Testing for genetic interactions with imperfect information about additive causal effects

Gibran Hemani1,\*, Joseph E. Powell2,3,\*, Huanwei Wang4, Konstantin Shakhbazov, Harm-Jan Westra5, Tonu Esko6,7,8, Anjali K. Henders9, Allan F. McRae4, Nicholas G. Martin9, Andres Metspalu6, Lude Franke5, Grant W. Montgomery4, Michael E. Goddard10, Greg Gibson11, Jian Yang4,\*, Peter M. Visscher4,\*

Affiliations:

1. MRC Integrative Epidemiology Unit at the University of Bristol, Oakfield House, Oakfield Grove, Bristol, UK. BS8 2BN.
2. UNSW Cellular Genomics Futures Institute, School of Medical Sciences, University of New South Wales, Sydney, NSW, Australia
3. Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia
4. Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
5. Department of Genetics, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9700 RB Groningen, The Netherlands.
6. Estonian Genome Center, University of Tartu, Tartu 51010, Estonia
7. Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142, USA
8. Divisions of Endocrinology, Children's Hospital, Boston, Massachusetts 02115, USA.
9. Queensland Institute of Medical Research, Brisbane, Queensland 4006, Australia.
10. Faculty of Veterinary and Agricultural Science, University of Melbourne, Melbourne 3052 VIC, Australia
11. School of Biology and Centre for Integrative Genomics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA.

\* These authors contributed equally

## Abstract

The extent to which genetic interaction (epistasis) contributes to the genetic architecture of human complex traits is expected to be low but remains unknown. A common approach to evaluating this question is to perform targeted or hypothesis-free searches of pairwise interactions, using a statistical test that evaluates if a model with the interaction terms explain more phenotypic variance than that with the marginal terms alone. Using results of 501 genetic interactions that we previously reported to influence gene expression, alongside theory and simulations, this paper seeks to explain why the standard approach is unreliable. We show that it can lead to increased false discovery rates in both discovery and replication data sets, and that without sequence level data and assumptions of homogenous additive effects across individuals being met, the problems are difficult to guard against.

## Introduction

An important component of understanding the genetic architecture of complex traits is the extent to which the effect of a particular locus depends on the genotype at another locus. This phenomenon is referred to as genetic interaction or epistasis. Knowledge of epistatic influences on complex traits may inform biological understanding of their aetiology, contribute towards improved prediction accuracy, and have implications on natural selection1,2. Along with other non-additive genetic components, the contribution of epistatic variance to complex traits is likely to be small3,4. However, beyond analyses of clonal organisms5, broad sense heritability is not easily to estimate where the non-additive variance components can be captured and decomposed6. Instead, researchers have sought to detect interacting genetic loci through association analyses to report instances of epistatic influences on complex traits.

In a recent review of the literature, it was concluded that, among many empirical papers that search for epistasis influencing human complex traits, it was unlikely that any reported statistical genetic interactions represented robust examples of biological epistasis2. The analysis that we presented in Hemani et al 20144 attempted to be rigorous, using a classical statistical test for epistasis and setting strict significance thresholds for discovery and replication. In that study we reported discovery of genetic interactions influencing gene expression levels, and replication in independent datasets, but since then doubt has been cast on even these findings7,8. Here, we aim to provide a detailed examination of the statistical test for epistasis, which is employed widely, but has potential issues which have not yet been described.

## The traditional statistical test for 2-locus genetic interactions

Should epistatic interactions influence complex traits, their detection is known to be difficult for two reasons. First, the statistical power for an interaction term to reach significance is low in comparison to a marginal additive effect of similar magnitude. This is because the statistical test typically has a larger number of degrees of freedom (d.f.), and if the causal variants are not available in the data then loss of signal with decaying linkage disequilibrium (LD) between the causal variant and the observed variant is squared or quadratic, in comparison to a linear loss for additive effects9. Second, the feature space for two-locus epistasis is O(m2), where m is the number of markers being tested, hence a much stricter multiple testing correction is required than association tests under the additive model. If the computational capability does not exist to test the entire set of pairwise interactions, then the incomplete coverage likely translates into loss of power.

Many methods exist that attempt to circumvent these problems2. One analytical strategy has been to bypass statistical power issues by selecting traits to analyse that are likely to have some large effects. In such traits, genetic perturbation could have a more proximal effect in comparison to complex diseases. Recent studies have focused efforts on analysing gene expression levels for epistatic interactions partly for this reason4,10. In Hemani et al 2014 a brute-force search strategy was performed, applying a 4 d.f. linear model test for each pairwise combination of 528,509 genotyped autosomal single nucleotide polymorphisms (SNPs) against each of 7,339 gene expression levels. The statistical test attempted to capture any joint effect of two independent variants that was not explained by the marginal additive or dominance effects of either of the variants11:

where µ is the phenotype mean and is the phenotypic mean of pairwise genotype class for genotype at locus A and genotype at locus B, is the marginal phenotypic class mean for genotype A and is the marginal phenotypic class mean for genotype B. Here the additive by additive, additive by dominance, dominance by additive and dominance by dominance terms are jointly estimated in the interaction term. This effect decomposition is fundamental to basic quantitative genetic theory12, and has been used routinely in the linkage study era and the GWAS era2,13. The level of epistasis can be tested for statistical significance using an *F*-test with degrees of freedom assuming individuals are present in all pairwise genotype classes. A simpler variation is to parameterise the interaction term to include only the additive by additive term, though what follows in this paper applies to that approach also.

The 4 d.f. method was used in Hemani et al 2014 with a sample of 846 individuals, which yielded 501 pairwise interactions that surpassed a permutation-derived threshold of . Throughout this paper we will refer to these associations as the H2014 interactions, using them as examples to explore issues with the test statistic. The majority of these interactions were long-range ‘cis-trans’ associations, where one interacting variant was close to the gene whose expression level was influenced, and the other interacting variant was on a different chromosome. In two independent datasets, together comprising 2,131 individuals, 30 of these interactions replicated at a Bonferroni multiple testing correction (), and 46 replicated at FDR < 0.05.

## A summary of the problems with the original findings

Soon after publication, these findings were further statistically replicated in an independent dataset by Wood *et al*. (2014)7. However, with the availability of sequence level genetic data, Wood *et al*. were able to fine-map the additive effects for each gene expression level where the H2014 signals were discovered. Typically, the fine-mapped additive effects were distinct from the *cis*-acting interacting variant. Upon including the fine-mapped additive effects as covariates in the interaction models they found that most of the interaction effects substantially attenuated. Following this finding, we found a similar attenuation of effects in the original data by using fine-mapped imputed additive effects as covariates8. In light of this exchange, it became clear that certain statistical findings in H2014 were less likely to represent biological epistasis than originally thought. Importantly, it raised the question of why such a fundamental method was giving rise to unreliable results.

Wood *et al*. (2014) interpreted the original discovery interactions as haplotype effects, a well-understood mechanism by which two loci can appear epistatic but be due to a simple additive effect. Here, the observed loci flank a causal variant and are in incomplete linkage disequilibrium with each other and the causal variant. A statistical interaction between the observed loci can capture more of the additive variance of the causal variant than the marginal additive effects of both the observed loci combined. This explanation for the H2014 was unproven but plausible for the set of *cis-cis* interactions reported, those where the two interacting loci were each close to the gene whose expression levels they were associated with. Other work has since appeared that investigate the same process14. However, this explanation is not possible for *cis-trans* interactions, where the two interacting loci are on different chromosomes, and a fine-mapped *cis*-additive effect attenuates the interaction effect. This *cis-trans* scenario forms the majority of the H2014 signals. In this paper we explore the question of how a single unobserved *cis*-additive effect can give rise to *cis-trans* statistical associations. The mechanism that we go on to describe can also apply to *cis-cis* and *trans-trans* interactions. We then go on to explore how this mechanism influences replication rates, and discuss potential methods for avoiding the problem.

## Inflation of test statistics

If the test statistic for the interaction term can be attenuated with the inclusion of a single additive term, this implies that the interaction test statistic is inflated under the null hypothesis of no epistasis. To begin, we test this assumption by estimating the genomic inflation factor for each of the 501 H2014 signals. In each case, we ran a genome-wide analysis where we performed an interaction test of the detected *cis*-SNP against every other SNP excluding those on the *cis* chromosome. The genomic inflation factor was then calculated for the interaction test statistics across the set of genome-wide tests. Figure 1 shows that some loci have no obvious genomic inflation, while for many loci the inflation factor is much larger than that expected under the null. This is consistent with the idea that for many of the loci the test statistics are inflated.

There are other possible explanations that could give rise to high genomic inflation factors, such as an epistatic polygenic component, though this is unlikely given the discovery sample size15. We also observe that many loci that had discovered epistatic associations have low genomic inflation factors, which raises the question of whether this metric is a reliable indicator of false positives.

## Theory

The explanations for inflated genetic interaction test statistics, put forth by Wood et al (2014) and elaborated upon theoretically by De los Campos et al. (2019), depended on both variants being in LD with a causal additive variant. We set out to understand the mechanism by which the classic test statistic for interaction test can be inflated when only one of the interaction variants is in LD with a causal additive variant. Such a mechanism would be sufficient to explain our observations of genome-wide inflation for some markers, found in the previous section. In the Appendix, we derive the properties and consequences of a statistical test for interaction for this case. To keep the theory tractable and intuitive, we used a haploid model, and additive effects that are so large that the phenotype has a binomial distribution. We find that when one of the interaction variants is in LD with a large additive causal variant, the residuals from a linear model are a mixture of normal and binomial distributions. This leads to systematic inflation or deflation of the test statistic. The more variance explained by the additive variant, the larger the proportion of variance of the residuals arises from the binomial distribution. We show that both the mean and the variance of the expected F value from the classical interaction model are increased. The mechanism described in the Appendix is entirely separate from the sources of test statistic inflation that have been previously suggested.

## Test statistic inflation in the diploid case

Having demonstrated that the mean and the variance of the test statistic is higher than expected under a simplified haploid model, we now use simulations to explore the behavior of the test statistic in the diploid context, using the H2014 signals as examples (Methods). Because of the scale of the original analysis it is difficult to mimic the conditions that gave rise to the H2014 signals, but we can evaluate the liability for test statistic inflation with respect to the variants reported.

There were 846 samples in the discovery and a combined 2,131 in the replication datasets used in Hemani et al 2014. We reported *MBNL1* gene expression being influenced by 11 *cis-trans* epistatic interactions, where the *cis* variant was rs13069559. Five of these replicated at the Bonferroni level (p < 0.05/501) but the genomic inflation factor for this locus was 3.15 and the additive *cis-*variant explained 10.5% of the phenotypic variance. Fitting the fine-mapped additive *cis*-variant rs67903230 attenuated the *cis-trans* signals involving rs13069559. We used real genotype data of the rs67903230 variant to simulate a phenotype with a large additive effect, and then performed the 4 d.f. interaction test for association between the originally discovered *cis* variant rs13069559 and 502510 genotyped markers, excluding the *cis* chromosome (chromosome 3). We used the rs13069559 variant based on the reasoning that if it was detected due to test statistic inflation then it was ascertained for its LD properties with the rs67903230 *cis*-additive causal variant. In the simulations, any pair of loci that had interaction test statistics surpassing a Bonferroni correction were taken forward to replication. As in the original analysis, we only allowed one *trans*-effect per autosome, thus the maximum number of *cis-trans* interactions for a simulation was 21. We performed 40,000 simulations, allowing the phenotypic variance explained by the rs67903230 additive effect to range uniformly from 0 to 50% across the set of simulations.

Figure 2 shows that the genomic inflation factor related strongly to the variance explained by the additive effect. Figure 3 demonstrates that as genomic inflation grows, the number of false positive interactions grows. We also observe that it is possible to obtain several false discovery signals per simulation even when the genomic inflation factor is low. This is consistent with the variance of the test statistic theoretically being inflated.

Extending these simulations to more scenarios by expanding to more loci amongst the H2014 signals resulted in slightly less inflation because we are no longer ascertaining for a locus that is known to have high inflation and high replication rates.

## Replication rate of false discovery signals

In Hemani et al 2014 we found a replication rate of 6%, after strict multiple testing correction (p < 0.05/501). An important result from that analysis was that the replication rate of true positive epistatic signals is expected to be low compared to additive effects, due to three processes: First, there is a more extreme winner’s curse for the epistatic signals due to a more stringent significance threshold. Second, there is a winner’s curse in the discovery sample operating on the LD between the causal interacting loci and each of the two observed markers. Third, the decay of the causal interaction signal with lowering LD in the replication sample is rapid. However, under the null hypothesis being true, it is not clear what is to be expected in terms of the replication rate of false positives from the discovery sample.

Using the simulations described in the previous section, we were able to compare the genomic inflation factor obtained in the discovery data against the replication data. Figure 4 demonstrates a strong relationship, though the F-statistics from the discovery and the replication were uncorrelated. We next asked if a simulation had at least one significant interaction under the null then what was the replication rate of that significant interaction in the independent replication sample? We used three different significance thresholds for determining replication, 1) FDR within simulation, 2) Bonferroni within simulation and 3) Experiment-wide Bonferroni as used in H2014 (p < 0.05/501). While the relaxed thresholds (1) and (2) could reach replication rates as high as 15%, the experiment-wide threshold (3) required very high genomic inflation to obtain any detectable replication, and the rate when genomic inflation was very high did not surpass 2% (Figure 5).

These simulations were designed to be as favorable as possible to generate false positive interaction terms. Analysing less extreme scenarios (Methods) gave lower replication rates. Further simulations, in which the scenarios incorporating more loci (Methods), resulted in lower experiment wide replication rates (Figure 5). These simulations do not perfectly mimic the H2014 context but they do appear to exhibit much lower replication rates than was observed empirically for the *MBNL1* locus. One possibility is that the contextual differences between the empirical analysis and the simulations incurs differences in replication rates; a second is that there is a mixture of false positives and true epistatic effects amongst those discovered in Hemani et al 2014; and a third is that there are additional statistical issues with the classical test that we are not aware of, that could inflate the replication rate.

## Measurement error in the *cis* additive causal variant

We then asked whether it is possible to avoid the inflation that we see in interaction tests. An intuitive approach would be to use a two-stage strategy, where first the additive effects are fine-mapped for the phenotype, and second the interaction search is performed with the fine-mapped variants included as covariates in the model. In the previous simulations, however, we observed that even when there are very small additive effects it is possible to find false positive interaction test statistics. This implies that if there is incomplete tagging of large additive effects by the fine-mapping strategy, we would fail to completely protect against inflated test statistics. This is confirmed through a basic simulation showing that interaction test statistic inflation occurs when the causal variant is included in the linear model, but there is measurement error of the causal variant (Figure 6).

To evaluate how this problem might transpire empirically, we performed a new set of simulations in which we constructed a phenotype using a variant typed in the UK10K sequence dataset as the *cis* additive causal effect. We then developed four datasets in which to perform the analysis16 - 1) retaining SNPs only present on Illumina CoreExome array, 2) variants imputed from this array data set to the HapMap2 reference panel17, 3) variants imputed from this array data set to 1000 genomes reference panel18, and 4) the original sequence data19. In each case we identified the top variant and tested for interaction against remaining SNPs. Figure 7 demonstrates that only when the sequence level data is available is it possible to prevent inflation of the test statistic.

## Additive effect heterogeneity

We have shown that the additive *cis*-causal variant must be measured without error and included as a covariate in order to avoid test statistic inflation of the interaction term. However, there is a scenario in which even this will not be sufficient. Typically, we assume that the estimate of the causal effect in a linear model represents a homogeneous influence of the variant on all individuals. However, if there is variation in the effect, meaning that the true effect varies across different individuals, we can only capture the average effect. Here, the residual error term becomes a mixture of variance not captured by the causal variant, and variance not captured by the average effect estimate of the causal variant20,21. We demonstrate through simple haploid simulations that if there is any effect heterogeneity across individuals, even when the causal variant is included as a covariate, the interaction term will be inflated (Figure 8).

## Discussion

We have shown that a large additive effect can contaminate standard approaches to obtaining genetic interaction test statistics. The problem arises because neighboring variants that imperfectly tag that additive effect are liable to exhibit higher mean and substantially higher variance of the test statistic. Ultimately this can lead to high false discovery rates, and those are likely to be correlated between datasets which can lead to some degree of replication.

To guard against this problem, it may be insufficient to fit the fine-mapped additive effects if there is any measurement error, suggesting that sequence level data may be required. We also show that genomic inflation factors are not always reliable metrics for detecting cases where false positives may have arisen. When applying genomic inflation correction to the H2014 discovery test statistics, 140 of the original 501 pairs remain significant, and this does not necessarily clarify whether they are real signals. In Hemani et al (2014) we argued that replication was a good standard for demonstrating statistic robustness. Our simulations demonstrate that this does not completely protect against these statistical issues, though the per-test inflation is almost uncorrelated between independent samples. Triangulating results for epistasis from statistical analysis alongside analytical approaches with non-overlapping limitations is of heightened importance in light of these results.

There is a long history of problems arising in genetic analysis due to the interplay between statistical tests and background genetic architecture being poorly understood or experimental design being misaligned. Linkage studies on complex traits have exhibited poor replication, and it was recently shown that under a polygenic architecture the test statistic could be inflated, thus the standard threshold was not sufficiently conservative22. Candidate gene studies also have a legacy of poor replication due to what is likely to be a combination of low power under a polygenic model and publication bias23. In the case of the *F*-statistic used for detecting epistasis, the problem of inflation that we describe here arises due to two forces. First, when there is imperfect tagging between a large additive causal variant and a nearby locus, the mean and the variance of the test statistic for interaction terms of the tagging locus will be inflated. Second, an exhaustive search for epistasis will allow the ascertainment of loci that have the appropriate tagging qualities to maximise test statistic inflation. This problem will exist even in a more targeted search for interactions with a known additive locus against other loci, if the true additive effect is not completely captured. Such a scenario can arise even when the causal variant is known but is poorly typed *e.g.* due to imperfect imputation.

How we reliably perform tests for epistasis going forwards remains a challenging question. We have shown that adjusting for fine-mapped *cis*-additive effects does not completely protect the interaction test statistic at tagging loci from inflation if genotyping accuracy is not guaranteed, nor if the assumption of homogeneous causal effects across individuals is violated. It does however improve matters substantially, and so we advocate that this should be done routinely.

A second approach is that for any loci that are detected with interaction effects, a permutation test of that specific interaction could be performed in which the phenotype and *cis*-variant are held constant while the *trans*-variant (*i.e.* the variant that is not tagging an additive effect) is shuffled. This would give an empirical distribution of the test statistic in the context of potential inflation. Non-parametric tests, if computationally tractable, may also be immune to this form of interaction test statistic inflation.

Over and above these approaches, using sequence data offers the most robust solution to protecting against inflation, though we warn that even here errors may remain where there is heterogeneity in the additive effect between individuals. If there is no large additive effect, as is the case with most complex traits and for most trans regions of ’omic variables, then the problem of the residual being a mixture of binomial and normal distributions is unlikely to exist.

## 

## Methods

### Genomic inflation in the discovery data

For each of the 501 interactions reported in Hemani et al 2014, we used the original discovery data to estimate the genomic inflation factor of the interaction test statistic, where we tested for interaction of the cis-locus against all trans-loci. This resulted in approximately 500,000 interaction test statistics (4 d.f. F tests) per analysis (varying depending on the number of SNPs on the *cis*-chromosome, as that was omitted for the test). We calculated the genomic inflation factor distribution of test statistics by obtaining the median p-value, converting it to a 1 d.f. chi-square value, and dividing by the expected median 1 d.f. chi-square value of 0.455.

### Simulations of discovery-replication scenarios

Here, our objective is to evaluate the expected behavior of replication of interaction tests under the null hypothesis that there is no interaction (and therefore any discovered interactions are false positives). To do this we create two datasets, one representing the discovery sample (n = 846) and another representing the replication (n = 2,131). We use genotype data from the Avon Longitudinal Study of Parents and Children (ALSPAC) study24,25 to create the two genetic datasets, such that realistic LD structures are present and there is genotype resampling between the discovery and replication. The ALSPAC data comprised 8871 children genotyped on the Illumina HumanHap550 quad chip genotyping platform and imputed to 1000 genomes reference panel (Phase 3 version 1). This was used to simulate the phenotype under an additive model, where a large additive effect was caused by a single variant.

#### Phenotype simulation

We want to simulate a phenotype that is due to a single large additive effect, and then perform interaction tests with the causal variant absent from the set of markers that are tested. In Hemani et al (2014) we reported *MBNL1* gene expression being influenced by several *cis-trans* epistatic interactions. In that scenario, rs67903230 was the fine-mapped additive cis-variant, which we will treat as the causal variant in these simulations. It was absent from the genotype data used to conduct the interaction analysis, and instead rs13069559 emerged as a cis-tagging variant, which showed interaction associations against variants on other chromosomes, and also replicated in independent datasets. We attempt to mimic this scenario here. We define the phenotype to be simulated as

where is the genotype value for individual at the causal variant. In this case we use the rs67903230 in the ALSPAC data. Fixing the variance of to be 1, the residual error term where additive effect . For each simulation replicate we sample the variance explained by the additive variant from .

#### Analyses

Once the phenotype was simulated using the causal variant, we were able to obtain F-statistics for the interaction term of the tagging variant (rs13069559) against every trans-variant (excluding those on the cis-chromosome). We retained only 502,510 autosomal markers to match the original discovery data, excluding those on chromosome 3 which is where *MBNL1* resides. The 4 d.f. interaction test was performed between rs13069559 and each of these markers sequentially. We did this both in the discovery and the replication dataset, so that we could compare the distributions of *F*-statistics between the two, where we expect variation to only arise due to resampling of genotype values and residual values between the discovery and replication data.

This process of creating a phenotype, performing the *cis-trans* analysis in the discovery and the cis-trans analysis in the replication, was repeated 40,000 times.

To mimic the discovery-replication process, for a particular simulation we tested if any *cis-trans* interactions (4 d.f. test) were significant at a Bonferroni corrected threshold, and then looked up their associations in the replication.

We refer to this set of simulations, where rs13069559 and rs67903230 are used as the cis variants, as ‘Scenario 1’. The resampling of genotype values between simulations was not possible here due to the limited ALSPAC sample size, though based on further simulations in the scenarios described below where some degree of genotype resampling is achieved, it is unlikely to have a major impact.

#### Further scenarios

Similar results were found in slightly modified simulation scenarios to the initial one described above. If all H2014 signals are false positives then the above simulation scenario represents a relatively bad scenario (meaning more susceptible to false positives), because the only one locus was analysed in which the genomic inflation factor was high and the replication rate was high. We now look at more potentially realistic scenarios below by expanding the set of loci analysed.

**Scenario 2:** Here we sample across a broader set of loci, using 277 cis-trans interaction examples from the H2014 signals. We used the variant with the largest cis-additive effect from the eQTLGen analysis26 as the causal variant when generating a phenotype, allowing the variance explained to range from 0 to 50% across the simulations. For each of the 277 examples, we performed 200 simulations in which residual noise was resampled, and a test for interaction between the cis-acting variant detected in H2014 against all trans SNPs was performed.

**Scenario 3:** We then built upon Scenario 2 further. We again used the 277 cis-trans interaction examples from H2014, but this time the causal variant explained a fixed amount of variance, based on what was found empirically in the eQTLGen analysis. Each of the 277 examples was repeated 200 times with resampled residual noise.

### *Cis*-adjustment simulations

Here, our objective is to evaluate the test statistic inflation when the cis-additive causal variant is included as a covariate in the interaction model, but there is some degree of measurement error in the causal variant. We used the haploid simulation scenario for simplicity, in which there are four variants the causal variant, the observed causal variant which has some measurement error , a tagging cis-variant which has some LD with the causal variant , and an unlinked variant . When , this represents poor imputation accuracy at the causal variant or a fine-mapped tagging variant in incomplete LD with the causal variant. We simulated a continuous phenotype where thus two thirds of the variance were explained by . The following statistical test was performed to test for interaction between the tagging *cis*-variant and the unlinked locus , after fitting the measured fine-mapped variant as a covariate

Simulations were performed for 1000 haploid samples, and over a combination of values for and where each scenario was repeated 500 times.

### Sequence data simulations

We performed further simulations based on sequence data to gauge the extent to which adjusting for the fine-mapped additive effects could be insufficient due to imputation error of the causal variant. This analysis follows largely the strategy described in Yang et al. (2015)16 using whole-genome sequence (WGS) data from the UK10K project19. We randomly sampled a sequence variant on chromosome 21 as the causal variant and generated the phenotype based on the additive model described in the ALSPAC simulation above. We varied the variance explained by the causal variant from 2% to 80% and repeated the simulation 540 times for each setting. The analysis was performed using four different data sets: 1) WGS data of a subset of variants in common with those on an Illumina CoreExome SNP array; 2) data from imputing the CoreExome array genotypes to the HapMap 2 references27; 3) data from imputing the array data to the 1000 Genomes Project references18; 4) the entire WGS data. In each data set, we first searched for the top associated variant based on an additive model and then tested the interaction effect between the top associated additive-effect variant and all variants on chromosome 22.

### Additive effect heterogeneity simulations

Typically, the causal effect parameter is treated as constant across all individuals. However, if there is heterogeneity in this parameter, such that linear models only estimate the average causal effect, then the error variance is a combination of variance not captured by the causal variant, and variance not captured due to misestimation of the per-individual effect size. The objective of this set of simulations is to demonstrate that even when there is knowledge of the causal variant and that causal variant is measured perfectly, test statistic inflation can arise due to assumption that the causal additive effect is constant across individuals. Let the causal effect be

and the phenotype for individual

where and . We constructed the error variance to add additional noise on top of that due to causal effect heterogeneity induced by , such that in a linear model explained of the variance. Thus, the variance due to effect heterogeneity , and the variance of the residual error

Across the simulations we fixed and , but used a range of . Therefore, in each scenario, the variance explained by the causal variant remains the same, but the proportion of the residual variance due to point estimation error varies with changing . Using this framework we tested for interaction between the perfectly measured causal variant and an unlinked locus using the model

Across the simulations, we used 1000 samples and performed 500 replicates per scenario

## Figures



Figure 1: Genomic inflation factors for each of the 501 SNP pairs that passed the significance filters in H2014. Values on the x-axis were calculated by converting the median p-values for the 4 d.f. test to chi-square values with 1 d.f., and dividing by the expected median 1 d.f. chi-square value of 0.455.

A screenshot of a social media post

Description automatically generated

Figure 2: Relationship between variance explained by the cis additive locus (x-axis) and genomic inflation factor for the interaction test statistic (y-axis) across three different simulation scenarios (rows of figures) as described in the Methods. Each point represents one simulation, where the genomic inflation factor was calculated from approximately 500,000 interaction tests. Scenario 1 involves a single locus at MBNL1 that was shown to have high inflation in Figure 1. Scenario 2 is a mixture of all loci, where the causal variance explained is allowed to vary. Scenario 3 is the same as in scenario 2, but the causal variance for each cis effect is fixed based on results from the eQTLGen analysis. Clusters of plots represent the variation in genomic inflation for a particular locus. We note that scenarios 2 and 3 appear to include loci that do not have systematic inflation, which is consistent with observations in Figure 1. Note that the relatively large variability of the genomic inflation factors is because one of the loci in the pairwise interaction test is always the same, which creates a correlation in the test statistics across loci, over and above a correlation due to LD

A picture containing object, measuring stick

Description automatically generated

Figure 3: The number of independent and significant interaction terms under an additive model (y-axis) with respect to the variance explained at the additive cis locus in the simulation (x-axis). Rows of plots represent different simulation scenarios as depicted in Figure 2 and described in the methods. For visual clarity we binned the x-axis variable into 30 classes. Each box represents the distribution of the number of discoveries for simulations within those bins.

A picture containing text, map, flock

Description automatically generated

Figure 4: Relationship between genomic inflation factor in the discovery (x-axis) and replication datasets (y-axis) where each point represents one simulation replicate. Rows of plots represent the simulation scenario (Methods).

A close up of text on a white background

Description automatically generated

Figure 5: Rate of replication of false positives in an independent dataset (y-axis) as a function of the variance explained by the cis additive effect (x-axis). Colours represent the replication significance threshold used, where ‘experiment’ is the one used in H2014 (p < 0.05/501), and Bonferroni and FDR pertain the multiple testing correction within simulation, as each simulation can give rise to multiple independent false positives. Columns of plots represent the simulation scenario (Methods).

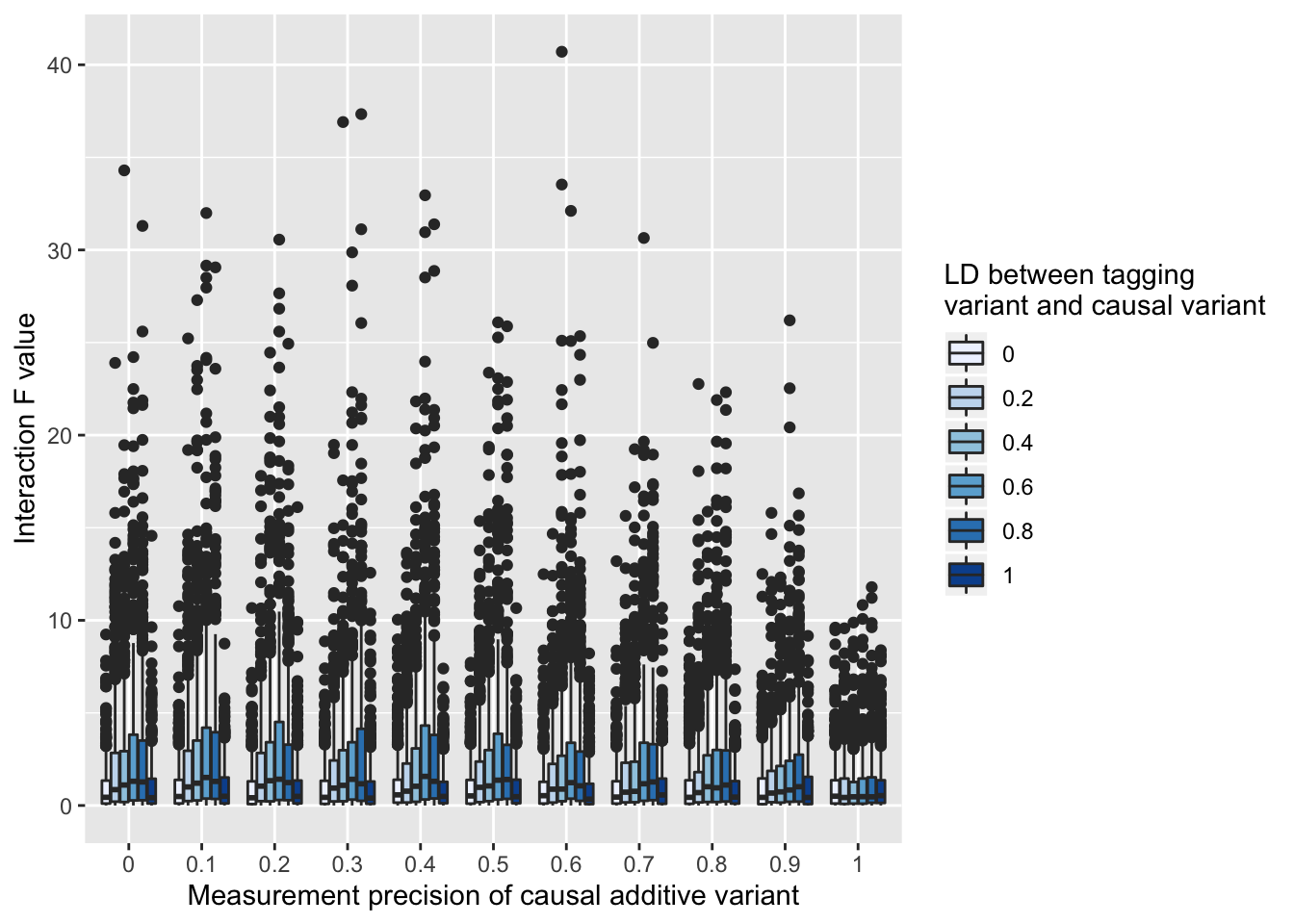


Figure 6: Test statistic inflation (y-axis) due to measurement error of the additive causal variant. Here a tagging variant and an unlinked variant for interaction, and the causal variant is included as a covariate. There is varying amounts of measurement error of the causal variant (x-axis), and LD between the the tagging variant and causal variant (colours). When there is no measurement imprecision, there is no test statistic inflation. When there is measurement imprecision the only scenarios in which there is no test statistic inflation is if the tagging variant (which is being tested for interaction) has LD of 0 or 1 with the true causal variant.

Figure 8: Genomic inflation factors (y-axis) estimated for interaction test statistics across a range of values for the variance explained by the additive effect (x-axis). Each line (colours) represents a different data scenario.

Figure 7: Genomic inflation factors (y-axis) estimated for interaction test statistics across a range of values for the variance explained by the additive effect (x-axis). Each line (colours) represents a different data scenario.

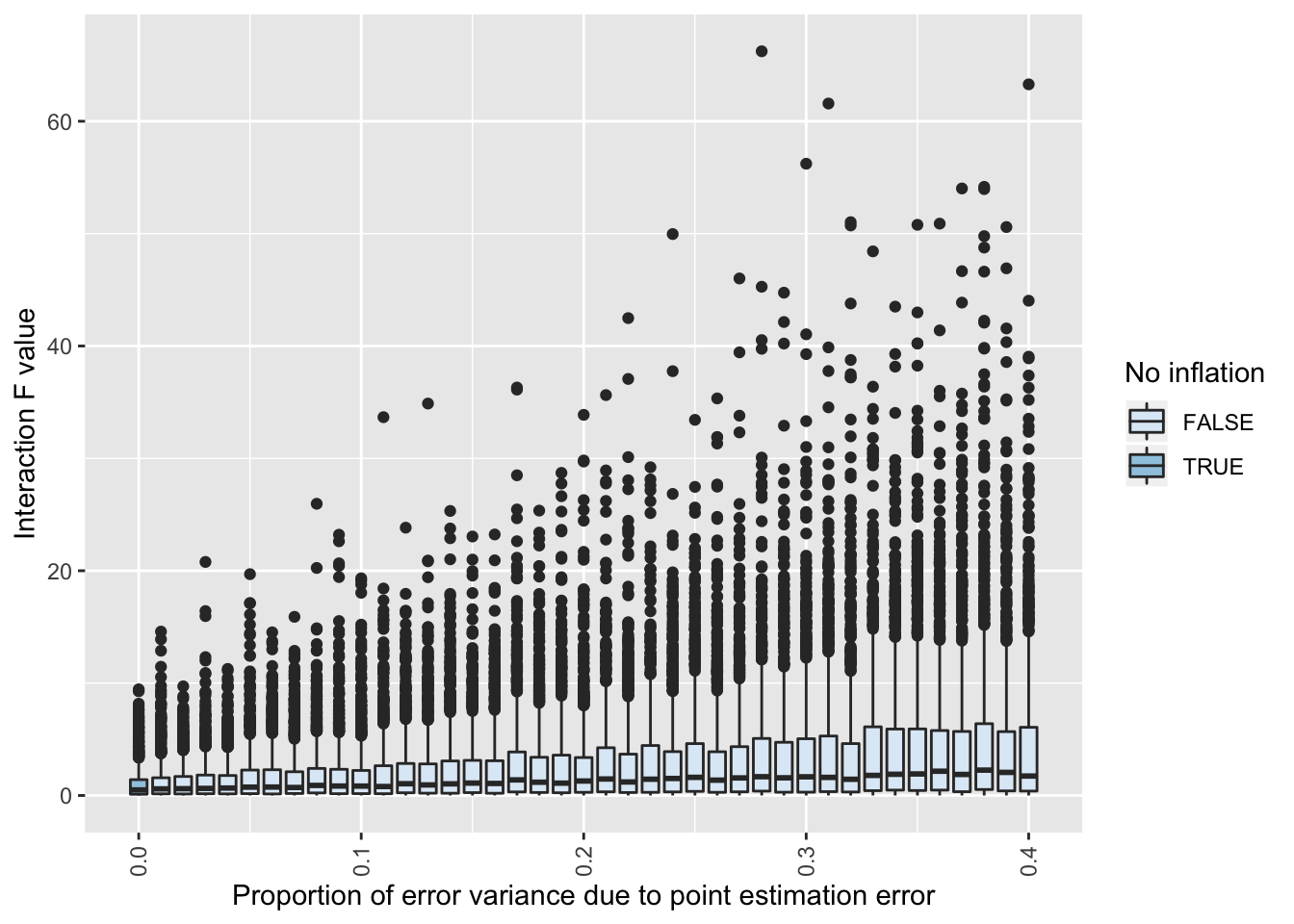


Figure 8: Heterogeneity in the additive causal effect across individuals can induce test statistic inflation (y-axis), even when the true causal variant is measured perfectly and included as a covariate in the linear model. Here the average causal effect explains 50% of the variance of the phenotype, but the proportion of this that is due to heterogeneity of the causal effect on the phenotype varies (x-axis). Only when there is no heterogeneity is there no inflation of the interaction test statistic (colours).

## Appendix: Distribution of a test statistic for interaction effects under an additive model

We consider 3 loci, where locus 1 is causal but not observed, locus 2 is correlated to locus 1 by LD, and locus 3 is uncorrelated to either locus 1 or locus 2. We consider a haploid system to keep the calculation as simple as possible. Let be the allelic dosage (which can take values of 0 or 1) at locus , with where the locus .

For LD parameter , the haplotype frequencies for loci 1 and 2 are

The LD correlation between loci 1 and 2 is . From these parameters, we can quantify the expected value at the causal variant as a function of the observed values at the linked and unlinked loci.

y2 y3 E(y1|y2, y3) = E(y1|y2)

0 0 p1 - D/(1-p1)

0 1 p1 - D/(1-p2)

1 0 p1 + D/p2

1 1 p1 + D/p2

Note that there is no main effect of the unlinked locus (y3), nor an interaction between the linked and unlinked loci (y2 and y3), consistent with the claim of de los Campos et al. (2019).

For a phenotype *Z*, the interaction test amounts to an *ANOVA* between the models and :

with the regression coefficient from the full (model b) and reduced model (model a). In this haploid example and a single additive by additive interaction term, the interaction test statistic is

Under the null hypothesis, and and being normally distributed, follows a central *F*-distribution with and degrees of freedom.

Now assume that the phenotype of interest is solely a function of , i.e. , and further that the additive effect of locus 1 is so large that . Therefore the distribution of is binomial.

Let the genotypic value of where , and let the counts for each genotype combination be:

The above *Fa,b* statistic is equivalent to the statistic /var( with a test for interaction defined as

with . The expectation of is 0, since from above, does not depend on , so the expected mean values cancel out. We now calculate

where the subscript *LM* stands for linear model, since the variance of  that is used in the test statistic is from a linear model, assuming homogeneous variance.

Under a linear model we expect that the error variance of each to be the same, and a pooled estimate is used. This assumption is violated when the trait has a binomial distribution. From the given haplotype frequencies:

and similarly

Each of the terms has a binomial variance

and frequencies

Putting all the terms together gives the exact variance of the test statistic as

where . In the linear model, the error variance is assumed to be the same in each cell of the 2x2 table and a pooled estimate is used. As a result

Thus, when using a linear model an incorrect variance of the interaction test is assumed, and this can lead to inflated (or deflated) type-I error rates. The ratio of the exact and linear model variances is the expected value of the linear model F-test. If we first rearrange,

we can obtain the ratio of test statistic variances

Unless , or the exact variance is different from that under the linear model. This also shows that the inflation (or deflation) term does not depend on the allele frequency at the unlinked *trans* locus.

We conducted basic simulations of this model, where we set , , and the LD was either or 0 to generate the values of , and . We then tested for interaction to obtain , and repeated the process 10,000 times to obtain a distribution of when there is LD between and , and when they are uncorrelated. When uncorrelated the mean and variance of are approximately 1 and 2, following expectation. However, when and are correlated the mean and variance of are approximately 3.4 and 23, consistent with the theory derived above - for the parameters used in this simulation .

The theory here is based upon an extreme case where the trait of interest is fully explained by a single locus and therefore has a binomial distribution. If the gene expression traits have large additive cis-eQTL that are not perfectly tagged then their distribution will be a mixture of binomial and normal errors. Nevertheless, as shown in the main text with both empirical data and simulations, the same principles as derived here hold.

## Acknowledgements

GH was funded by the Wellcome Trust and Royal Society [208806/Z/17/Z]. JEP is supported by National Health and Medical Research Council Fellowship (APP1175781). JY and PMV are supported by the Australian Research Council (DP160101343, DP160101056, FT180100186 and FL180100072) and the Australian National Health and Medical Research Council (1078037, 1078901, 1113400, and 1107258). We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

## References

1. Carlborg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* **5**, 618–25 (2004).

2. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* (2014). doi:10.1038/nrg3747

3. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genet.* **4**, (2008).

4. Hemani, G. *et al.* Detection and replication of epistasis influencing transcription in humans. *Nature* **10**, 249–53 (2014).

5. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234–237 (2013).

6. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era [mdash] concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

7. Wood, A. R. *et al.* Another explanation for apparent epistasis. *Nature* **514**, E3–E5 (2014).

8. Hemani, G. *et al.* Another Explanation for Apparent Epistasis. *Nature* **514**, E5 (2014).

9. Hemani, G., Knott, S. & Haley, C. An evolutionary perspective on epistasis and the missing heritability. *PLoS Genet* **9**, e1003295 (2013).

10. Brown, A. A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* 10.7554/eLife.01381 (2014). doi:http://dx.doi.org/10.7554/eLife.01381

11. Hemani, G., Theocharidis, A., Wei, W. & Haley, C. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics* **27**, 1462–5 (2011).

12. Cockerham, C. C. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**, 859–882 (1954).

13. Cordell, H. J. Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).

14. de los Campos, G., Sorensen, D. A. & Toro, M. A. Imperfect Linkage Disequilibrium Generates Phantom Epistasis (&amp; Perils of Big Data). *G3&amp;#58; Genes|Genomes|Genetics* **9**, 1429–1436 (2019).

15. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–12 (2011).

16. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* (2015).

17. The International Hapmap Consortium. The International HapMap Project. *Nature* **63 Suppl 1**, 29–34 (2005).

18. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

19. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

20. Angrist, J. D. Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. *Econometrica* **66**, 249 (1998).

21. Elwert, F., Elwert, F. & Winship, C. Effect Heterogeneity and Bias in Main-Effects- Only Regression Models. in *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (eds. Dechter, R., Geffner, H. & Halpern, J. Y.) (College Publications, 2010).

22. Hemani, G. *et al.* Inference of the Genetic Architecture Underlying BMI and Height with the Use of 20,240 Sibling Pairs. *Am. J. Hum. Genet.* **93**, 865–875 (2013).

23. Gaunt, T. R. & Davey Smith, G. ENOS and coronary artery disease: Publication bias and the eclipse of hypothesis-driven meta-analysis in genetic association studies. *Gene* **556**, 257–258 (2015).

24. Fraser, A. *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42**, 97–110 (2013).

25. Boyd, A. *et al.* Cohort profile: The ’Children of the 90s’-The index offspring of the avon longitudinal study of parents and children. *Int. J. Epidemiol.* **42**, 111–127 (2013).

26. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018). doi:10.1101/447367

27. Frazer, K. a *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–61 (2007).