**Reviewer 1**

In conclusion, this work provides strong evidence that pair-wise epistasis contributes significantly to variation in transcript levels, encouraging similar efforts applied to other complex traits including diseases. The topic of epistasis is of great interest to the field of complex trait genetics. The paper is in general clearly and succinctly written. The methodology is sound. The results are convincing.

We thank for reviewer for these positive comments

1. Abstract: “2.5% confidence interval of the distribution under the null hypothesis” ... I don’t think that this is what the authors mean. Should be rephrased here and throughput the text.

We have rephrased this to 2.5% confidence interval of the quantile-quantile plot of interaction p-values.

2. Abstract: “... within 2MB of regions ... “ Some contradictions here and throughout the text. Is it 2 or 2.5? According to Suppl. Fig. 11, its should be 2.5Mb

We have modified this to be consistent throughout the text.

3. Main text, page 3: “... 316 of the remaining 404 discovered SNPs ...” come confusiuon here and in other places. These are 404 SNP pairs not SNPs.

We have changed the text to specify SNP pairs where appropriate.

4. I am not sure that trimming SNPs based on LD within pairs completely avoided haplotype effects. Yet, as the majority of interactions are cistrans this is not a measure issue. It would be nice to provide an idea of the average distance (+ range) for cis-cis interactions.

The data on distance and LD was provided in Table S1, but we have included a second table that shows these statistics for cis-cis interactions only and summarized in the text.

5. Suppl. Methods: Equ. 2: Is it correct that the diagonal includes both $\sigma^2$ $A + \sigma^2 E$?

This has been altered.

6. Suppl. Methods: Page 4, line 1: How do the authors explain that the epistatic component was significant for a much larger proportion of SNP pairs for which one or both SNP had a highly significant marginal effect?

One of the conclusions of this study is that typically the epistatic variance for most SNP pairs is low, so we often only see those that reach significance if they have a relatively large main effect. We can speculate that we would expect to see more interactions without large main effects in larger studies, where obtaining a significant result does not depend on there being a large main effect.

7. Suppl. Fig. 1: I would have liked to see the right panel (null) generated not with random SNPs but with the reshuffled 434 or 404 SNP pairs. The legend refers to 2.5\% FDR ... Is this truly FDR?

We performed an analysis on 434 SNP pairs in the replication datasets that were constructed by reshuffling the original discovery SNPs, as was suggested. The figure has been updated with these results and they are consistent with the previous inference that randomly drawn SNPs are not enriched for interaction terms. We have also modified the text to say “2.5% confidence interval” instead of FDR.

Finally, despite the inclusion and use of Hi-C data, the paper really does not elaborate about possible molecular mechanisms that might underlie the observed epistatic effects. Were the trans-SNPs more often causing cis-eQTL effects than expected by chance alone? Were they (and SNP in LD) enriched in coding SNP? Were they located in the vicinity of specific types of genes (transcription factors)? Etc.

We have performed numerous bioinformatics analyses on the discovery SNPs to try and elucidate possible mechanisms underlying epistasis, and this has proven difficult. In these types of analysis the hypothesis is that there is a general trend across all SNPs for a particular mechanism or feature. (e.g. non-synonymous mutations, GWAS and known eQTL overlap, genome segmentation, chromosome interactions, transcription factor binding sites). We find strong enrichment for hematopoietic specific transcription factors for cis-SNPs, and cis-SNPs are enriched for various regulatory elements. They also frequently have main effects for cis-expression probes. However, we have been unable to find enrichment for trans-effects specifically. We believe that rather than speculating on possible biological mechanisms further study is now warranted here, which is beyond the scope of this manuscript.

**Reviewer 2**

Authors carry out epistasis analyses on human eQTL GWAS data. Across 7339 gene expression levels in blood in a cohort of 846 individuals, they detect 501 pairwise SNP interactions, some of which replicate in at least one of two replication data sets. Authors perform some bioinformatic enrichment analyses of interaction SNPs.

Abstract and introduction background: epistasis has been reported in many mapping studies of natural trait variation in multiple species, including for gene expression levels.

The question of whether epistasis influences complex traits in humans is a very widely debated question, and it remains unresolved due to an absence of any empirical evidence (e.g. numerous reviews on the subject). We think this study presents an important first demonstration of robust statistical evidence that epistasis does arise from natural variation, and that though there may be many instances of epistasis the contribution to phenotypic variance is likely to be quite small.

We respect that epistasis has been reported in non-human organisms, sometimes on a large scale. However, one reason that there is no scientific consensus on the question of epistasis in humans is because the reports of epistasis that do come from artificial selection, artificial gene knockout studies, and hybridization experimental designs in model organisms are largely not relevant in this case.

We are also aware that epistasis has been shown in humans, e.g. ankylosing spondylitis. But the reports from human studies are often presented on the observed scale, and epistatic terms will disappear when measured on the liability scale.

What is the evidence that transcription levels are less polygenic than higher level phenotypes?

The main reason that large effect sizes are observed in eQTL studies is that the mutational target size is often small. Higher-level phenotypes are often mediated by genetic effects via the modifications of multiple expression levels (in addition to other mechanisms), so by definition higher-level phenotypes are more polygenic. Furthermore, we have recently shown that congruence between orthogonal pedigree and SNP-based estimates of heritability is very strong when just the top eSNP is fitted, but can also be shown to improve considerably by fitting additional independent eSNPs from secondary and tertiary eQTL (Powell et al. PLoS Genetics 2013).

Page 3, "remarkable similarity in GP maps" needs to be quantified.

We agree with this, and have removed the statement from the text. We have attempted to quantify the similarity of the GP maps across cohorts in a statistically rigorous manner. We decomposed the 2 locus genotypic effect into orthogonal epistatic effects and tested the concordance of the direction of the effects between discovery and replication datasets. We believe that these new results further strengthen the conclusions in the study.

Page 4, "cis-cis" interactions are defined as "both SNPs on same chromosome as expression gene". These can be very far away and unlikely to be cis, especially if filter of any SNPs in LD is applied here.

Within the literature the term ‘cis’ is somewhat arbitrary. Often eQTL studies impose different thresholds for cis and trans effects, so the definition of cis and trans are statistically relevant. However, in this study we do not treat cis effects differently from trans effects in a statistical sense (i.e. the same threshold is applied throughout).

We had included the distance between SNPs in a SNP pair in table S5.

Interaction results between SNPs on same chromosome are frequently artefactual due to small sample size of "recombinant" haplotype classes because of LD.

We were very concerned about haplotype effects driving cis-cis interactions. For this reason we have filtered on LD r^2 and D’ in the discovery and replication datasets.

Page 4, genes and SNPs involved in very many interactions are not expected given the sparseness of interactions detected, and are likely to reflect technical artefacts.

We strongly disagree with this opinion. The release of genetic variation at multiple loci through ‘hub’ genes is a known phenomenon in artificial genetic studies, and is a key mechanism for epistasis e.g.

Carlborg et al. Nature Genetics 2006

Quietsch et al. “HSP90 as a capacitor of phenotypic variation”. Nature 2002

Bergman and Siegal. “Evolutionary capacitance as a general feature of complex gene networks”. Nature 2003

This is an indication of phenotypic robustness, something that is to be expected in complex traits. In terms of technical artifacts, we have been very careful in this regard. For example, we discarded any expression probes that mapped or partially mapped to multiple positions in the genome. Specific suggestions about how technical artifacts might arise that haven’t been addressed in the manuscript would be welcomed.

Bottom of page 5 and top of page 6, enrichement analyses are weakly informative at best. From weak enrichment of cis-acting SNPs vs trans-acting SNPs for transcriptionally active regions in haematopoietic cells it seems unreasonable to draw conclusions about their biological relevance.

We have amended the text

Page 5, there is no justification for applying interaction threshold to additive effects. Should match false discovery rates or effect sizes but not thresholds for classes with very different statistical properties in regards to multiple testing and power. Leads to huge underestimate of additive effects.

The purpose of this exercise on page 5 was to provide some indication of the contribution of epistatic effects *relative* to additive effects to phenotypic variation. Of course, ideally one would do this by comparing the contribution of all effects (e.g. through whole genome variance estimation), but this is not feasible for epistasis. So instead we posed the question, “At a given significance threshold, what is the ratio of additive to epistatic variance?” To this end we imposed the same statistical threshold for both additive and epistatic effects. We believe that this is a question of interest to the field. We are also aware that it has its limitations, and we have listed them in the same section.

Methods to identify epistatic QTL are confusing.

…

Significance threshold for the full vs null model is cited in main text on page 3 before stating that 501 interactions were discovered. This is not exactly appropriate, as this was threshold for full vs null, not the criteria used to determine if there was significant epistasis. It is unclear what "filters 1 and 2" are on (methods page 4).

…

It is not clear how filtering out SNPs with significant additive or dominant effects (methods page 3) is consistent with results in the first full paragraph of page 4 (main text), which notes many interaction SNP pairs with significant main effects.

We have rewritten much of the text that describes the statistical procedure, both in the main text and the supplementary methods.

Test of full vs null model should capture significant additive, epistatic, and/or dominance effects and post hoc methods could be used to disentangle which terms are contributing, but significance after post hoc filters hard to evaluate.

…

It is not clear how many of the 501 interactions are actually significant, nor what the false discovery rate is for this set.

…

An appropriate FDR threshold for the tests of the (full) model vs the (additive and dominance) model would be more informative than the Bonferroni threshold used for the post hoc determination of epistatic pairs.

In this manuscript we felt it was important to be highly conservative regarding identification of epistatic SNP pairs. To this end we employed Bonferroni corrections both during the discovery and also replication phases. Most inferences made in our manuscript are based on only epistatic pairs that are significant at a Bonferoni level in the replication datasets.

Nevertheless, we agree with these comments and estimating the type 1 error rate is important. We performed simulations to evaluate the type 1 error rate of the two-stage experimental design. This has now been included in the manuscript. We show that the type 1 error rate at stage 2 is dependent upon the (unknown) power at stage 1. Assuming that power is close to zero, using the Bonferroni threshold in stage 2 we would expect a type 1 error rate of 0.14, and assuming power of 0.5 the type 1 error rate is around 0.07.

Should at least use additive-by-additive epistasis model alongside the full model, to increase statistical power and generate better context via comparison to previous work, where this is what is standardly done.

…

What is the relevance of the statement that "patterns of epistasis used for statistical decomposition are not designed to resemble biological function" in the context of that paragraph (end of first full paragraph page 4).

We justified the use of the 8 d.f. test in the text – because empirical evidence for epistasis is unknown the best way to parameterize the search for epistasis is also unknown. By using an AxA model, a large proportion of the signals that we did discover would simply not have been uncovered. We have emphasised the point that orthogonal decomposition of 2 locus genotypic effects into AxA, AxD, DxA and DxD is a statistically convenient parameterization, and choosing just one as the model (e.g. AxA) has no biological justification.

How is the "null distribution of no epistatic effects" (bottom of page 3) determined?

The null distribution of no epistatic effects simply assumes that the distribution of interaction p-values will be uniform.

(top of page 4) Is the dependence on LD between observed SNPs and causal variants the most noteworthy explanation for the lack of replication between discovery and replication samples.

Assuming a low type 1 error rate, and relatively low statistical power, LD is likely to be a major reason behind the failure to replicate many of the signals.