Response to reviewers

Your Matters Arising "Testing for genetic interactions with imperfect information about additive causal effects" has now been seen by the reviewers, and their reports are included below. In light of the enclosed advice, we can offer online publication of a revised version of your comment, once you have addressed the points raised by our referees (please submit a point-by-point response) and have correctly formatted your paper for this section.

Specifically, we hope that you will agree to make the manuscript a bit more accessible to a broader readership, as recommended by Reviewer #3. As discussed before, we can be somewhat flexible with the word limit for the main text and limit on display items/SI items and if it is necessary to exceed these limits, then we can allow that.

We aim for a tone of scholarly debate in exchanges published in Matters Arising, so immoderate language and use of lengthy quotations from the original Nature paper should be avoided.

Detailed guidelines about the Matters Arising format can be found at https://www.nature.com/nature/for-authors/matters-arising. Note that the length of your main text should not exceed 1200 words; an extra 100 words can be added to the end of the text to describe the Methods if necessary. Up to two simple figures or tables are permitted, with captions of about 100 words each. A maximum of three more complex figures or tables may be provided as Extended Data.

Because we are particularly concerned that your comment should be accessible to the non-specialist readers of this section of the journal, care should be taken to ensure that any technical terminology is fully defined and kept to a minimum.

Titles should not exceed 43 characters, inclusive of spaces; assertions or statements must be avoided as the title should encompass the whole debate (i.e. your comment and the reply to it).

The opening paragraph should be recast in four short sentences. The first sentence sets the scene; the second summarizes the results of the Nature paper under discussion; the third presents your contradictory view/results; and the fourth states the implications.

The reference list should not exceed 15 references.

***Thank you for the feedback on this draft of the manuscript. We have now attempted to improve the accessibility of the text, specifically by adding the opening paragraph as suggested and rewording sections throughout.***

Referee #3 (Remarks to the Author):

I have previously commented on the retraction question. My comments now pertain to the submitted manuscript.

In this manuscript, the authors report on their understanding of what may have gone wrong in the 2014 epistasis analysis. As I commented before, the authors are to be commended for making an effort to get to the bottom of this -- it's certainly not fun to revisit old papers in this way.

The manuscript itself is of course more technical, and of much narrower interest than a typical Nature paper, but given the history of these articles I can be supportive of publishing a revised version of this in Nature.

I have three main comments.

(1) Both the manuscript and supplement are very hard to follow.

The main manuscript has sentences like "Under a simplified haploid model with completely penetrant additive genetic effects we find that the residuals from a linear model are a mixture of normal and binomial distributions" that can only be understood in detail if one either reads the supplement or already knows a lot about these models.

This topic is inherently somewhat technical but the overall presentation contains a lot of jargon and is overall not clearly presented (I'm including both the main text and the supplement in this comment). It's a lot of work for even a statistically-minded reader to wade through it and understand the key issues.

Overall this could be improved by reducing jargon; making the main points more clearly in both the main text and supplement; linking more clearly from paragraphs in the main text to corresponding sections in the supplement; making the order of the supplementary sections more natural so that one can read through in a more continuous way.

***Thank you for taking the time to review this revised version of our original paper describing the problem. We have now added an opening section which briefly summarises the entire saga. Readers who are interested in the details about the problem with the F-statistic, the ways in which it will manifest, and how different potential solutions might perform, can get brief summaries from the main text, and more detailed summaries from the signposted supplementary notes. We have tried to clarify wording throughout but it remains a technical document as that is the very nature of the topic at hand.***

(2) One important result in H2014 was the sign concordance between discovery and replication: "Sign concordance between the discovery and both replication datasets was observed in 22 out of the 30 significantly replicated interactions (expected value = 7.5 under the null hypothesis of no interactions, p = 3.76 × 10−8)."

But the present paper comments that hits "were seldom independently replicated at the Bonferroni threshold" and I cannot find any comments on the sign of replication. If this is mainly an effect of inflated variance, it would seem that sign replication might be random; is there a good explanation for why it does not appear to be?

***We do not have a good explanation for the rate of replication found in the original H2014 signals. In Supplementary Note 4 we state:***

***"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."***

***Throughout this saga we have avoided dogmatically defending our original analyses, amongst the co-authors opinion is divided on the extent to which the H2014 signals are true or false positives. But what is unanimous is that the widely used F-statistic is unreliable and can lead to false positives, and we initially wrote this paper and embarked upon this process with the main purpose that we wanted to warn the community of this issue.***

(3) Moving forward, readers may be less interested in the post mortem than in how to design future tests that are more robust. You touch on this at the end of the main paper but there's no corresponding section in the supplement to back this up with details.

***We agree that a novel statistical method would be ideal, but that is perhaps beyond the scope of a Matters Arising. We have conducted a further simulation to evaluate the efficacy of permutation analysis as a sensitivity analysis for discovered associations and draw attention to the difficulty in their interpretation - namely that F-statistic inflation appears to be correlated with the permutation test statistic (Supplementary Figure 9).***