**Testing for genetic interactions with imperfect information about additive causal effects**

Hemani et al.

[Supplementary methods 2](#_Toc56461643)

[Supplementary note 1: The traditional statistical test for 2-locus genetic interactions 6](#_Toc56461644)

[Supplementary note 2: Distribution of a test statistic for interaction effects under an additive model 7](#_Toc56461645)

[Supplementary Note 3: Test statistic inflation in the diploid case 9](#_Toc56461646)

[Supplementary Note 4: Replication rate of false discovery signals 10](#_Toc56461647)

[Supplementary Note 5: Measurement error in the](#_Toc56461648) *[cis](#_Toc56461648)* [additive causal variant 11](#_Toc56461648)

[Supplementary Note 6: Additive effect heterogeneity 12](#_Toc56461649)

[Supplementary figures 13](#_Toc56461650)

[References 21](#_Toc56461651)

## **Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.**

## Supplementary 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

Because of the scale of the original analysis it is difficult to mimic exactly 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.

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) study1,2 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 the 1000 Genomes Project 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 both the discovery and 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 analysis3 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)4 using whole-genome sequence (WGS) data from the UK10K project5. 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 references6; 3) data from imputing the array data to the 1000 Genomes Project references7; 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

## Supplementary note 1: 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 effects8. 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 problems9. 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 reason10,11. 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) for 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 effect of either of the variants12:

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 at locus A and is the marginal phenotypic class mean for genotype at locus B. Here the additive by additive, additive by dominance, dominance by additive and dominance by dominance terms are jointly assessed in the interaction term. This effect decomposition is fundamental to basic quantitative genetic theory13, and has been used routinely in the linkage study era and the GWAS era9,14. 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 is discussed in this paper applies to that approach also.

## Supplementary note 2: Distribution of a test statistic for interaction effects under an additive model

Having observed that many of the H2014 signals appear to exhibit systematic *F*-statistic inflation, we hypothesized that testing assumptions were violated due to the binomial distribution of the SNP variables involved in the hypothesis test. To construct a theoretical framework that mimics the data context of the *cis-trans* H2014 signals 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.

## Supplementary Note 3: 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 (Supplementary Methods). Because of the scale of the original analysis it is difficult to mimic exactly 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.

Supplementary Figure 2 shows that the genomic inflation factor related strongly to the variance explained by the additive causal effect. Supplementary 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.

## Supplementary Note 4: Replication rate of false discovery signals

In Hemani et al 2014 we found an overall 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 locus 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. Supplementary 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% (Supplementary 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 (Supplementary 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.

## Supplementary Note 5: 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 interactions. 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 (Supplementary 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 analysis4 - 1) retaining SNPs only present on Illumina CoreExome array, 2) variants imputed from this array data set to the HapMap2 reference panel15, 3) variants imputed from this array data set to 1000 Genomes Project reference panel7, and 4) the original sequence data5. In each case we identified the top variant and tested for interaction against remaining variants. Supplementary Figure 7 demonstrates that only when the sequence level data is available is it possible to prevent inflation of the interaction test statistic.

## Supplementary Note 6: 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 variant16,17. 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 (Supplementary Figure 8).

## Supplementary figures



Supplementary 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

Supplementary 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

Supplementary 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

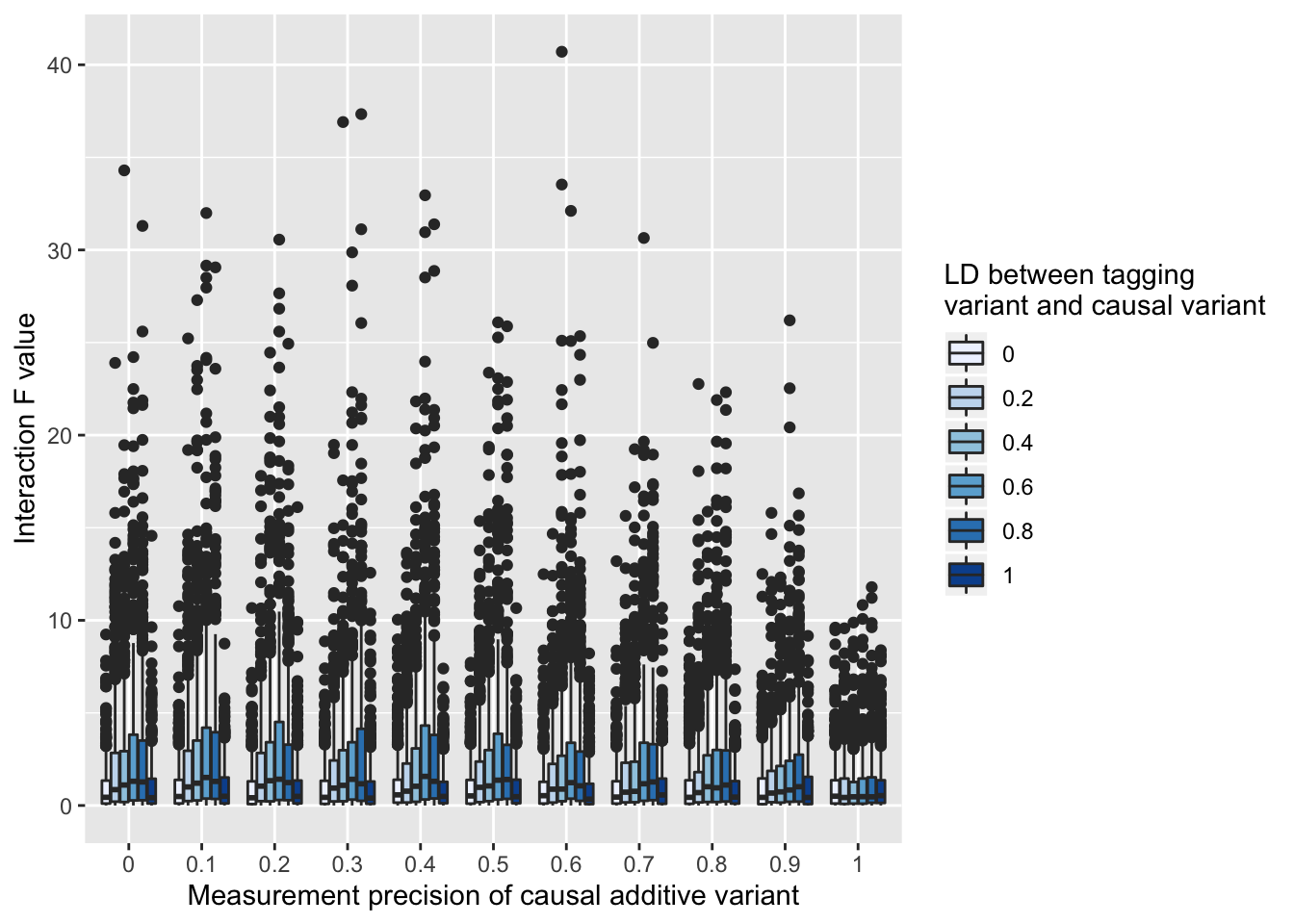
Description automatically generated

Supplementary 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

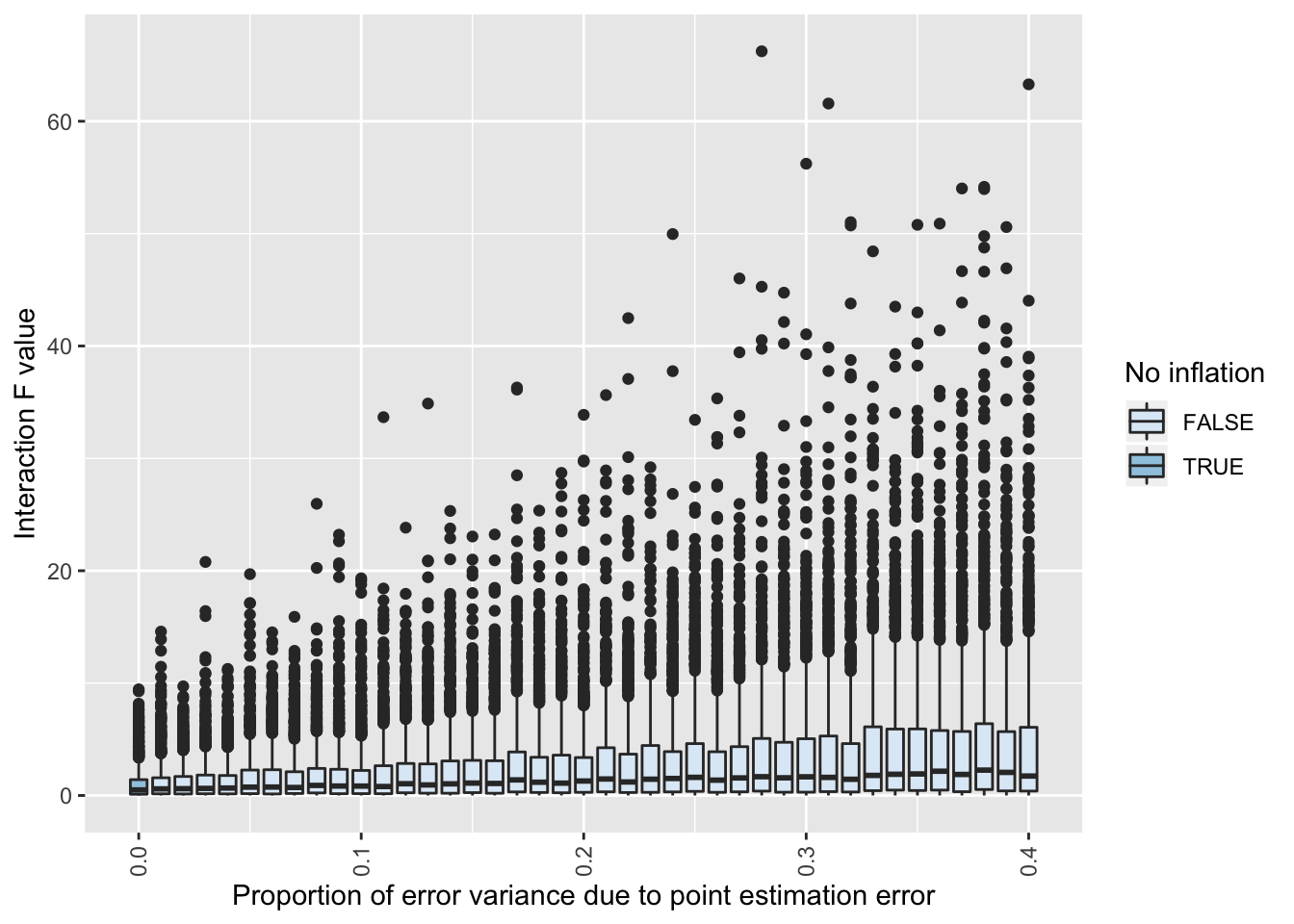
Supplementary Figure 5: Rate of replication of false positives in an independent dataset (y-axis) as a function of the variance explained by the cis additive effect (x-axis). Colours represent the replication significance threshold used, where ‘experiment’ is the one used in H2014 (p < 0.05/501), and 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).



Supplementary 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 are varying amounts of measurement error of the causal variant (x-axis), and LD between 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.

Supplementary 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.



Supplementary 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).

*Chart, scatter chart

Description automatically generated*

*Supplementary Figure 9: Relationship between 4 d.f. F-statistics (x-axis) and permutation test p-values (y-axis on the -log10 scale) under different models of genetic effects, where the F-statistics have been ascertained to surpass a p-value threshold of 1e-8. The top row of plots represents no additive causal variant while the bottom row of plots represents a large partially tagged additive causal variant. The left column of plots represents the null model of no interaction terms, while the right column of plots represents a true interaction term. Permutations are obtained by conducting 1000 4.d.f. tests in which the trans-acting locus is randomly shuffled without replacement each time, thereby maintaining the cis-acting additive effect. In the situation of large additive effect and no interaction term (bottom left) the permutation p-value is substantially attenuated compared to the discovery p-value, but there is a strong relationship between higher discovery test statistic and more extreme permutation p-value. Further, the pattern of relationship when there is both a true interaction term and a large additive effect follows a similar pattern to the null. Hence, the permutation approach as a sensitivity analysis is unlikely to be straightforward to interpret.*

## References

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

2. 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).

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

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

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

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

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

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

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

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

11. 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

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

13. 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).

14. 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).

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

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

17. 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).