al., 2019), shedding light on their genetic architecture (Lappalainen et al., 2024; Watanabe et al., 2019) and enabling the prediction of disease risk based on genetic data through polygenic scores (PGSs) (Lambert et al., 2021; Torkamani et al., 2018).

The overwhelming majority of genetic associations identified so far involve variation at a single locus (Tam et al., 2019). Comparatively less attention has been given to gene-by-gene (G×G) interactions, or epistasis, between pairs or larger groups of mutations, where the simultaneous presence of mutations at two or more distinct loci has a combined average effect on the phenotype that is different from the sum of their individual effects in isolation (Cordell, 2009; Phillips, 2008). Few such effects have been found (Lappalainen et al., 2024; Wei et al., 2014), and it has been established that epistasis has a limited role in explaining phenotypic variance (Hill et al., 2008; Hivert et al., 2021; Sella and Barton, 2019), so is unlikely to explain a substantial fraction of 'missing heritability' (Manolio et al., 2009), or allow for improved PGS performance. However, detecting a statistical interaction between two parts of the genome suggests a functional interaction between them and so could improve our understanding of the biological function underlying genetic associations (Domingo et al., 2019; Mackay, 2014; Phillips, 2008), a fundamental open problem in the field (Lappalainen et al., 2024).

The search for genetic interactions is made challenging by the vast search space of possible variant combinations, which, for pairwise interactions, grows quadratically in the number of variants considered (Wei et al., 2014). This has raised obstacles along two dimensions. First, exhaustive testing for all possible pairwise combinations of variants has a large computational cost. Second, even if this computational barrier is surmounted, performing a great number of statistical tests

requires a stringent multiple testing correction to avoid rampant false positives, resulting in loss of statistical power; this is exacerbated by such tests having a higher-dimensional parameter space than tests for direct effects. Multiple software tools have been developed that make exhaustive testing for epistasis possible (Upton et al., 2016), however they have not yielded much fruit, likely due to a lack of power. As an alternative, methods that first perform a strategic reduction of the search space by leveraging statistical or biological features of the loci under consideration (Sun et al., 2014) appear more promising, and have enabled the identification of a few robust interactions (GAPC & WTCCC2, 2010; TASC et al., 2011; Masuda et al., 2015; Lyon et al., 2022).

Recently, a novel statistical challenge to the reliable detection of interactions has been recognised and explored. Termed *phantom epistasis*, this phenomenon arises when the statistical evidence for the existence of a genetic interaction fades when the *additive* effect of a third variant is considered. Phantom epistasis was first clearly identified as a problem following a study which examined possible interactions impacting transcription levels in humans: Hemani et al. (2014, retracted) considered the expression of 7,339 genes, measured through RNA sequencing, and detected 30 statistically significant interactions affecting 19 different genes that replicated in two external datasets. However, Wood et al. (2014) showed with different data that none of these interactions remained statistically significant once the most strongly associated variant with the corresponding phenotype was included in the model. A crucial difference between the datasets used in the two studies was sequencing depth: the former study used genotyping array data ($\sim$530,000 single-nucleotide polymorphisms, SNPs) while the latter used a whole-genome sequencing (WGS) dataset. Wood et al.'s findings suggest that the interactions originally reported were, in effect, tagging a third variant whose effect on the phenotype was fully additive (and which was not included in the data of Hemani et al.).

More recently, de los Campos et al. (2019) have demonstrated, using a simplified statistical model of genetic interactions, that phantom epistasis can arise as the result of mutual but imperfect linkage disequilibrium (LD) between the two interacting SNPs and a third, unobserved, causal quantitative trait locus (QTL) which has a fully additive effect. They state that it is not possible to empirically test for this condition, as the QTL genotype is unknown. Hemani et al. (2021) reproduced spurious interactions in simulations and recommend adjusting for fine-mapped additive effects when testing for epistasis. Thus, the challenge posed by phantom epistasis remains unsolved, with the best solution currently available being to carefully account for observed additive effects, with no guarantee that this will be sufficient to avoid false-positive interactions (since this still ignores the possible presence of unobserved causal QTLs, for instance structural variants or SNPs appearing in repetitive regions which are difficult to sequence).

In this article, we develop a method that can provide evidence against the existence of an unobserved additive effect which could explain a putative interaction. Our key observation is that, while we cannot by definition observe a missing genetic variant, it is possible to calculate a proxy genotype with which any such unobserved QTL would have to be highly correlated for phantom epistasis to emerge, and to search for evidence of its presence – or of its absence – within a genealogical framework. We do this by considering the conditions for phantom epistasis in terms of conditions on the existence of clades of samples within the underlying genealogy. For instance, if the set of samples carrying both of the detected interacting SNPs itself forms a clade within the genealogy, and a variant shared by these samples is present but not included in the statistical analysis, the interaction term will act as a proxy for this unobserved additive effect variant (likely resulting in a false positive in a test for interactions).

We present a computational tool (Spectre: Searching for Phantom EpistatiC interactions using TREes) implementing two methods of searching for evidence of phantom epistasis, which can be applied to summary statistics from association studies conducted using large-scale genotype data, and genealogies (in the form of ancestral recombination graphs, ARGs) reconstructed using a WGS dataset; we use the 1000 Genomes Project (1KGP, 1000 Genomes Project Consortium, 2015). We proceed by calculating genotype proxies for unobserved causal (additive effect) QTLs, and then (1) searching for clades in the genealogy that are highly correlated with these proxies, and (2) searching for evidence *against* the existence of such highly correlated clades. We show that, as previously

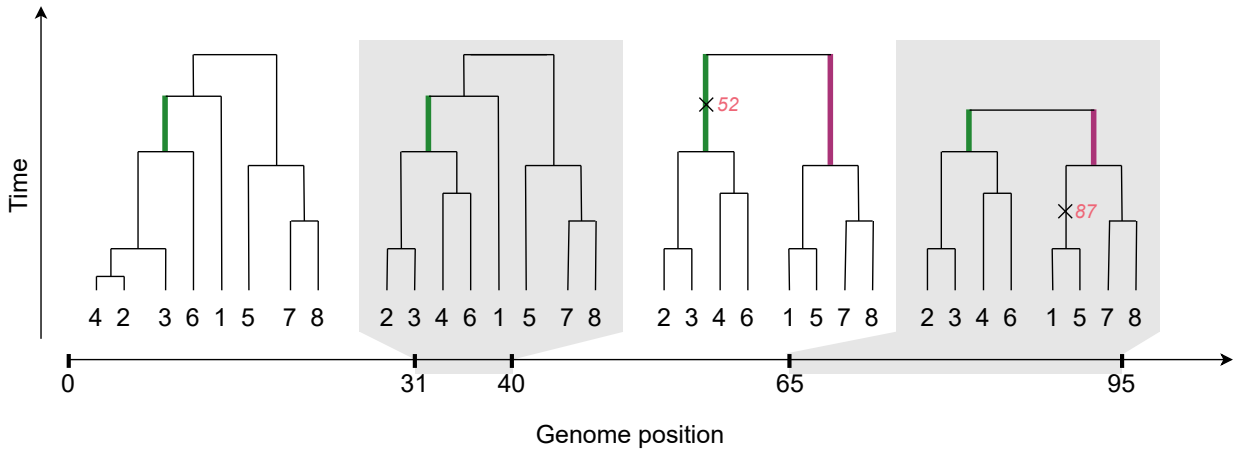

Figure 1: Example of an ARG with $n = 8$ samples, with four local trees shown covering the genomic region $[0, 95]$. The set of samples $A = \{2, 3, 4, 6\}$ forms a clade $\widehat{A}$ within the entire interval $[0, 95]$ (since there is a branch in each local tree subtending exactly this set of samples, shown in green); this clade is supported by one mutation at position 52 (shown as $\times$). The set of samples $B = \{1, 5, 7, 8\}$ forms a clade $\widehat{B}$ within the interval $[40, 95]$ so has a genomic span of 55bp (in local trees 3 and 4, the branches subtending this clade are shown in purple); this clade is not supported by any mutations. The set $C = \{1, 3, 4\}$, for instance, does not form a clade anywhere within the shown region.

noted, phantom epistasis is relatively likely to arise when the two SNPs are in close proximity. However, our method allows to explicitly test for the presence of phantom epistasis even for SNPs in *cis*, without requiring an arbitrary cut-off on the genetic distance between them as has been suggested elsewhere in the literature. We use simulation studies to demonstrate the effectiveness of these methods, apply them to the interactions identified by Hemani et al. (2014), which were shown to be false positives by Wood et al. (2014), and test a known interaction between variants in the *ABO* and *FUT2* genes affecting alkaline phosphatase levels (which has been replicated in multiple studies). Our work presents the first practical computational tool for assessing the likelihood that detected genetic interactions are due to phantom epistasis.

Code implementing the methods is available at `github.com/a-ignatieva/spectre`.

2 Methods

2.1 Overview of ARGs

The genealogy of a sample of sequences can be fully encoded in the form of an ARG, which can be represented as a series of local trees, describing the local genealogy at each locus (for recent overviews, see for instance Brandt et al., 2024; Lewanski et al., 2024; Wong et al., 2024). Adjacent local trees are highly correlated, as edges tend to persist across multiple local trees before being broken up by recombination. A set of samples can form a clade in the ARG if at some genomic position the set is subtended by a branch of the corresponding local tree. Each clade has a well-defined genomic span, being an interval including all the genomic positions where, in the corresponding local tree, there is a branch subtending exactly the set of samples in the clade. This is illustrated in Figure 1. ARGs reconstructed from a given sequencing dataset thus provide not only an estimate of the age of each mutation, but also an estimate of the genomic span of the clade of its carriers. For any given ARG, it is thus possible to traverse the local trees left-to-right, and record the genomic span of each encountered clade of samples.

Each clade and mutation within the ARG uniquely defines a corresponding set of samples (those belonging to the clade, or carrying the mutation, respectively). A given clade $\widehat{Z}$ containing the set of samples Z, in turn, defines a corresponding (haploid) genotype vector $\boldsymbol{z} = (z_1, \ldots, z_N)$, with

$z_i = 1$ if $i \in Z$ and $z_i = 0$ otherwise. Likewise, a given genotype vector $\boldsymbol{v}$ uniquely defines a coresponding set of samples $V = \{i : v_i = 1\}$, but it is not necessarily the case that this set forms a clade $\widehat{V}$ within the ARG.

2.2 Methods for detecting epistasis

The issue of phantom epistasis concerns only *statistical epistasis*, that is, statistical interaction terms between genetic variables in (parametric) linear models of phenotypic architecture (Phillips, 2008). By construction, these effects are measured while averaging over all other genotypes in a population. There are two widely-used models of phenotypic architecture that include epistatic effects. Consider first the simplest linear model of a quantitative phenotype $\boldsymbol{y}$ with a single genetic interaction between two loci (with genotype vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$):

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i}, \tag{2.1}$$

where i indexes the samples. The canonical test for a genetic interaction is to fit a linear regression model and perform a t-test of significance for β_{12} (Cordell, 2009) (using logistic regression when the phenotype is binary). A more parsimonious model separates additive from dominance effects and allows interactions between these two classes of effects:

$$y_i = \beta_1 x_{1i} + \beta_{1d} \mathbb{I}_{x_{1i}} + \beta_2 x_{2i} + \beta_{2d} \mathbb{I}_{x_{2i}} + \beta_{12} x_{1i} x_{2i} + \beta_{1d2} \mathbb{I}_{x_{1i}} x_{2i} + \beta_{12d} x_{1i} \mathbb{I}_{x_{2i}} + \beta_{1d2d} \mathbb{I}_{x_{1i}} \mathbb{I}_{x_{2i}},$$

where $\mathbb{I}_x = 1$ if $x = 1$ and 0 otherwise. Under this model, we can test for the existence of an epistatic effect of any of these four types (additive-by-additive, additive-by-dominance, etc.) at once through an F-test with four degrees of freedom. This is the test performed by Hemani et al. (2014) using the computational implementation of Hemani et al. (2011).

Hemani et al. (2021) show through simulation that the latter test is prone to phantom epistasis and state that the same is true of the t-test approach, which is the model de los Campos et al. (2019) considered in their work. Given the high-dimensionality of the search space and the fact that very few interactions have been identified using human genetic data, we explore the issue of phantom epistasis using the simpler model (2.1). This approach makes tractable the derivation of a key quantity used in our method.

2.3 Phantom epistasis

Consider testing for an interaction between SNP1 and SNP2 (at positions m_1 and m_2, with genotype vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and sample sets X_1 and X_2, respectively) within the linear regression model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i, \tag{2.2}$$

where $\boldsymbol{\varepsilon}$ is an error term. Assuming that neither the SNPs nor their interaction has a causal effect on the phenotype (such that $\beta_1 = \beta_2 = \beta_{12} = 0$), it may nevertheless be the case that our estimate of the coefficient β_{12} is statistically significant if a third SNP which does have a causal effect on $\boldsymbol{y}$ is correlated with the interaction term (this follows from standard linear regression results, and a detailed derivation is given in SI, Section S1.3).

Comparing this setting to a standard GWAS where the individual effects of SNPs on a phenotype are tested, there is both a similarity and an important difference worth noting. Whenever we test for the direct effect of a variant on a phenotype, it is understood that a statistically significant effect does not necessarily imply that the variant in question has a causal effect on the phenotype; rather, we recognise that it may be merely correlated with such a causal effect in its vicinity, a phenomenon known as *tagging*, and that fine-mapping is required to be confident that the true causal variant has been identified. In a sense, a similar claim could be made about a statistically significant interaction. The crucial difference is that, if a third variant with an *additive* effect explains an apparent interaction signal, this changes the nature of the putative interaction effect being reported. The claim that a statistical interaction has an effect on a phenotype implicitly

implies that this effect cannot be fully accounted for by any combination of simple additive effects, and that a fundamentally more complex phenomenon is at work (in other words, a purely additive model is not a fully accurate statistical representation of the genetic basis of this phenotype). If it turns out that the additive effect of a third variant can explain the interaction signal, this is no longer true.

To begin examining this issue from a genealogy-based perspective, let us assume a haploid model for simplicity and consider the following example matrix of genotypes:

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{s}$	$\boldsymbol{z}$
0	0	0	0
1	0	0	0
1	1	1	1
1	0	0	0
0	0	0	0
1	1	1	1
0	1	0	0
0	1	0	0

where $\boldsymbol{s} := \boldsymbol{x}_1 \circ \boldsymbol{x}_2$ ('$\circ$' denoting element-wise product) represents the simultaneous occurrence of the two mutations, and $\boldsymbol{z}$ is a third mutation that happens to be perfectly correlated with $\boldsymbol{s}$. If $\boldsymbol{z}$ is the true causal variant for the phenotype but only $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{s}$ are included in the model, $\boldsymbol{s}$ will likely appear statistically significant as it is perfectly correlated with the true (additive) causal effect, when in fact no epistasis is present.

We note that variants like $\boldsymbol{z}$ are not the only kind that can lead to spurious interactions. Suppose that a significant interaction is detected between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Then define the four *target sets* as

$$\begin{aligned} S &= \{i : i \in X_1 \text{ and } i \in X_2\}, \\ K_1 &= \{i : i \in X_1 \text{ and } i \notin X_2\}, \\ K_2 &= \{i : i \notin X_1 \text{ and } i \in X_2\}, \\ O &= \{i : i \notin X_1 \text{ and } i \notin X_2\}. \end{aligned}$$

Phantom epistasis can arise if any of these sets forms a clade in the ARG (at *any* position along the genome), but no SNP corresponding to this clade is included in the model (implying that a causal variant shared by the samples in this clade exists but has not been sequenced). This is because inclusion of such a SNP would mean that the interaction term can be represented by a linear combination of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and the genotype vector of this SNP. For instance, consider the following matrix corresponding to the ARG in Figure 2:

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{s}$	$\boldsymbol{k}_1$	$\boldsymbol{k}_2$	$\boldsymbol{o}$
0	0	0	0	0	1
1	0	0	1	0	0
1	1	1	0	0	0
1	0	0	1	0	0
0	0	0	0	0	1
1	1	1	0	0	0
0	1	0	0	1	0
0	1	0	0	1	0

The interaction term can be written equivalently as

$$\boldsymbol{x}_1 \circ \boldsymbol{x}_2 = \boldsymbol{s} = \boldsymbol{x}_1 - \boldsymbol{k}_1 = \boldsymbol{x}_2 - \boldsymbol{k}_2 = \boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{o} - \mathbf{1}$$

(where $\mathbf{1}$ is a vector of 1's), so inclusion of any of $\boldsymbol{s}, \boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{o}$ would result in the model being unidentifiable. On the other hand, if the genotype of the true causal variant is highly correlated

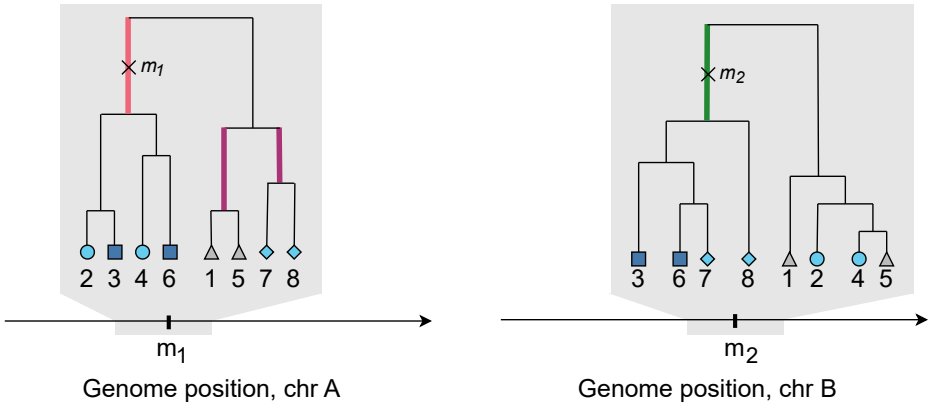

Figure 2: The two mutation events, SNP1 at m_1 and SNP2 at m_2, are shown as $\times$ on the corresponding branches (in red and green) of the local trees at m_1 and m_2, respectively. Here $X_1 = \{2, 3, 4, 6\}$, $X_2 = \{3, 6, 7, 8\}$, and the corresponding target sets are $S = \{3, 6\}$ (dark blue squares), $K_1 = \{2, 4\}$ (light blue circles), $K_2 = \{7, 8\}$ (light blue diamonds), $O = \{1, 5\}$ (grey triangles). In these two trees the set S does not form a clade, since there is no branch that subtends only the samples in S; the same is true for set K_1. The sets K_2 and O form clades in the local tree at m_1; branches subtending these clades are shown in purple. Thus, if the causal mutation occurs on one of the purple branches but is not sequenced, phantom epistasis could arise and result in a significant interaction term between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

with $\boldsymbol{s}, \boldsymbol{k}_1, \boldsymbol{k}_2$ or $\boldsymbol{o}$, omitting this variant from the model will cause spurious inference of a significant interaction effect.

Phantom epistasis can thus arise with the existence of any clade $\widehat{Z}$ such that the corresponding set of samples Z is highly enough correlated (positively or negatively) with a target set. Note that, theoretically, it does not matter where along the genome $\widehat{Z}$ arises. Under the coalescent model, the probability that a set of $|Z|$ samples forms a clade in a random local tree of size n can be easily calculated (SI, Section S1.2); this is negligible unless $|Z|$ is small, or the required level of correlation with a target set is low. We can thus, in practice, ignore the probability of $\widehat{Z}$ arising at a random position along the genome by chance. However, this probability is elevated at genomic positions near m_1 and near m_2: since $S = X_1 \cap X_2$ is a subset of X_1, for instance, near m_1 we need to instead consider the probability that Z forms a sub-clade of a clade of size $|X_1|$, which is higher than the probability that Z forms a clade in a tree of size n. Note that this is true even when m_1 and m_2 are on different chromosomes. When m_1 and m_2 are close together, however, we know that S is a subset of X_1 and a subset of X_2, which both form clades in the region $[m_1, m_2]$, so the probability of interest will be further elevated. This reasoning aligns with phantom epistasis being more likely to affect SNPs in *cis*; it further implies that, for each target set, we only need to consider the probability of a highly correlated clade existing at genomic positions near m_1 and m_2 (whether the interactions are between SNPs in *cis* or in *trans*).

2.4 Genealogy-based testing for phantom epistasis

Typically, interaction tests are performed using large (unphased) genetic datasets, and hence it is difficult to calculate the target sets for each pair of tested SNPs as above, and not possible to directly estimate the probability that they form clades. We thus take the following approach. For each pair of SNPs with a significant interaction term in an association study, we calculate the corresponding target sets using a WGS dataset, in particular, the 1KGP (subsetted to the $n = 1006$ samples from European populations, to match the association testing data). We then use the ARG reconstructed for this data with Relate (Speidel et al., 2019) to quantify the likelihood of phantom epistasis, using two approaches. We first directly search for clades that are highly correlated with one of the target sets, which would imply the significant interaction term may be the result of omitting an additive effect variant from the genotype data. Then, we use the ARGs to quantify

how *unlikely* it is that such a highly correlated clade exists somewhere along the genome.

2.4.1 Test 1: Searching for additive effects not included in the model

Our first method of checking for phantom epistasis is by using the 1KGP ARG to check for clades (within ±1cM of each SNP) which, if included in the model, can cause the interaction term to become significant (we refer to these as *significant clades*). We construct a test for whether a clade satisfies this condition, as detailed in SI, Sections S1.3 and S1.4. Briefly, building on the results of de los Campos et al. (2019) and using the regression model (2.2), we derive an explicit inequality that a clade $\widehat{Z}$ must satisfy for the interaction term to become significant if the variant shared by this clade (which is assumed to have a true additive effect) is omitted from the model:

$$V_1 \cdot \mathbb{E}(x_{1i}z_i) + V_2 \cdot \mathbb{E}(x_{2i}z_i) + V_3 \cdot \mathbb{E}(g_i z_i) \geq \frac{\Phi^{-1}(1-\alpha/2)}{b}\sqrt{\frac{1-h^2}{N}} =: F, \tag{2.3}$$

where $\boldsymbol{g}$ is the genotype vector corresponding to one of the target sets ($\boldsymbol{s}, \boldsymbol{k}_1, \boldsymbol{k}_2$ or $\boldsymbol{o}$), V_1, V_2, V_3 are terms involving the variances and covariances of $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{g}$, α is the significance threshold and N the sample size in the test for interactions, b is the true (additive) effect size and h^2 is the trait heritability. Since we cannot know or reliably estimate b, the threshold F cannot be readily obtained. To solve this problem, we make use of the fact that we know the interaction term in the association study has a significant effect size, so $\boldsymbol{s}$ itself *must* satisfy this inequality, giving an upper bound on F. In practice, we approximate this upper bound by substituting $\boldsymbol{s}$ for both $\boldsymbol{g}$ and $\boldsymbol{z}$ into the left hand side and taking the resulting threshold F to be this value (multiplied by a factor of 0.8 to be conservative).

2.4.2 Test 2: Quantifying evidence against existence of unobserved additive effects

While this approach will identify observed sequenced variants (and those potentially shared by samples in inferred clades) which can be the cause of spuriously inferred significant interactions, it does not rule out the possibility that such a variant exists. This can be the case if, for instance, the causal variant is a structural variant, or is present in a repeat-heavy region which is difficult to sequence, or simply due to the unavoidable presence of topological uncertainty in ARG reconstruction (Hayman et al., 2023).

We thus propose a method to consider each clade in the ARG, and, within its genomic span, quantify the evidence it provides *against* the possibility of an omitted variant that causes the interaction term to have a significant effect size. To illustrate the basic idea, notice that in Figure 2, the fact that the set $\{2, 3\}$ forms a clade within the genomic span of the local tree at m_1 provides evidence against the possibility of the set $\{3, 6\}$ forming a clade within this span. Using similar reasoning, for each target set G, we consider in turn each clade $\widehat{A}$ of the ARG (within ±1cM of each SNP), and check whether the existence of $\widehat{A}$ disproves the existence of all other possible clades $\widehat{Z}$ that are highly correlated with G and satisfy the inequality (2.3) (SI, Section S1.5). This provides evidence against the existence of any such clade $\widehat{Z}$ within the genomic span of $\widehat{A}$. We can thus, at each genomic position, calculate the number of clades in the ARG that provide evidence against the existence of such significant clades. In regions where this number is zero (which we call *uncovered regions*), we cannot rule out that a significant clade exists, and hence the total span of uncovered regions quantifies the probability of phantom epistasis.

3 Results

3.1 Simulation studies

We first simulated interactions by randomly selecting two SNPs from the 1KGP data and assuming that they interact. Note that we do not simulate phantom epistasis, where by design an apparent interaction would be explained by a true additive effect. Instead, we examine how

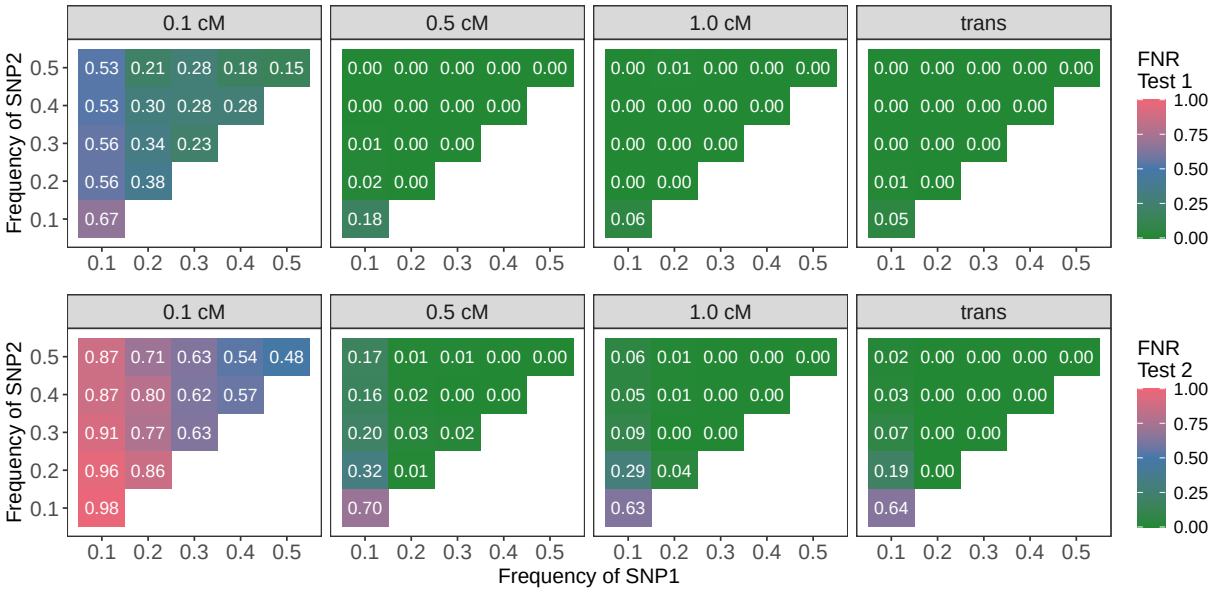

Figure 3: Simulation results using 1KGP ARG with European samples ($n = 1006$). Panels are labelled by the distance between the two SNPs. Top row: proportion of cases where at least one significant clade is identified by Test 1. Bottom row: proportion of cases where span of uncovered regions identified by Test 2 is greater than 0. Each tile is averaged over 100 simulations. This, in essence, quantifies how likely it is that a true interaction would be statistically indistinguishable from phantom epistasis.

the presence of variants highly correlated with a target set can in some cases make it impossible to identify a true interaction as such, since the possibility of phantom epistasis cannot be ruled out. We chose this approach over simulating data and phenotypes and performing association and interaction testing, specifically to avoid confounding the results by the particular choice of epistasis detection method and parameters controlling genetic architecture. We considered the regions 1cM to the left and to the right of each SNP using the 1KGP ARG, and applied the two methods described in Section 2.4. We varied the frequency of SNP1 and SNP2 (requiring that each target set has a frequency of at least 0.5%) and simulated SNPs in *cis* and in *trans* (varying the genetic distance between them in the former case), as detailed in SI, Section S1.7.

The results (Figure 3) show that, as expected, when the two SNPs are in *cis* and in close proximity, with high probability a clade exists which satisfies (2.3), so inclusion of the corresponding genotype in the model could explain away the significant interaction effect. This highlights the important point that it is not possible to state in these cases that the significant interaction is a 'false positive' as such, but rather just that it is not possible to distinguish whether epistatic or additive effects are present (due to the fundamental unidentifiability of the model when a variant exists whose set of carriers is highly correlated with a target set). As expected, this decays relatively quickly with increasing genetic distance, and (unless the two SNPs have small sample sets, relative to the size of the ARG), the probability of significant clades arising by chance is relatively low even for SNPs that are 0.5cM apart. If one or both SNPs have relatively low frequency, however, phantom epistasis can arise even if they are on different chromosomes, through the existence of significant clades by random chance near one of the SNPs.

The results also demonstrate, importantly, that our method can provide convincing evidence against an interaction being significant due to phantom epistasis, even for SNPs that are moderately close together (as long as their frequency is not very low). This provides a way of checking whether detected interactions in *cis* are likely to correspond to true epistatic effects.

SNP1 (Hemani et al.)	SNP2 (Hemani et al.)	Distance (cM)	Uncovered regions (bp)	Significant clades	Supporting SNPs	Top SNP (This paper)
10:76446305 (0.7)	10:75929517 (0.6)	0.1	489,945	289	5	10:76248814
21:48027084 (0.6)	21:47764477 (0.2)	0.5	346,482	238	56	21:47732271
17:80890638 (0.6)	17:80827903 (0.6)	0.1	312,469	1975	273	17:80893992
3:152234166 (0.3)	3:152116652 (0.5)	0.0	295,500	172	7	3:152249794
11:88117962 (0.2)	11:88077479 (0.5)	0.0	253,085	510	99	11:88070563
1:203877662 (0.4)	1:203780591 (0.7)	0.0	231,351	527	166	1:203805538
19:36268923 (0.4)	19:36147315 (0.8)	0.1	213,617	309	36	19:36237245
8:144663661 (0.6)	8:144613680 (0.4)	0.1	140,054	455	57	8:144629876
21:45230974 (0.8)	21:45198355 (0.7)	0.0	77,301	547	81	21:45246723
18:74747424 (0.5)	18:74732087 (0.5)	0.0	34,856	125	1	18:74723459
21:48063862 (0.2)	21:47776382 (0.2)	0.6	18,299	6	3	21:48057719
19:19810050 (0.3)	19:19738554 (0.5)	0.1	1,623	0	0	-
7:94648239 (0.5)	3:152187431 (0.2)	-	0	0	0	-
6:114067127 (0.5)	3:152187431 (0.2)	-	0	0	0	-
5:22101322 (0.4)	3:152187431 (0.2)	-	0	0	0	-

Table 1: Results for tested interacting SNPs from Hemani et al. (2014). Columns 1 and 2: GRCh37 positions, with frequencies given in brackets. Column 3: total span of uncovered regions within ±1cM of the two SNPs. Column 4: number of significant clades detected. Column 5: number of SNPs supporting the significant clades. Column 6: supporting SNP giving highest observed value of LHS of (2.3).

3.2 Interactions from Hemani et al. (2014)

In order to examine the power of our method applied to real data, we consider the interactions reported by Hemani et al. (2014), which were detected using data from 846 individuals. These were, through additional sequencing of 450 individuals, shown to be the result of phantom epistasis by Wood et al. (2014). We consider the 15 (out of 30) interactions that were replicated by Wood et al. at a nominal 0.05 significance threshold (since our method is not designed to discern other causes of false positives, which can be assumed to have resulted in the lack of replication).

An example of the results for two of the interactions is shown in Figure 4. At increasing distance away from the two SNPs, in both cases there is a relatively large number of clades which suggest that a significant clade is unlikely to exist (shown on the y-axis). In both cases, however, there are uncovered regions around m_1 and m_2 (shown in purple), where the number of such clades drops to zero, suggesting that we cannot rule out phantom epistasis. Indeed, in both cases there is a large number of clades (many supported by SNPs) which are highly correlated with one or more of the target sets and satisfy the significance inequality (2.3) (shown in light blue and green, respectively).

A summary of the full results is presented in Table 1. Of the 15 interactions tested, in 12 cases we detected the presence of uncovered regions, and in 11 cases the presence of significant clades (supported by corresponding SNPs). All of the three cases where we do not identify uncovered regions involve *cis–trans* interactions with a variant in the *MBNL1* gene, which Wood et al. (2014) note occurs in the probe sequence used to measure expression levels (positing that these interactions could be the result of a technical artefact).

3.3 *ABO–FUT2* interaction

We next consider the known *ABO–FUT2* interaction between variants rs635634 (9:136155000) and rs281379 (19:49214274) affecting levels of alkaline phosphatase circulating in blood serum, identified by Masuda et al. (2015) and replicated using UK Biobank data by Lyon et al. (2022). This interaction corresponds to a known biological pathway: the *ABO* variant affects the determination of blood group, and the *FUT2* variant the determination of ABH secretor status; the interaction corresponds to a differential effect of blood antigens on alkaline phosphatase levels depending on secretor status. Applying our method to these SNPs, we do not find evidence of phantom epistasis, with no significant clades or uncovered regions identified (Figure 5).

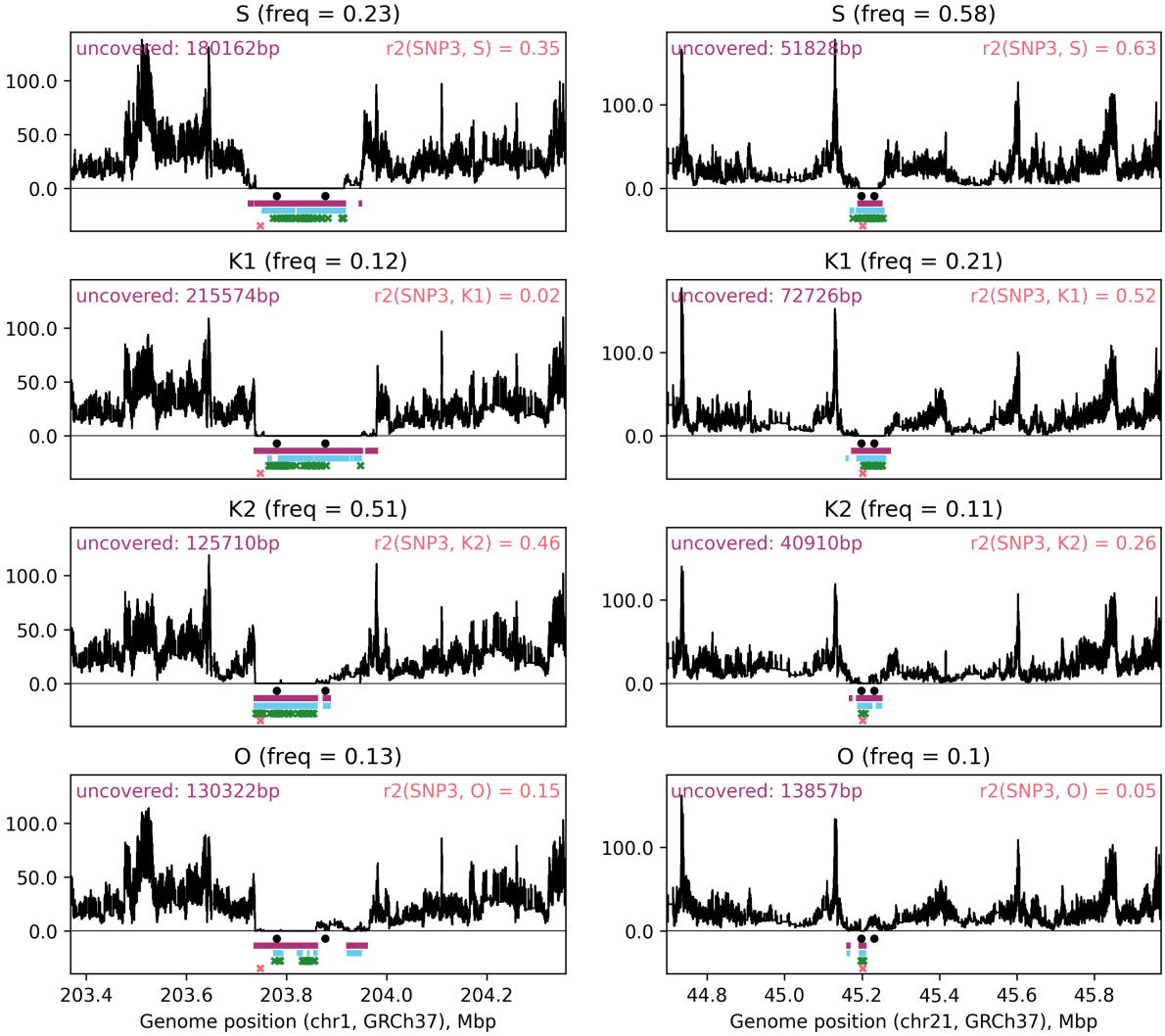

Figure 4: Two examples of results for Hemani et al. (2014) interactions between SNPs in *cis*. Black points show positions of SNP1 and SNP2. Genomic span of significant clades (using Test 1) shown in light blue; positions of SNPs that support these clades (that is, occur on a branch subtending the clade in the corresponding local tree) shown in green. At each genomic position within the interval $[m_1 - 1\text{cM}, m_2 + 1\text{cM}]$, y-axis shows number of clades providing evidence against the existence of a significant clade; uncovered regions are shown in purple (using Test 2). Position of SNP3 from Wood et al. (2014) is shown in red (its r^2 with each target set given in top right corner).

4 Discussion

The solutions for addressing phantom epistasis suggested in the literature so far typically involve including fine-mapped variants as covariates in the model. This, however, does not fully resolve the problem, as it still leaves open the possibility that another (unsequenced) additive-effect causal variant exists. Our method is the first approach, to our knowledge, that quantifies the probability of this, by formulating the issue of phantom epistasis from the point of view of genealogies, and leveraging the extra information contained in reconstructed ARGs: specifically, inferred clades and their estimated genomic spans. Such estimates can be sensitive to errors and artefacts in ARG reconstruction (Ignatieva et al., 2024), so the accuracy of the method will improve in tandem with the rapidly ongoing developments in ARG reconstruction methodology.

By formally deriving the conditions that a clade must satisfy to cause phantom epistasis, we again show that it is not enough to ensure that LD between the two tested SNPs is below some threshold, or that they are far apart in genetic distance. A limitation of our approach, constrained

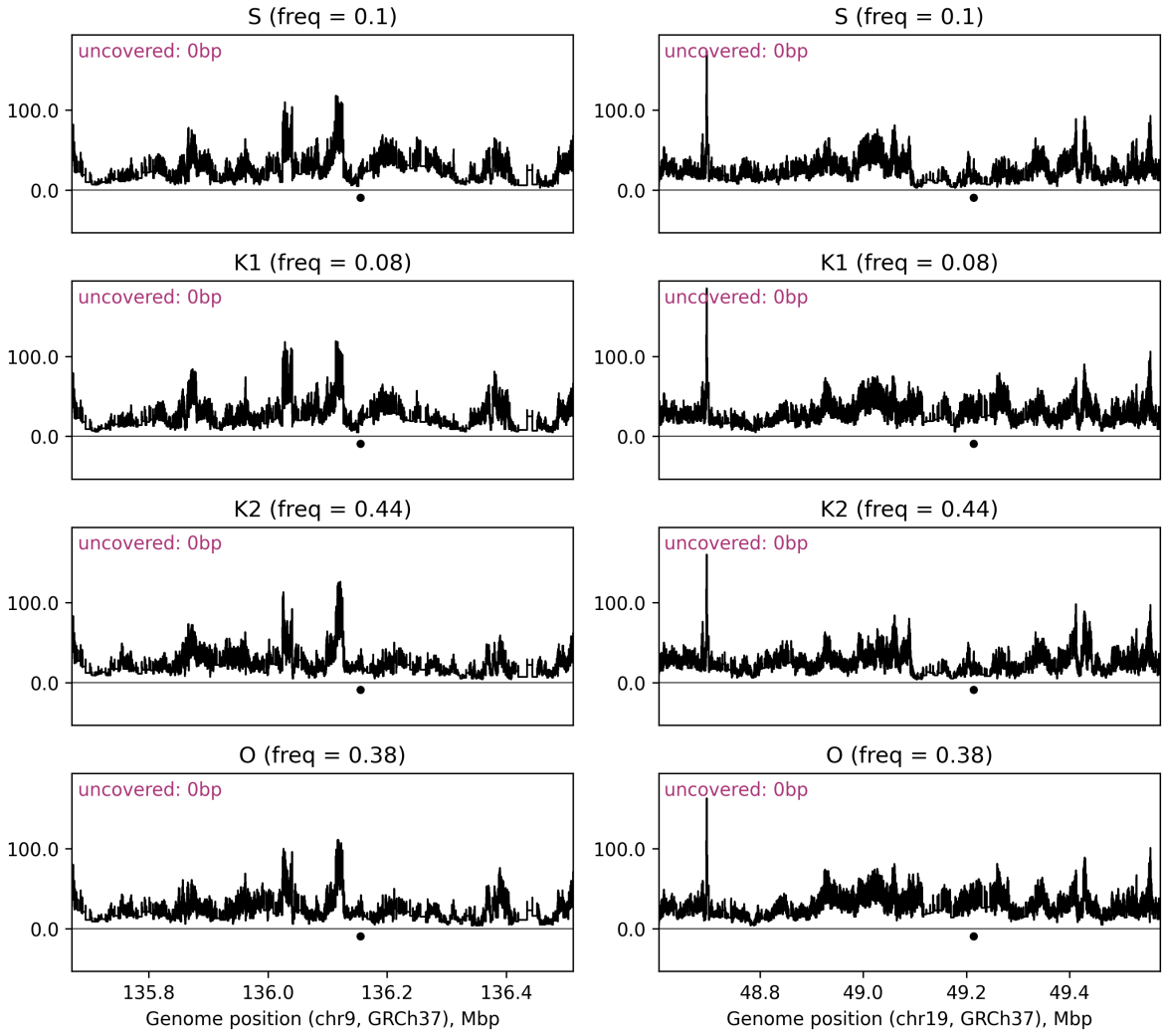

Figure 5: Results for *ABO–FUT2* interaction. See legend of Figure 4.

by the size of the currently available ARGs, is that the tested SNPs must have relatively high frequency (to allow them to be matched to the SNPs present in the 1KGP ARG). This, again, will likely be resolved when ARGs reconstructed for biobank-scale data become available, an area of very active research (e.g. Zhang et al., 2023).

Our approach also does not require access to the genetic and phenotypic data used in association testing. It is important to stress that the method takes as input pairs of SNPs that have been detected to have significant interaction effects. It can thus only specifically detect whether these results are likely to be due to phantom epistasis as defined here (the presence of an omitted additive effect variant). The method will not identify, for instance, false positives due to general inflation of test statistics or other technical issues with the test for interactions. Thus, the method does not remove the necessity of constructing sensitive and powerful tests for detecting epistasis, checking for the inflation of test statistics and replicating the results across multiple datasets.

A key strength of our method is that, as noted previously in the literature, phantom epistasis appears likely to arise if the two tested SNPs are in close proximity, and significant interactions between such SNPs are typically assumed to be false positives. In general, true epistasis between nearby SNPs may well be commonplace, and the problem is only that detecting this is very difficult, in part due to the fundamental unidentifiability of the model if a variant exists whose set of carriers is correlated with a target set. Our method explicitly quantifies how likely this is to be the case, regardless of the distance between the two SNPs, and thus offers a way of distinguishing which interactions between nearby SNPs might represent true biological effects.

5 Data availability

Data sharing is not applicable to this article as no new data were generated or analysed.

6 Acknowledgements

LAFF was supported by the Wellcome Trust [22334/Z/21/Z].

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, **526**(7571): 68, 2015.

Brandt, D. Y., Huber, C. D., Chiang, C. W., and Ortega-Del Vecchyo, D. The promise of inferring the past using the ancestral recombination graph. *Genome Biology and Evolution*, **16**(2): evae005, 2024.

Cordell, H. J. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, **10**(6): 392–404, 2009.

de los Campos, G., Sorensen, D. A., and Toro, M. A. Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data). *G3: Genes, Genomes, Genetics*, **9**(5): 1429–1436, 2019.

Domingo, J., Baeza-Centurion, P., and Lehner, B. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics*, **20**(1): 433–460, 2019.

Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics*, **42**(11): 985–990, 2010.

Hayman, E., Ignatieva, A., and Hein, J. Recoverability of ancestral recombination graph topologies. *Theoretical Population Biology*, **154**: 27–39, 2023.

Hemani, G., Powell, J. E., Wang, H., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A. K., McRae, A. F., Martin, N. G., Metspalu, A., et al. Phantom epistasis between unlinked loci. *Nature*, **596**(7871): E1–E3, 2021.

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A. K., McRae, A. F., Yang, J., Gibson, G., Martin, N. G., Metspalu, A., et al. Retracted article: Detection and replication of epistasis influencing transcription in humans. *Nature*, **508**(7495): 249–253, 2014.

Hemani, G., Theocharidis, A., Wei, W., and Haley, C. EpiGPU: Exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, **27**(11): 1462–1465, 2011.

Hill, W. G., Goddard, M. E., and Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics*, **4**(2): e1000008, 2008.

Hivert, V., Sidorenko, J., Rohart, F., Goddard, M. E., Yang, J., Wray, N. R., Yengo, L., and Visscher, P. M. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *The American Journal of Human Genetics*, **108**(5): 786–798, 2021.

Ignatieva, A., Favero, M., Koskela, J., Sant, J., and Myers, S. R. The length of haplotype blocks and signals of structural variation in reconstructed genealogies. *bioRxiv*, 2024.

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, **53**(4): 420–425, 2021.

Lappalainen, T., Li, Y. I., Ramachandran, S., and Gusev, A. Genetic and molecular architecture of complex traits. *Cell*, **187**(5): 1059–1075, 2024.

Lewanski, A. L., Grundler, M. C., and Bradburd, G. S. The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *PLOS Genetics*, **20**(1): e1011110, 2024.

Lyon, M. S., Millard, L. A. C., Smith, G. D., Gaunt, T. R., and Tilling, K. Hypothesis-free detection of gene-interaction effects on biomarker concentration in UK Biobank using variance prioritisation. *medRxiv*, 2022.

Mackay, T. F. C. Epistasis and quantitative traits: Using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, **15**(1): 22–33, 2014.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. Finding the missing heritability of complex diseases. *Nature*, **461**(7265): 747–753, 2009.

Masuda, M., Okuda, K., Ikeda, D. D., Hishigaki, H., and Fujiwara, T. Interaction of genetic markers associated with serum alkaline phosphatase levels in the Japanese population. *Human Genome Variation*, **2**(1): 1–6, 2015.

Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**(11): 855–867, 2008.

Sella, G. and 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**(1): 461–493, 2019.

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, **51**(D1): D977–D985, 2023.

Speidel, L., Forest, M., Shi, S., and Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, **51**(9): 1321–1329, 2019.

Sun, X., Lu, Q., Mukherjee, S., Crane, P. K., Elston, R., and Ritchie, M. D. Analysis pipeline for the epistasis search – statistical versus biological filtering. *Frontiers in Genetics*, **5**, 2014.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, **20**(8): 467–484, 2019.

The Australo-Anglo-American Spondyloarthritis Consortium (TASC), the Wellcome Trust Case Control Consortium 2 (WTCCC2), Evans, D. M., Spencer, C. C. A., Pointon, J. J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics*, **43**(8): 761–767, 2011.

Torkamani, A., Wineinger, N. E., and Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, **19**(9): 581–590, 2018.

Upton, A., Trelles, O., Cornejo-García, J. A., and Perkins, J. R. Review: High-performance computing to detect epistasis in genome scale data sets. *Briefings in Bioinformatics*, **17**(3): 368–379, 2016.

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J. C., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, **51**(9): 1339–1348, 2019.

Wei, W.-H., Hemani, G., and Haley, C. S. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, **15**(11): 722–733, 2014.

Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A. W., and Kelleher, J. A general and efficient representation of ancestral recombination graphs. *Genetics*, **228**(1): iyae100, 2024.

Wood, A. R., Tuke, M. A., Nalls, M. A., Hernandez, D. G., Bandinelli, S., Singleton, A. B., Melzer, D., Ferrucci, L., Frayling, T. M., and Weedon, M. N. Another explanation for apparent epistasis. *Nature*, **514**(7520): E3–E5, 2014.

Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F., and Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, **55**(5): 768–776, 2023.

Supplementary Information

S1 Supplementary methods **1**

S1.1 Summary of notation . . . 1

S1.2 Probability that a set of samples forms a clade in a local tree . . . 1

S1.3 Statistical conditions for phantom epistasis . . . 2

S1.4 Searching for clades correlated with a target set . . . 6

S1.5 Evidence against existence of clades correlated with a target set . . . 7

S1.6 Processing the 1KGP ARG . . . 8

S1.7 Simulations . . . 9

S1 Supplementary methods

S1.1 Summary of notation

Variable	Description
n	Total sample size of the ARG
N	Total sample size used in statistical testing for interactions
A	Set of samples $\{a_1, \ldots, a_k\}$
$\|A\|$	Size of set A, $\|A\| = k$
$\widehat{A}$	Clade in the ARG containing exactly the samples in set A
$\widetilde{\boldsymbol{a}}$	Haploid genotype vector $\widetilde{\boldsymbol{a}} = (\widetilde{a}_1, \ldots, \widetilde{a}_n)$ where $a_i = 1$ if $i \in A$, 0 otherwise
$\boldsymbol{a}$	Normalised genotype vector $\boldsymbol{a} = \widetilde{\boldsymbol{a}} - \sum_{i=1}^{n} a_i/n$
$A \cap B$	Intersection set $\{i : i \in A \text{ and } i \in B\}$
X_1, X_2	Sets of samples carrying SNP1 and SNP2, respectively
m_1, m_2	Genomic positions of SNP1 and SNP2, respectively
S	Set of samples corresponding to the interaction term, $S = X_1 \cap X_2$
$\widetilde{\boldsymbol{x}}_1$, $\widetilde{\boldsymbol{x}}_2$, $\widetilde{\boldsymbol{s}}$	Genotype vectors for SNP1, SNP2 and their interaction, with $\widetilde{\boldsymbol{x}}_1 = (\widetilde{x}_{11}, \ldots, \widetilde{x}_{1n})$ and $\widetilde{x}_{1i} = 1$ if $i \in X_1$ and 0 otherwise
$\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{s}$	Normalised genotype vectors for SNP1, SNP2 and their interaction
$\boldsymbol{X}$	Genotype matrix $[\boldsymbol{x}_1 \| \boldsymbol{x}_2 \| \boldsymbol{s}]$
σ_1^2, σ_2^2, σ_s^2	Variance of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{s}$, respectively
σ_{12}, σ_{1s}, σ_{2s}	Covariance of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\boldsymbol{x}_1$ and $\boldsymbol{s}$, and $\boldsymbol{x}_2$ and $\boldsymbol{s}$, respectively
$\boldsymbol{y}$	Vector of phenotype values (standardised to have zero mean and unit variance)

Note that, for readibility, in the main text we do not distinguish between unnormalised and normalised genotype vectors (since this does not cause ambiguity in the narrative).

S1.2 Probability that a set of samples forms a clade in a local tree

Consider the probability that the target set S forms a clade in a random coalescent tree with n samples. This requires that the samples in S coalesce together before they coalesce with any of the other samples, giving the recursions

$$p_{i,j} = \frac{\binom{i}{2}}{\binom{i+j}{2}} p_{i-1,j} + \frac{\binom{j}{2}}{\binom{i+j}{2}} p_{i,j-1},$$

where i is the number of lineages currently in the tree subtending S, j is the number of lineages currently in the tree *not* subtending S, and $p_{i,j}$ is the probability that S forms a clade given i and j. The boundary conditions are $p_{1,j} = 1\ \forall j$ and we solve for $p_{|S|,n-|S|}$. This can be solved explicitly:

$$p_{|S|,n-|S|} = \frac{2}{(|S|+1)\binom{n-1}{|S|-1}},$$

see, for instance, Hein et al. (2004) [p. 84, eq. (3.26)]. For large n, unless $|S|$ (or $n - |S|$) is small, this probability is very small.

However, we know that S is a subset of X_1, which forms a clade $\widehat{X}_1$ at position m_1. So the relevant probability of interest is instead $p_{|S|,|X_1|-|S|}$, which is larger than $p_{|S|,n-|S|}$. In fact, in some *region* around m_1, this probability will be elevated (before the local trees are broken up by recombination). The same argument holds for the other target sets. Notice that this does not depend on m_1 and m_2 being close together (or even on the same chromosome). However, if m_1 and m_2 *are* close together, the probability that S forms a clade will be much higher, since S a subset of X_1 and a subset of X_2, which both form clades within the region $[m_1, m_2]$.

S1.3 Statistical conditions for phantom epistasis

In practice, it is enough for a set Z which is sufficiently highly correlated with S to form a clade $\widehat{Z}$ in order for phantom epistasis to emerge. We now derive the exact conditions that such a set must satisfy.

S1.3.1 The true model

Suppose that we have data on a sample of N individuals, indexed by i, for whom we observe a quantitative phenotype y_i. We assume the same theoretical setup considered by de los Campos et al. (2019) and so examine the effects of variation at three loci, denoted by z_i, x_{1i} and x_{2i}, on this phenotype. For clarity, we first re-derive the results of de los Campos et al. on the least-squares estimator of the interaction coefficient.

The first variant, z_i, is assumed to be the only 'causal locus': together with a zero-mean error term δ_i with constant variance σ_δ^2 for all individuals, it completely determines the phenotype:

$$y_i = bz_i + \delta_i.$$

From this theoretical setup, it is clear that there is no epistasis affecting y_i. However, we assume that the true causal locus z_i is not observed; instead, we only observe x_{1i} and x_{2i}, two variants that may be correlated with z_i.

S1.3.2 The model we fit

We fit the following linear regression model which allows for an epistatic relationship between x_{1i} and x_{2i} with an effect on the phenotype:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i.$$

Since the true additive effect z_i is unobserved, fitting the model above may suggest an epistatic relationship between x_{1i} and x_{2i} if the estimate for the coefficient β_{12} is significant.

We will now derive a threshold for the correlation between the simultaneous occurrence of variants x_{1i} and x_{2i} and the true additive effect z_i above which we expect to obtain a significant estimate for β_{12} as the sample size gets large.

S1.3.3 Notation

To simplify notation, we define auxiliary variables $s_i := x_{1i}x_{2i}$ and $\beta_s := \beta_{12}$, and so write the model to be estimated as follows:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_s s_i + \varepsilon_i. \tag{S1}$$

To aid with our derivations below, we also write the model in an equivalent matrix formulation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{S2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ (and similarly for $\boldsymbol{\varepsilon}$), the genotype matrix is

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{21} & s_1 \\ x_{12} & x_{22} & s_2 \\ & \cdots & \\ x_{1N} & x_{2N} & s_N \end{bmatrix},$$

and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_s)$.

S1.3.4 Assumptions

Without loss of generality, all variables (including s_i) are assumed to have mean zero. This simplifies derivations as it allows us to omit and ignore intercept terms.

We make four additional assumptions. First, all variables are independent and identically distributed across individuals. Second, the variances of, and covariances between, all the independent variables x_{1i}, x_{2i}, s_i and z_i exist and are finite. Third, there is no multicollinearity between the variables x_{1i}, x_{2i} and s_i; in other words, the matrix $\boldsymbol{X}$ is full rank. Finally, we assume zero correlation between the true error term δ_i and all independent variables in equation (S1) as well as the true causal variant z_i, i.e., $\mathbb{E}(\delta_i z_i) = \mathbb{E}(\delta_i x_{1i}) = \mathbb{E}(\delta_i x_{2i}) = \mathbb{E}(\delta_i s_i) = 0$ for all i.

The second and third assumptions are realistic in genetic applications: Bernoulli or binomial random variables always have finite variances and only non-identical genetic variants that segregate in the sample are considered, guaranteeing no multicollinearity. Note also that, while the model above assumes that the phenotype is influenced by only one genetic locus, we could derive the same results while including an additional polygenic component (e.g., $y_i = b_z z_i + \sum_{p=1}^{P} b_p w_{pi} + \delta_i$) as long as all these additional genetic variants (which would then be part of the error term in the model as it is stated) are uncorrelated with z_i, x_{1i} and x_{2i} so that the fourth assumption holds. The model is therefore more general than it may appear at first, with the key assumption being that *there is only one causal genetic locus* in the haplotypic block under consideration.

S1.3.5 Asymptotic behaviour of the least squares estimator of $\boldsymbol{\beta}$

We first emphasise that the true causal locus z_i is not part of the set of independent variables included in our regression (as it is not observed), which implies that it is effectively part of the error term. In other words, we may rewrite the model in equation (S1) as follows:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_s s_i + (b z_i + \delta_i), \tag{S3}$$

where by construction we have $\beta_1 = \beta_2 = \beta_s = 0$ and $\varepsilon_i = b z_i + \delta_i$.

From standard results for least squares estimators of linear models (see, e.g., Wooldridge, 2010, §4.2.1.) we know that the presence of a variable in the error term that has a non-zero effect on the dependent variable renders the estimator of the effect of an included regressor inconsistent as long as this variable is correlated with that regressor. It is this fact – that the least squares estimator may not converge to the true value for the parameter of interest β_s, which is zero – that can lead

to a false positive interaction being detected if z_i is correlated with s_i. Note that we cannot check empirically whether such correlations are non-zero as we do not observe z_i.

To derive the limit to which the estimator of the parameters $(\beta_1, \beta_2, \beta_s)$ converges, we begin by writing the linear projection (see, e.g., Wooldridge (2010) [§ 4.3.1]) of the unobserved variant z_i on the observed variants x_{1i}, x_{2i}, s_i:

$$z_i = \hat{\gamma}_1 x_{1i} + \hat{\gamma}_2 x_{2i} + \hat{\gamma}_s s_i + \hat{r}_i,$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_s) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}$ is the vector of least squares coefficients and $\hat{r}_i$ are the resulting residuals. Note that a unique projection is guaranteed to exist since the matrix $\boldsymbol{X}$ is full rank, and (by definition of a linear projection) the residuals $\hat{r}_i$ are uncorrelated with x_{1i}, x_{2i} and s_i. Plugging this expression for z_i into equation (S3), we obtain:

$$\begin{aligned} y_i &= (\beta_1 + b\hat{\gamma}_1)x_{1i} + (\beta_2 + b\hat{\gamma}_2)x_{2i} + (\beta_s + b\hat{\gamma}_s)s_i + (\delta_i + b\hat{r}_i) \\ \boldsymbol{y} &= \boldsymbol{X}(\boldsymbol{\beta} + b\hat{\boldsymbol{\gamma}}) + \boldsymbol{\delta} + b\hat{\boldsymbol{r}}. \end{aligned} \tag{S4}$$

(We note that, since both $\hat{r}_i$ and δ_i are uncorrelated with all regressors, $\boldsymbol{X}$ is full rank, and the error term has constant variance, the rewritten model in equation (S4) satisfies the standard assumptions required for consistency of the least squares estimators of both the coefficients of the model and of their variances (see, e.g., Wooldridge (2010) [p. 59, Theorem 4.2]) – this will be useful below.)

We now use the expression for $\boldsymbol{y}$ in (S4) to derive the expression for the estimator of the coefficients in our original model as written in (S2), denoted $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_s)$:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left(\boldsymbol{X}(\boldsymbol{\beta} + b\hat{\boldsymbol{\gamma}}) + \boldsymbol{\delta} + b\hat{\boldsymbol{r}}\right) \\ &= \boldsymbol{\beta} + b\hat{\boldsymbol{\gamma}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\delta} + b\hat{\boldsymbol{r}}) \\ &= \boldsymbol{\beta} + b(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\delta} + b\hat{\boldsymbol{r}}). \end{aligned}$$

We wish to derive the asymptotic behaviour of $\hat{\boldsymbol{\beta}}$. Slutsky's theorem allows us to write its limiting expression as follows (where $\xrightarrow{\text{p}}$ denotes convergence in probability and plim its limit):

$$\begin{aligned} \hat{\boldsymbol{\beta}} \xrightarrow{\text{p}} \boldsymbol{\beta} + b \times \operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{N}\right)^{-1} \times \operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\boldsymbol{z}}{N}\right) + \\ + \operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{N}\right)^{-1} \times \left[\operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\boldsymbol{\delta}}{N}\right) + b \times \operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\hat{\boldsymbol{r}}}{N}\right)\right]. \end{aligned}$$

We can now apply the weak law of large numbers, which ensures that sample means converge to expectations as $N \to \infty$. Defining a matrix T as follows:

$$T = \begin{bmatrix} \mathbb{E}(x_{1i}^2) & \mathbb{E}(x_{1i}x_{2i}) & \mathbb{E}(x_{1i}s_i) \\ \mathbb{E}(x_{1i}x_{2i}) & \mathbb{E}(x_{2i}^2) & \mathbb{E}(x_{2i}s_i) \\ \mathbb{E}(x_{1i}s_i) & \mathbb{E}(x_{2i}s_i) & \mathbb{E}(s_i^2) \end{bmatrix},$$

we have:

$$\operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{N}\right)^{-1} = T^{-1}.$$

Similarly,

$$\operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\boldsymbol{z}}{N}\right) = \begin{bmatrix} \mathbb{E}(x_{1i}z_i) \\ \mathbb{E}(x_{2i}z_i) \\ \mathbb{E}(s_i z_i) \end{bmatrix}.$$

Since we assume no correlation between δ_i and x_{1i}, x_{2i}, s_i, and $\hat{r}_i$ is uncorrelated with these regressors by construction, we have:

$$\operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\boldsymbol{\delta}}{N}\right) = \begin{bmatrix}\mathbb{E}(x_{1i}\delta_i)\\ \mathbb{E}(x_{2i}\delta_i)\\ \mathbb{E}(s_i\delta_i)\end{bmatrix} = \mathbf{0}$$

$$\operatorname*{plim}_{N\to\infty}\left(\frac{\boldsymbol{X}'\hat{\boldsymbol{r}}}{N}\right) = \begin{bmatrix}\mathbb{E}(x_{1i}\hat{r}_i)\\ \mathbb{E}(x_{2i}\hat{r}_i)\\ \mathbb{E}(s_i\hat{r}_i)\end{bmatrix} = \mathbf{0}.$$

Plugging in these limits, we obtain:

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_s\end{bmatrix} \xrightarrow{\text{p}} \begin{bmatrix}\beta_1\\ \beta_2\\ \beta_s\end{bmatrix} + b \times \begin{bmatrix}\mathbb{E}(x_{1i}^2) & \mathbb{E}(x_{1i}x_{2i}) & \mathbb{E}(x_{1i}s_1)\\ \mathbb{E}(x_{1i}x_{2i}) & \mathbb{E}(x_{2i}^2) & \mathbb{E}(x_{2i}s_1)\\ \mathbb{E}(x_{1i}s_1) & \mathbb{E}(x_{2i}s_1) & \mathbb{E}(s_i^2)\end{bmatrix}^{-1} \begin{bmatrix}\mathbb{E}(x_{1i}z_i)\\ \mathbb{E}(x_{2i}z_i)\\ \mathbb{E}(s_iz_i)\end{bmatrix},$$

which is the same as equation 3 in de los Campos et al. (2019).

Focusing now on our parameter of interest β_s and its estimator, we note that by construction $\beta_s = 0$ and therefore from the expression above this estimator reflects only the direct effect of the unobserved causal locus z_i (given by the coefficient b) which is weighted by the coefficient of the interaction s_i in the linear projection of z_i above (loosely, this reflects how strongly correlated s_i and z_i are after accounting for the correlations between z_i and x_{1i} and x_{2i}). We see that our estimate of β_s will tend to be larger the larger the effect of the true locus and the stronger its correlation with the interaction term of interest after accounting for the main effects. The limiting expression for $\hat{\beta}_s$ is:

$$\hat{\beta}_s \xrightarrow{\text{p}} \begin{bmatrix}t_{31} & t_{32} & t_{33}\end{bmatrix} \begin{bmatrix}\mathbb{E}(x_{1i}z_i)\\ \mathbb{E}(x_{2i}z_i)\\ \mathbb{E}(s_iz_i)\end{bmatrix} b,$$

where t_{31}, t_{32}, t_{33} are the elements of the third row of the inverse of T.

The final step is to invert T. To simplify notation, we now define $\sigma_1^2 \coloneqq \mathbb{E}(x_{1i}^2)$, $\sigma_{12} \coloneqq \mathbb{E}(x_{1i}x_{2i})$ and so on. The determinant of T is:

$$\begin{aligned}\det(T) &= \sigma_1^2\sigma_2^2\sigma_s^2 + \sigma_{12}\sigma_{2s}\sigma_{1s} + \sigma_{1s}\sigma_{12}\sigma_{2s} - \sigma_{1s}\sigma_2^2\sigma_{1s} - \sigma_{12}\sigma_{12}\sigma_s^2 - \sigma_1^2\sigma_{2s}\sigma_{2s}\\ &= \sigma_1^2\sigma_2^2\sigma_s^2 + 2\sigma_{12}\sigma_{2s}\sigma_{1s} - (\sigma_{1s})^2\sigma_2^2 - (\sigma_{12})^2\sigma_s^2 - \sigma_1^2(\sigma_{2s})^2.\end{aligned}$$

The third row of T^{-1} is then:

$$\begin{aligned}t_{31} &= \det(T)^{-1}(\sigma_{12}\sigma_{2s} - \sigma_2^2\sigma_{1s})\\ t_{32} &= \det(T)^{-1}(\sigma_{12}\sigma_{1s} - \sigma_1^2\sigma_{2s})\\ t_{33} &= \det(T)^{-1}(\sigma_1^2\sigma_2^2 - (\sigma_{12})^2).\end{aligned}$$

Finally, we can write the limit of $\hat{\beta}_s$:

$$\hat{\beta}_s \xrightarrow{\text{p}} \det(T)^{-1}[(\sigma_{12}\sigma_{2s} - \sigma_2^2\sigma_{1s})\mathbb{E}(x_{1i}z_i) + (\sigma_{12}\sigma_{1s} - \sigma_1^2\sigma_{2s})\mathbb{E}(x_{2i}z_i) + (\sigma_1^2\sigma_2^2 - (\sigma_{12})^2)\mathbb{E}(s_iz_i)] \times b.$$

We have now derived the limit of the least squares estimator, which agrees with that derived by de los Campos et al. (2019). To complete our analysis, we must also consider the behaviour of its standard error as the sample size increases, as it is the ratio of these two estimators that will determine whether or not we detect a statistically significant effect for the interaction s_i.

S1.3.6 Asymptotic behaviour of the standard error of $\boldsymbol{\beta}$

Having rewritten our model such that it obeys standard least squares assumptions in equation (S4), we can use known results to write the limit of the usual estimator of the variance of $\hat{\boldsymbol{\beta}}$:

$$\widehat{\mathrm{Var}}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N-3} (\boldsymbol{X}'\boldsymbol{X})^{-1} \xrightarrow{\mathrm{p}} \mathrm{Var}(\delta_i + b\hat{r}_i) \times (NT)^{-1}.$$

Now since $\hat{r}_i$ is a function of x_{1i}, x_{2i} and z_i, it is non-random when conditioning on these explanatory variables (which we implicitly do throughout these derivations). Therefore $\mathrm{Var}(\delta_i + b\hat{r}_i) = \mathrm{Var}(\delta_i) = \sigma_\delta^2$ and so:

$$\widehat{\mathrm{Var}}(\hat{\beta}_s) \xrightarrow{\mathrm{p}} \sigma_\delta^2 \frac{t_{33}}{N} = \sigma_\delta^2 \det(T)^{-1} (\sigma_1^2 \sigma_2^2 - (\sigma_{12})^2) \frac{1}{N}.$$

S1.3.7 Deriving a threshold for false positive interactions

Having obtained the limits of both the estimator of the coefficient $\hat{\beta}_s$ and its standard error, we now consider the asymptotic behaviour of the test statistic. A standard t-test of the null hypothesis $\beta_s = 0$ is:

$$t\text{-statistic} = \frac{\hat{\beta}_s}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_s)}}.$$

This will be compared against the quantile of a Student's t distribution with $N-3$ degrees of freedom which, for large samples, approximates a standard Normal.

Therefore, and assuming a positive estimate of β_s, we will reject the null hypothesis and find evidence for a putative epistatic effect if and only if:

$$\frac{\hat{\beta}_s}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_s)}} \geq \Phi^{-1}(1 - \alpha/2) =: C,$$

where α is the significance level of the test (commonly 5×10^{-8} for standard GWASs but may be much lower for exhaustive tests of pairwise interactions).

Given our results above for the convergence of the numerator and denominator of the t-statistic, the inequality determining rejection of the null hypothesis is:

$$\frac{(\sigma_{12}\sigma_{2s} - \sigma_2^2\sigma_{1s})\mathbb{E}(x_{1i}z_i) + (\sigma_{12}\sigma_{1s} - \sigma_1^2\sigma_{2s})\mathbb{E}(x_{2i}z_i) + (\sigma_1^2\sigma_2^2 - (\sigma_{12})^2)\mathbb{E}(s_i z_i)}{\sqrt{\det(T)(\sigma_1^2\sigma_2^2 - (\sigma_{12})^2)}} \geq \frac{C}{b}\sqrt{\frac{\sigma_\delta^2}{N}}. \quad \text{(S5)}$$

S1.4 Searching for clades correlated with a target set

Note that in (S5), $\sigma_1^2, \sigma_2^2, \sigma_s^2, \sigma_{12}, \sigma_{1s}, \sigma_{2s}$, and therefore $\det(T)$, can be easily estimated from the genotype data used for interaction testing, or from the ARG. Moreover, σ_δ^2 can be written equivalently as $\sigma_y^2(1-h^2)$, where h^2 is the SNP heritability of the phenotype (and can therefore also be readily estimated from available data), and we assumed that the phenotype is standardised to have unit variance so $\sigma_y^2 = 1$.

However, given the inherent difficulty with obtaining a reliable estimate for the effect size b, as it is by definition unobserved, instead of computing an estimate for the threshold directly, in practice we calculate the normalised genotype vector $\boldsymbol{s}$ using the ARG (corresponding to the intersection of observed sets X_1 and X_2), and substitute this for $\boldsymbol{z}$ into the LHS of (S5). Since we know that the effect size of the interaction term is significant (otherwise no epistasis would have been detected), this gives an upper bound on the value of the RHS. We then multiply this by a factor of 0.8, producing a lower (more conservative) threshold. We validated this factor through simulation

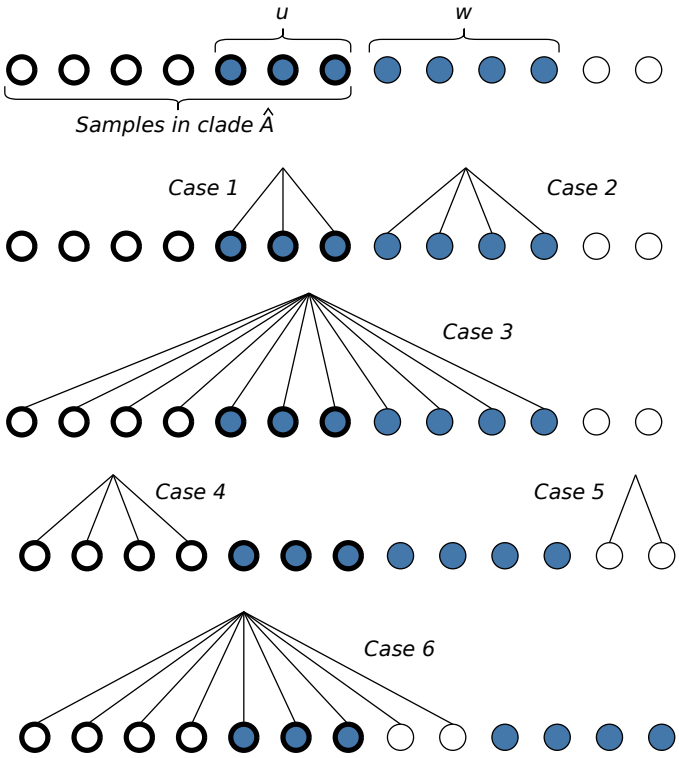

Figure S1: Samples of an ARG (ARG itself not shown) of size $n = 14$. Samples in clade $\widehat{A}$ shown with bold outline; samples in set S shown in blue. Lines connect samples that could possibly form a clade $\widehat{Z}$ in each considered case. Note that since we condition on A itself forming a clade, we cannot observe a clade formed of some samples from within $\widehat{A}$ and some samples from outside $\widehat{A}$ as this would violate that assumption (for instance, in the scenario shown, the set S itself cannot form a clade — only a set correlated with S can).

studies and on the Hemani et al. data, and checked that this is in line with the known range of estimates of b and h^2 for the phenotypes in question (where we used a high quantile of the estimated effect sizes across all significant eQTLs for a given phenotype as an estimate of b).

We can thus, for each clade $\widehat{Z}$ of the ARG, calculate the corresponding normalised genotype vector $\boldsymbol{z}$, calculate the covariances $\mathbb{E}(x_{1i}z_i), \mathbb{E}(x_{2i}z_i)$ and $\mathbb{E}(s_i z_i)$, and check whether (S5) is satisfied (and repeat this replacing $\boldsymbol{s}$ with each of the other target set genotype vectors). If so, this implies that the effect size of the interaction term in the model may be spuriously capturing the additive effect of the corresponding genotype $\boldsymbol{z}$, giving rise to phantom epistasis. This is the basis for the procedure described in Methods, Section 2.4.1.

S1.5 Evidence against existence of clades correlated with a target set

Let S be the set of samples in the ARG corresponding to the interaction genotype vector $\boldsymbol{s}$ (that is, the set of samples that carry both SNP1 and SNP2). Suppose clade $\widehat{A}$ appears in some local tree along the genome, with $|S \cap A| = u$, and let $w = |S| - u$ (as illustrated in Figure S1, top line). We would like to check whether $\widehat{A}$ provides evidence against the existence of another clade $\widehat{Z}$ (within the genomic span of $\widehat{A}$) such that the corresponding normalised genotype vector $\boldsymbol{z}$ would satisfy (S5).

Since, in realistic scenarios, ARG reconstruction is not perfect and there is potentially a large amount of uncertainty about the topology of each local tree, we condition only on the fact that the samples in set A form a clade $\widehat{A}$. Then there are six possible scenarios to consider which maximise the correlation between a possible clade $\widehat{Z}$ and the set S, shown in Figure S1. Cases 1–3 (resp. 4–6) correspond to maximising the positive (resp. negative) correlation between Z and S, when $\widehat{Z}$ is a subclade (Cases 1 and 4), superclade (Cases 3 and 6) or disjoint from $\widehat{A}$ (Cases 2 and 5).

Taking each case in turn, we calculate the values of $\mathbb{E}(x_{1i}z_i)$, $\mathbb{E}(x_{2i}z_i)$ and $\mathbb{E}(s_iz_i)$. Let $f_u := u/n$, $f_w := w/n$, $f_{x1} := |X_1|/n$, $f_{x2} := |X_2|/n$, $f_s := |S|/n$, $f_a := |A|/n$, $f_{x1a} := |A \cap X_1|/n$ and $f_{x2a} := |A \cap X_2|/n$. Then for Case 1, $|Z| = u$ and

$$\begin{aligned}
\mathbb{E}(x_{1i}z_i) &= \mathbb{E}(\widetilde{x}_{1i}\widetilde{z}_i) - \mathbb{E}(\widetilde{x}_{1i})\mathbb{E}(\widetilde{z}_i) = f_u(1 - f_{x1}),\\
\mathbb{E}(x_{2i}z_i) &= \mathbb{E}(\widetilde{x}_{2i}\widetilde{z}_i) - \mathbb{E}(\widetilde{x}_{2i})\mathbb{E}(\widetilde{z}_i) = f_u(1 - f_{x2}),\\
\mathbb{E}(s_iz_i) &= \mathbb{E}(\widetilde{s}_i\widetilde{z}_i) - \mathbb{E}(\widetilde{s}_i)\mathbb{E}(\widetilde{z}_i) = f_u(1 - f_s);
\end{aligned}$$

for Case 2, $|Z| = w$ and

$$\begin{aligned}
\mathbb{E}(x_{1i}z_i) &= f_w(1 - f_{x1}),\\
\mathbb{E}(x_{2i}z_i) &= f_w(1 - f_{x2}),\\
\mathbb{E}(s_iz_i) &= f_w(1 - f_s);
\end{aligned}$$

for Case 3, $|Z| = |A| + w$ and

$$\begin{aligned}
\mathbb{E}(x_{1i}z_i) &= f_{x1a} + f_w - f_{x1}(f_a + f_w),\\
\mathbb{E}(x_{2i}z_i) &= f_{x2a} + f_w - f_{x2}(f_a + f_w),\\
\mathbb{E}(s_iz_i) &= f_s(1 - f_a - f_w).
\end{aligned}$$

Similarly, for Case 4, $|Z| = |A| - u$ and

$$\begin{aligned}
\mathbb{E}(x_{1i}z_i) &= f_{x1a} - f_u - f_{x1}(f_a - f_u),\\
\mathbb{E}(x_{2i}z_i) &= f_{x2a} - f_u - f_{x2}(f_a - f_u),\\
\mathbb{E}(s_iz_i) &= -f_s(f_a - f_u);
\end{aligned}$$

for Case 5, $|Z| = n - |A| - w$ and

$$\begin{aligned}
\mathbb{E}(x_{1i}z_i) &= f_{x1} - f_{x1a} - f_w - f_{x1}(1 - f_a - f_w),\\
\mathbb{E}(x_{2i}z_i) &= f_{x2} - f_{x2a} - f_w - f_{x2}(1 - f_a - f_w),\\
\mathbb{E}(s_iz_i) &= -f_s(1 - f_a - f_w);
\end{aligned}$$

for Case 6, $|Z| = n - w$ and

$$\begin{aligned}
\mathbb{E}(x_{1i}z_i) &= f_{x1} - f_w - f_{x1}(1 - f_w) = -f_w(1 - f_{x1}),\\
\mathbb{E}(x_{2i}z_i) &= -f_w(1 - f_{x2}),\\
\mathbb{E}(s_iz_i) &= f_u - f_s(1 - f_w).
\end{aligned}$$

Thus, for each clade $\widehat{A}$ in the ARG, if (S5) does not hold for any of these six cases, this provides evidence against the existence of a clade that is highly correlated with S and which could cause the interaction term to have a significant effect (within the genomic span of $\widehat{A}$). Note that we make a key simplifying assumption that Z being highly correlated with S is more likely to cause (S5) to be satisfied, which we expect to hold in practice unless S is small. We repeat these calculations for each of the other target sets K_1, K_2 and O. This is the basis for the procedure described in Methods, Section 2.4.2.

S1.6 Processing the 1KGP ARG

We converted the ARG reconstructed using Relate for the 1KGP (Phase 3, GRCh37) data into tskit format (Kelleher et al., 2016; Wong et al., 2024), and subsetted this to $n = 1006$ samples from European populations to match the data used by Hemani et al. (2014) (Powell et al., 2012).

We defined each clade by its set of descendant samples, and traversed the 1KGP ARGs for each chromosome, recording the genomic position at which each observed clade first arose, and at which it disappeared (due to recombination). We only considered SNPs that map uniquely to a branch in the corresponding local tree, and used the HapMapII GRCh37 recombination map to calculate genetic distances.

S1.7 Simulations

To simulate ('true') interactions between SNPs in *cis*, we selected two SNPs at random from the 1KGP ARG from a randomly chosen chromosome, varying the frequency of each SNP (0.1 to 0.5) and the distance between the two SNPs (0.1 to 2.0 cM). We ran 100 simulations for each combination of parameters (requiring a frequency of at least 0.5% for each target set), and applied the two methods described in Sections 2.4.1 and 2.4.2, searching the region $[m_1 - 1\text{cM}, m_2 + 1\text{cM}]$ in each case. We esimated the threshold F as described in Section 2.4.1. To simulate interactions between SNPs in *trans*, we ran the same simulations as above, but selecting the two SNPs on different randomly chosen chromosomes.

References

de los Campos, G., Sorensen, D. A., and Toro, M. A. Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data). *G3: Genes, Genomes, Genetics*, **9**(5): 1429–1436, 2019.

Hein, J., Schierup, M., and Wiuf, C. *Gene genealogies, variation and evolution: A primer in coalescent theory.* Oxford University Press, USA, 2004.

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A. K., McRae, A. F., Yang, J., Gibson, G., Martin, N. G., Metspalu, A., et al. Retracted article: Detection and replication of epistasis influencing transcription in humans. *Nature*, **508**(7495): 249–253, 2014.

Kelleher, J., Etheridge, A. M., and McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, **12**(5): e1004842, 2016.

Powell, J. E., Henders, A. K., McRae, A. F., Caracella, A., Smith, S., Wright, M. J., Whitfield, J. B., Dermitzakis, E. T., Martin, N. G., Visscher, P. M., et al. The Brisbane Systems Genetics Study: Genetical genomics meets complex trait genetics. *PLOS One*, **7**(4): e35430, 2012.

Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A. W., and Kelleher, J. A general and efficient representation of ancestral recombination graphs. *Genetics*, **228**(1): iyae100, 2024.

Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data.* The MIT Press, Cambridge, Massachusetts, 2nd edn., 2010.