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 the 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 thank the reviewer for this comment. We have now modified this to be consistent throughout the text.

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

Agreed. 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 cis-trans this is not a measure issue. It would be nice to provide an idea of the average distance (+ range) for cis-cis interactions.

We agree that the possibility of haplotype effects should be considered carefully. The data on distance and LD was provided in Table S1, but we have now included a second table (Table S2) that shows these statistics for interactions between SNPs on the same chromosome only and summarized the findings in the main text (line 151-153). The conclusion remains the same, i.e. that given the SNP distances of the cis-cis interactions and low LD it is unlikely that haplotype effects drive the results.

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

This has now 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?

We agree that this is an interesting question. 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 in addition to an epistatic effect. This is consistent with theory (e.g. Marchini et al. Nature Genetics (2005) 37(4) 413-417; Hill et al. PLoS Genetics (2008) 4(2) e1000008; Hemani et al. PLoS Genetics (2013) 9(2) e1003295). A limitation of any mapping study is that any inference on genetic architecture is based on ascertained results, in this instance the question of whether or not large epistatic effects exist without large main effects is unknown because we ascertain epistatic effects with large main effects.

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 appreciate this suggestion. 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. Identification of any clear biological mechanism has proven difficult and as such we have avoided speculation in the manuscript. In these types of analyses 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 we have performed all obvious bioinformatics analyses regarding a possible mechanism (and report these) and that we have a good balance between reporting the main findings on discovery and replication of epistatic interactions and performing post-hoc enrichment analyses that may point to mechanism.

Reviewer 2

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

We agree with the reviewer that epistasis has been reported in mapping studies, indeed a number of these have been cited in the manuscript. However, the question of whether epistasis arises from variation in natural populations to influence complex traits, and to what extent, is very much an open question because no such systematic search using this study design has ever been performed. We believe that this remains an important and unresolved question in which the scientific community holds a long-standing interest.

Systematic epistatic searches have been reported for complex traits in model organisms through artificial gene knockouts (e.g. for yeast in Costanzo et al. Science (2010) 327(5964) 425-431); for artificial line crosses and hybridisation (e.g. Bloom et al. Nature (2013) 494 234-237; Huang et al. PNAS (2012) 109 15553-15559); and even at the speciation scale (e.g. Breen et al. Nature (2012) 490(7421) 535-538). But we do not believe that reports of statistically robust and replicable findings of epistasis exist from genetic variation that has arisen in natural populations.

Though it is difficult to prove the absence of evidence, a number of high profile reviews on the subject have stated that empirical evidence for this question is indeed lacking. For example:

- Phillips, PC. Genetics (1998) 149(3) 1167-1171; Discusses the theoretical importance of epistasis over the history of quantitative genetics, and demonstrates examples from artificial studies on the importance of epistasis in evolutionary theory.
- Carlborg and Haley. Nature Reviews Genetics (2004) 5(8) 618-624; Calls for increased emphasis on mapping epistasis in complex traits and discusses methods that may be used
- Moore and Williams. American Journal of Human Genetics (2009) 85(3) 309-320; Discusses methods for detection of epistasis and translational possibilities to personalized medicine
- Phillips, PC. Nature Reviews Genetics (2008) 9(11) 855-867; Highlights a number of examples of epistasis from model organisms, including the example of coat colour in mammals
- Cordell, H. Nature Reviews Genetics (2009) 10(6) 392-404; Demonstrates the statistical challenges of detecting epistasis in humans and highlights that convincing examples do not exist
- Crow, JF. Phil Trans Roy Soc London B (2010) 365(1544) 1241-1244); Demonstrates that convincing examples do not exist and that based on inference from artificial selection studies and theory, the contribution of epistasis is likely to be small.

We believe that the results presented in this manuscript are the first to begin to empirically address the question of epistasis

- on a large, genome-wide scale using a sufficiently powered experimental design,
- in a statistically robust manner (including replication),
- and in human populations for natural trait variation and natural genetic variation.

We greatly respect that epistasis continues to be reported in non-human organisms, sometimes on a large scale (Bloom et al. Nature (2013) 494 234-237; Huang et al. PNAS (2012) 109 15553-15559; Costanzo et al. Science (2010) 327(5964) 425-431). 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 experimental designs that use artificial selection, artificial gene knockouts, hybridization of inbred lines, or model organisms are largely not translatable to the question of the influence of epistasis on human complex traits.

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

We agree that the use of the term 'polygenic' is slightly inaccurate in this case, and we have amended the text accordingly (lines 80-83). In the original manuscript we used the term in consideration of the much larger effect sizes identified in expression mapping studies in comparison to high order phenotypes such as common diseases. Heritable traits with large effect sizes imply that the mutational target size is relatively small, and therefore likely to be less polygenic.

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

We thank the reviewer for this suggestion. We have now removed the statement from the text. In addition, we have quantified 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. The results are given in Tables S3 and S4. They show that, using any one of several different methods of quantifying sign agreement between discovery and replication datasets, that there is a very significant enrichment for the epistatic effects in the discovery dataset sharing the same direction of effect in the two replication datasets. For example, taking the largest epistatic variance component of all 434 discovery interactions, 221 were in the same direction in both independent replication datasets ($p = 5\times10^{-31}$). We believe that these new results further strengthen the conclusions in the study, and have been discussed in the main text (line 117-119).

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.

We agree with the reviewer regarding the issue of the definition of 'cis'. Within the literature there is ambiguity over the term 'cis'. For example, some studies define cis regions as the same chromosome as the expression gene (Price *et al*. PLoS Genetics 2011 e1001317), and other distance from the Transcription Start Site (TSS) ranging from +/-250kb-2MB (Stranger *et al*. Nature Genetics 2007; Dimas *et al*. Science 2009; Nica *et al*. PLoS Genetics 2011). Often eQTL studies impose different thresholds for cis and trans effects, so the definition used for cis and trans is 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 have amended the results to define cis-SNPs as being within 1Mb of the transcription start site of the gene, and trans-SNPs being all others.

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. The data on distance and LD was provided in Table S1, but we have now included a second table (Table S2) that shows these statistics for cis-cis interactions only and summarized the findings in the main text (line 156-158). The median distance between interactions on the same chromosome was almost 2Mb, and the average was over 18Mb. Given the strict filtering on LD we agree with the first reviewer that it is unlikely that haplotype effects are driving these interactions.

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 respectfully strongly disagree with this opinion. The release of genetic variation at multiple loci in the presence of a mutation at a 'hub' gene is a known phenomenon in artificial genetic studies, and is a key mechanism for epistasis e.g.

- Carlborg, O et al. Epistasis and the release of genetic variation during long-term selection. Nature Genetics 38 , 418-420 (2006)
- Quietsch et al. HSP90 as a capacitor of phenotypic variation. Nature 417, 618-624 (2002).
- Bergman and Siegal. Evolutionary capacitance as a general feature of complex gene networks. Nature 424, 549-552 (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. This has been explained in more detail in the supplementary methods (lines 150-161). In addition, if there were to be any study-specific technical artifacts then we would not expect them to be replicated, whereas our results clearly replicate in independent samples. Nevertheless, we have amended the text to state that although we have used strict quality control, it remains possible that technical artifacts may lead to the observation of statistical interactions (lines 94-97).

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 agree that caution is required and accordingly we have kept conclusions about this to a minimum. The main conclusion that we draw is that because there isn't enrichment for particular annotations for trans-acting SNPs then it is possible that the effects of cis-acting SNPs can be modified in diverse ways, rather than through one particular mechanism. We have clarified the sections where this is mentioned (e.g. line 182). In general, we have tried to balance reporting the main findings on discovery and replication of epistatic interactions and performing post-hoc enrichment analyses that may point to biological mechanisms.

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.

We thank the reviewer for this comment, and we agree that using p-values is not the best metric for comparing the relative contribution of epistatic and additive effects, and it led to an underestimate of the additive effect contribution. As suggested, we have modified this section to use the proportion of phenotypic variance explained as a threshold for comparing additive and epistatic effects instead of using p-values. The minimum estimated epistatic variance of the 501 discovery interactions was 2.1%. We found that 1848 eQTLs in the same study had an additive effect explaining at least 2.1% of the phenotypic variance, and that the total phenotypic variance explained by additive effects at this threshold was approximately 10 times higher than the total phenotypic variance explained by epistatic effects (main text lines 193-202; supplementary methods lines 295-314). We believe that this has greatly improved this section of the manuscript.

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 thank the reviewer for this comment. We have rewritten much of the text that describes the statistical procedure, both in the main text and the supplementary methods (line 85-100 of the manuscript and from line 68-123 of 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.

For this study 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. Hence we have used a very stringent and conservative testing procedure.

We agree with the reviewer's comments, and that estimating the type 1 error rate is important. We performed additional simulations to evaluate the type 1 error rate of the two-stage experimental design. This has now been included in the manuscript (lines 90-91 main text; lines 124-141 methods, Supplementary figure 1). We show that the type 1 error rate at stage 2 depends 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. Therefore, we believe that the type 1 error rate of the stage 2 discovery SNPs is likely to be higher than 5%, but is actually still low. We have amended the main text to include these estimates of the type 1 error rate at the discovery stage (lines 90-91).

We would like to reiterate that we focus our conclusions about the detection of epistasis, not on the discovery stage, but on the fact that there is replication in independent datasets.

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 thank the reviewer for this comment. Because empirical evidence for epistasis arising from natural variation is absent the best way to parameterize the search for epistasis is also unknown. By using an AxA model, one is explicitly excluding epistatic effects that are driven by AxD and DxD terms. This analysis was intended to be a survey of epistatic effects, and a large proportion of the signals that we did discover would simply not have been found if we parameterized on AxA only. We believe that this justifies the use of the 8 d.f. test. Performing the entire analysis again using AxA answers a much narrower question, and in addition to it being computationally unfeasible at this point, it also narrows the scope of the study.

In the manuscript we tried to explain that the statistical decomposition of 2 locus epistatic effects into orthogonal effects (AxA, AxD, DxD) is simply a statistical treatment of the model, and that by choosing to use only AxA to *search* for epistasis does not have a biological justification. This section has been re-written for clarity (lines 147-150).

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. This has been included in the text at lines 111-112.

(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.

We thank the reviewer for this interesting question. The question of LD is important because for additive variance the power to detect an effect is proportional to r^2 whereas for epistasis it is proportional to r^4, r^6, r^8. We have performed additional simulations to attempt to quantify the effect of LD between observed SNPs and causal variants on replication.

In the first simulation we wanted to answer the question of what the sampling variance of LD $\hat{r}$ is, given some population value of r. We observed that when true population r is high (e.g. $r^2 > 0.9$), as is expected in the discovery sample because otherwise detection power would be low, the sampling variance of r^x increases as x increases. This is important because if the significant epistatic effects in the discovery were ascertained to have high $\hat{r}$ between causal variants and observed markers, then higher sampling variance of increasing x would result in lower sampling $\hat{r}$ in an independent dataset.

This inference was tested in simulation 2. Here, we used 1000 genomes data to construct a scenario where we have ~500,000 observed markers and 100,000

unobserved causal variants. We show that if we ascertain for high $\hat{r}$ between observed markers and unobserved causal variants in a discovery sample, then the average decrease in sample $\hat{r}^x$ increases as x increases. As a direct consequence statistical power of replication of epistasis is likely to be lower than for additive effects.

The results for these simulations are shown in Supplementary figures 7-9, and the main text has been changed to describe these results (lines 122-135). A detailed description of the methods is provided in the supplementary methods (lines 234-283). We believe that the inclusion of these further analyses has improved the manuscript, and offers some explanation for why the epistatic signals replicate at a lower rate compared to additive effects. We thank the reviewer again for raising this point.