Reviewer #3 (Remarks to the Author):

My responses to the rebuttal are in red

This study uses a cohort of diversity outcross mice to perform a novel mediation analysis of obesity-related traits. The two main findings are that (I) distal eQTLs (ie trans eQTLs) seem to play a larger role in controlling transcripts relevant to obesity traits than do local (cis) eQTLs, and (ii) it is possible to define composite measures of genotype and transcriptome and phenotype and perform a mediation analysis on the composite measures, thereby establishing causality.

Both of these findings are interesting and noteworthy and would be of interest to the readership of Nature Comms. However, I also had a large number of queries about the manuscript. Whilst most of these are related to improving the presentation, there are a two more substantive queries which need to be dealt with satisfactorily.

Major points:

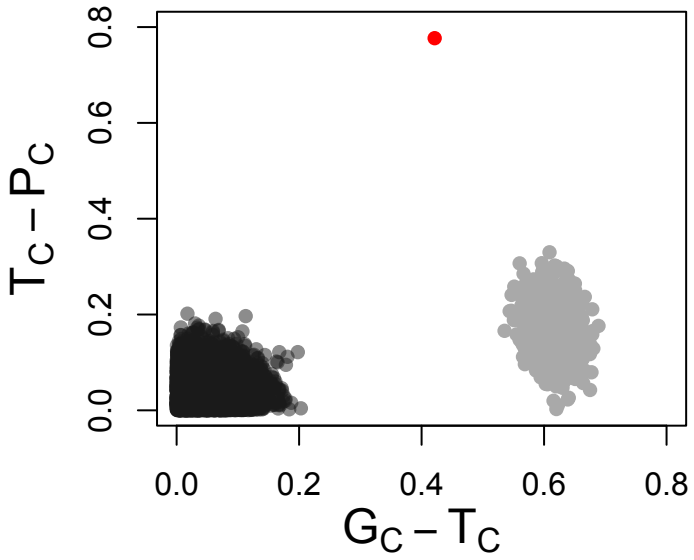
(i) Is the permutation procedure (Fig 3B,C) valid? It ignores relationships between individuals across omic levels (ie assumes all mice are exchangeable) and so might inflate the apparent significance. If I have understood correctly, the permutation procedure is performed such that it destroys any correlation between the three omic levels – including the existence of any QTLs or eQTLs, and this is likely too harsh a null hypothesis. It might be better to use multivariate generative models to simulate sets of genotypes, transcripts and phenotypes and evaluate performance on those rather than using permutation. An alternative might be to ask how unusual are the correlations cor(P_C , T_C) etc for the optimal choices of weights compared to randomly sampled weights using the unpermuted data, possibly augmented with a distribution fitted to the random correlations. In any event, that authors should justify their choice of permutation strategy and explain why it supports their thesis.

> We apologize for the lack of clarity on this critical point. The permutation test was performed by only permuting the transcriptomes. The correlations between genome and phenome were preserved in this test. Thus, we preserved the effects of any QTLs while breaking the association between eQTLs (and all other genetic effect on transcription) with QTLs. The essential goal of this procedure was to determine whether it was possible within a random transcriptome to identify a spurious transcriptomic profile that appeared to mediate the true genotype-phenotype correlations as well as our optimized transcriptomic profile in the true data. Our permutation test definitively rules this possibility out; the true transcriptome is more highly aligned with the genotype-phenotype map than expected at random. We stress that this is nontrivial, given that canonical correlation analysis (CCA)-based approaches have problems overfitting because they involved estimating the inverse of a covariance matrix and, hence, can

detect such spurious components. We have added clarification of this point, which can be seen in lines 176-179 in the manuscript with tracked changes.

The suggestion to consider the performance of HDMA under generative model hypotheses is an important one, and is the subject of ongoing research, but is beyond the scope of this paper. In this study, we use HDMA in a purely descriptive mode.

We randomly sampled the unpermuted weights 10,000 times and recalculated the correlations cor(G_C, T_C) and cor(T_C, P_C) as shown in Figure 3C. The results are shown below. We think it would add more confusion than information to add this to the manuscript figure, but shows clearly that the correlations identified by HDMA are well outside the null distribution. The original permutated values from the manuscript in Figure 3 are shown in gray, and the observed correlations are shown in red.



Thanks for clarifying the permutation procedure. So to summarise, the relationship between genotype and phenotypes (ie QTL) is preserved by permutation but there no longer any relationship between transcriptome and either genotype or phenome, (but presumably the internal gene expression network correlation structure is preserved). Under this permutation procedure we would not expect to see any mediation, therefore. Whilst I understand the rationale for this approach, it is very severe. It does not really address the question of whether the mediation observed (ie the correlation of T_C and P_C) is unusually strong, given the presence of eQTLs; is T_C better correlated with P_C than might be expected given the existence of lots of eQTLs.

At the very minimum, the manuscript should be very clear as to what the permutation procedure is demonstrating and its limitations. Ideally a different approach should be

adopted, although I understand that this would be a significant amount of work and quite difficult.

(ii) I'm a bit puzzled by the use of the CC-RIX mice as a validation set. It appears that actual body weights were not measured in the CC-RIX (why not? – weight is a standard phenotype) so they were imputed from local transcript data, which complicates their use for validation and makes it far less convincing. I really don't see what they add to the study. Why not simply keep back a random 10% of the DO mice, train the models on the 90% and test the predictions into the 10%, for many random samplings? (ie the standard machine learning cross validation procedure).

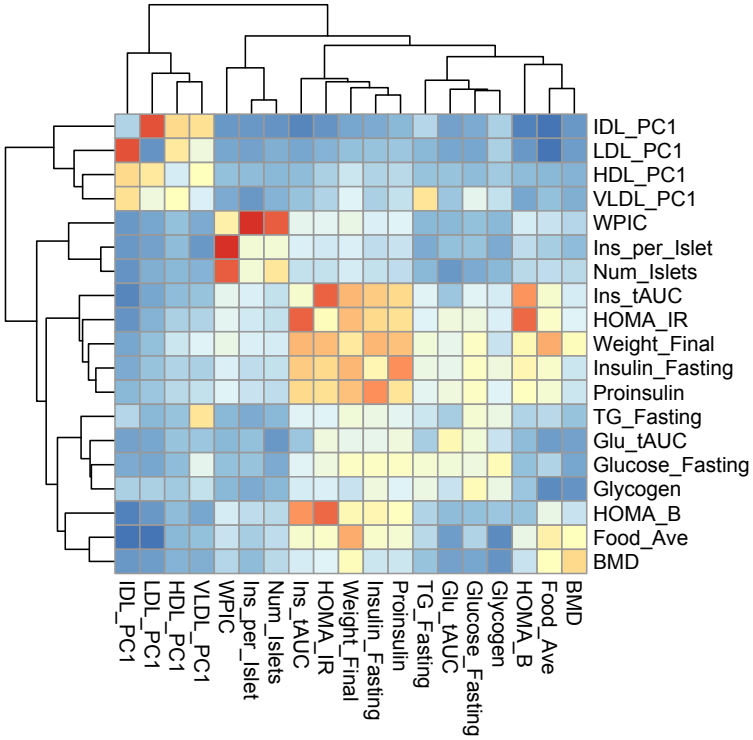
> We apologize for the lack of clarity on this point. Body weight was measured in the CC-RIX. We used measured body weight as the ground truth validation in the CC-RIX population. We estimated the metabolic disease index (MDI), which is largely based on body weight from the DO and compared it to the actual body weight as the validation (Figure 7B). To estimate MDI, we either used the measured transcriptome in the CC-RIX, or a predicted transcriptome based on local genotype. While the MDI based on the measured transcriptome correlated well with body weight, the MDI based on the locally imputed transcriptome was not correlated with body weight. We interpreted this result as support that genetic effects on genotype are not mediated through local gene regulation. It was important to use the CC-RIX data as opposed to validating within the DO data because the point of local and distal regulation is particularly critical when translating results between populations with different allele structure. We suggest that there is a failure of human TWAS results to translate between human populations because genetic effects are mediated through distal gene regulation, which are dramatically different across populations with different allele structures. In this mouse experiment, the two populations shared ancestral haplotypes, but had dramatically different allele structure, thus allowing us to tease apart the effects of local and distal gene regulation on phenotypic effects.

Thanks for the clarification. I still think there is a case for a training/test set analysis within the DO, in addition to the RIX, because of the challenges interpreting the replication experiments due to differences in diet, husbandry etc between the two populations. That said, it is quite impressive that the RIX replicates the DO despite all these differences and using RIX body weight as a surrogate phenotype.

Minor Points:

(i) Figure 1 is very good, except Figure 1G could be improved if the upper triangle of the heatmap displayed genetic correlations and the main diagonal heritabilities.

This is an interesting way to add more information to the plot. We tried it (shown below), but then realized that it might be confusing to have two different types of data in a single heat map. Because most of the cells show Pearson correlation, but the diagonal has very low values, we think this might cause more confusion, especially since the heritabilities are shown in Figure 1F.



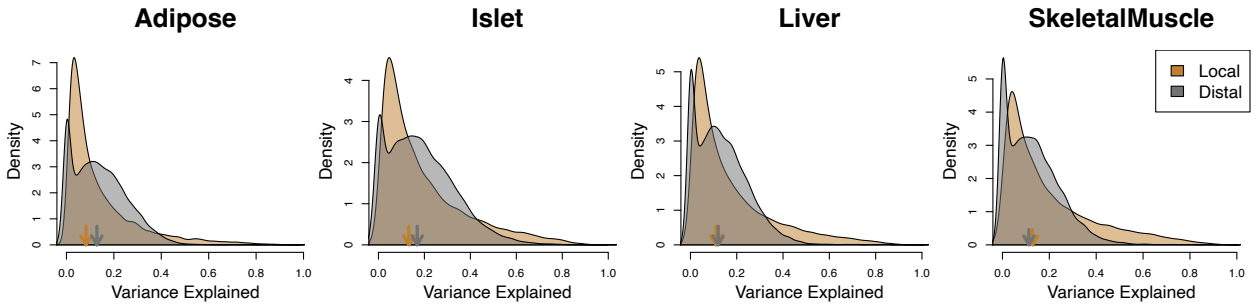
The Figure shown above in the response does not correspond to what was requested – I was asking for the upper triangle to be the genetic correlations and the lower triangle the phenotypic correlations, in addition to the main diagonal being the heritabilities. The Figure seems to have just the genetic correlations.

(ii) Figure 2 could be significantly improved by replacing the violin plots in Fig2A with overlapping distributions as in Figure 1A,B. Fig 2B is informative but I think the use of linear regression is not the best way of showing the shapes of the distributions for Local and Distal are different. Clearly a straight line does not fit any of the data very well. Can the authors think of another measure which quantifies the fact that strong local eQTL are more likely to have small trait correlations than strong distal eQTL?

We tried using overlapping distributions (below), but it became difficult to compare the distributions across the tissues. We would like Figure 2A to show a direct comparison of the variance explained by local and distal eQTLs across tissues, so we reverted to the violin plots.

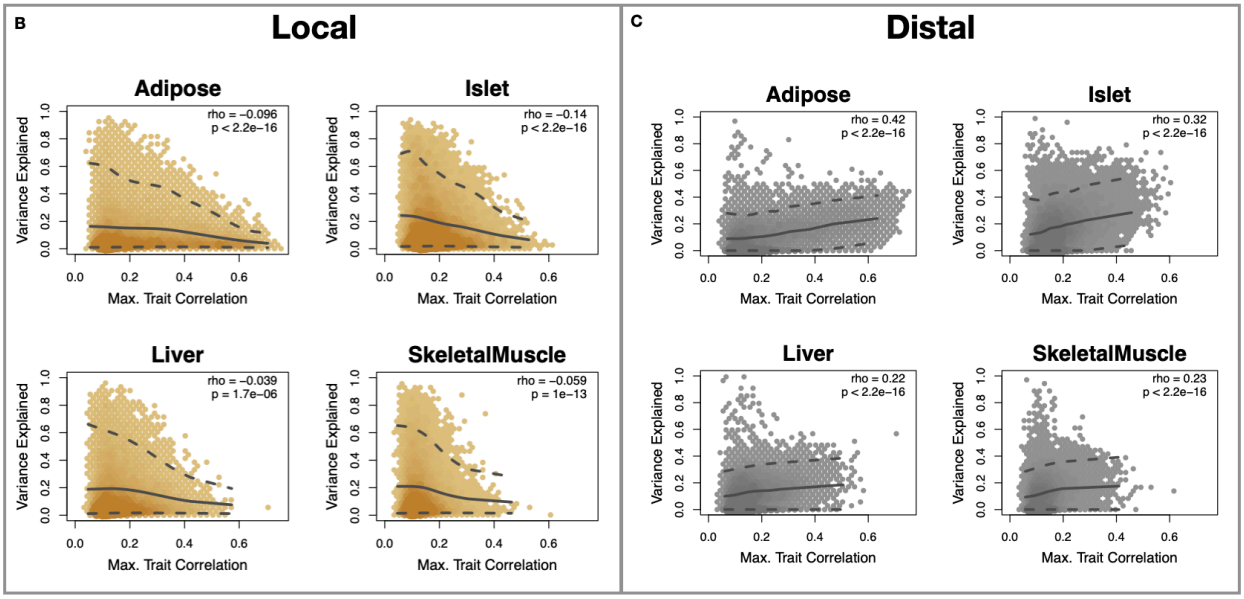
This is a matter of personal preference, so I don't insist on the use of overlapping distributions, but I do find the violin plots to be so similar it is hard to draw any

conclusions about the distributions. Incidentally the Figure legend now seems to be adjusted for the overlapping distribution plots, not the violin plots.



Figures 2B and C have been changed to avoid the linear test. We binned the transcripts into centiles based on their variance explained. We then calculated the mean and 95th and 5th percentiles of their maximum trait correlation. We smoothed these values using splines. The top and bottom lines in each panel shows the 95th and 5th percentile respectively and the middle line shows the mean. We reported statistics from Spearman rank correlation tests.

This seems OK



(iii) Fig 3 is hard to follow, particularly since it mentions Kernelization which is otherwise not mentioned in the main paper (it's mentioned in the Methods) Not sure what G_K, T_K , P_K signify.

We have added a short description of the kernelization terminology to the main text where we describe high-dimensional mediation. The new text appears in Lines 150-156 of the manuscript with tracked changes.

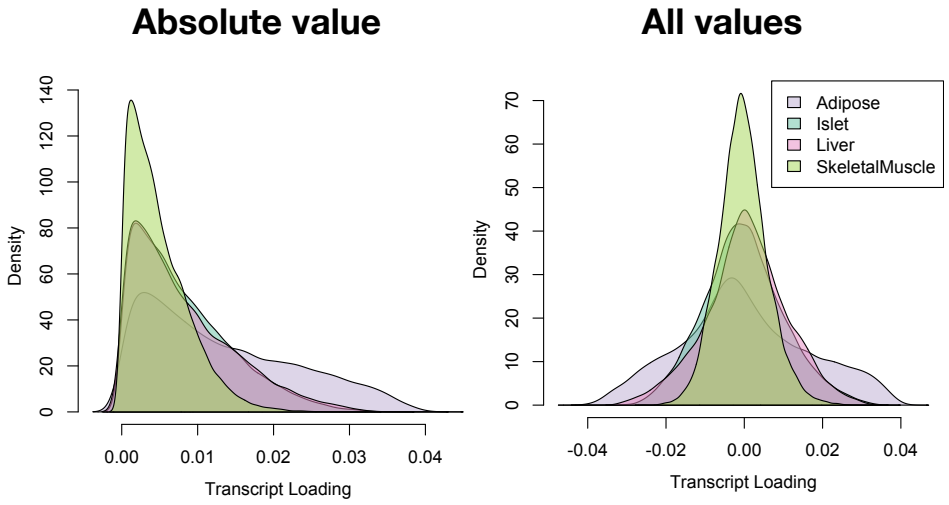
Thank you

(iv) Figure 4 D – Please include scatter plots of P_C vs T_C, G_C vs T_C, G_C vs P_C - this is surely key to understanding. Please replace violin plots with overlapping distributions as in Figure 1A,B

We tried using overlapping histograms to compare these distributions, but the distributions are so similar that the figure was difficult to read (see below). Instead, we replaced the distributions with separate panels showing the observed loading distribution for each tissue compared with the null. We shaded the area of the distributions that were more extreme than the null distribution and noted the number of genes in each extreme group. This adds a little more information to help distinguish the distributions.

The scatter plots took up a lot of room in Figure 3 without adding much information, so we have created a new Supplemental Figure S3 to show the details of these correlations.

I think the scatter plots are an excellent addition and should be in the main paper and the violin plots/overlapping distributions in the supplement.



(v) Fig 5 is good. It would be helpful to define exactly what is meant by TWAS in this study. I suggest a completely different color ramp is used to indicate tissue type from that used to indicate heritability – it's a bit confusing. It would also help to report the p-values of the t-tests for comparing the heritability distributions for distal vs local (in the three inset boxplots in the Fig 5)

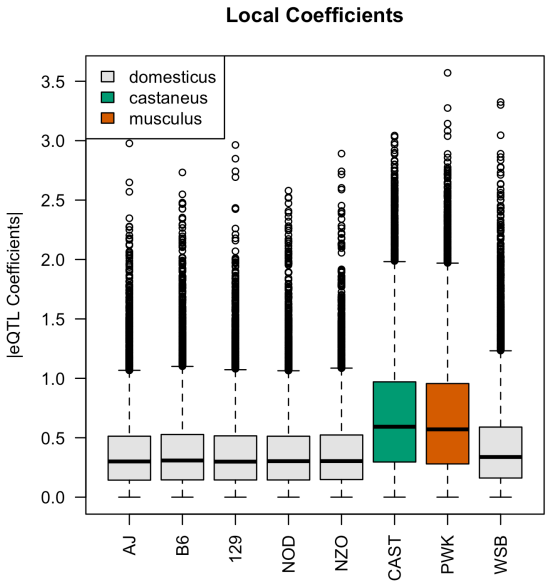
We have added a short description of our TWAS procedure to the results section that discusses Figure 5. We have also changed the colors for the tissues to select colors that do not overlap with the heat maps or the local/distal color scheme. We added t-test p values to the legend of Figure 5 for each of the box plots (page 14 in manuscript with tracked changes).

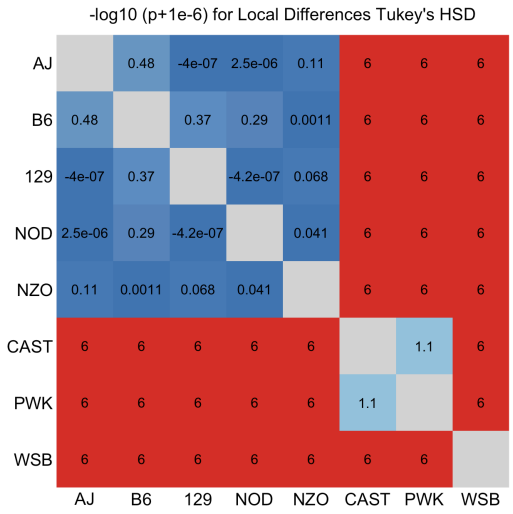
Thank you

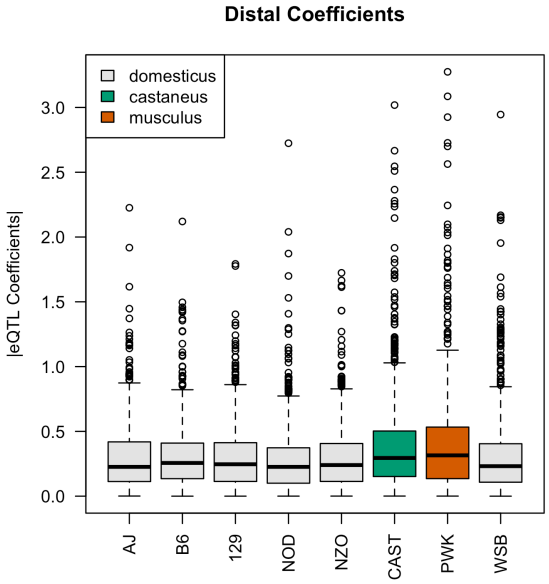
(vi) Given the diverse ancestry of the DO, involving alleles from three different murine subspecies, it would be interesting to know if the distal or local eQTLs more often involved alleles segregating between subspecies.

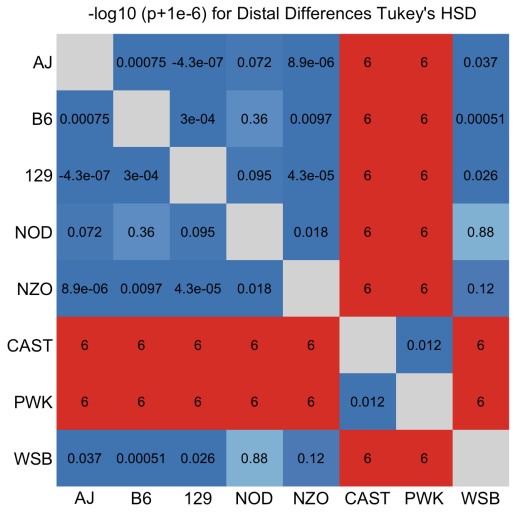
This is an interesting question and has been addressed extensively in the literature (PMIDs 25230953, 21406540, 34849860, 34425882, 33264334, 31961859). Because eQTLs were only an incidental piece of this manuscript (presented only in supplemental figures) and because this issue has been examined previously, we performed a limited analysis that is presented only in this response letter. We collected local and distal allele coefficients for all haplotypes. A comparison of the local and distal eQTL coefficients shows that the *castaneus* (green) and *musculus* (red) haplotypes tend to have larger local coefficients than the *domesticus* strains (gray). An ANOVA with Tukey's Honestly Significant Difference showed that the *castaneus*, *musculus*, and the wild-derived *domesticus* strain (WSB) had significantly stronger local eQTL coefficients than the other *domesticus* strains (top figures). For distal coefficients (bottom figures), both *castaneus*, and *musculus* strains had significantly higher coefficients than all *domesticus* strains.

This is a satisfactory response – I think this is an interesting finding, possibly worth adding as a supplemental figure.









(vii) Is there a reason for preferring the nomenclature “Local vs Distal” instead of the more usual “cis vs trans”?

We avoid the usage of cis and trans terminology because these terms have specific biochemical definitions that are not completely captured by physical location on the genome (See Box 2 in PMID: 18597885, ref 22 in the manuscript). Because we are not evaluating the biochemical nature of variants in this study and are classifying variants only by genomic position, we use the terms local and distal. We have added text to the introduction to define these terms and discuss the difference between cis/trans and local/distal terminology.

Although I prefer the use of cis/trans, so long as the paper defines these terms carefully this seems satisfactory

(viii) The statements around line 280: The mean loadings for alpha-cell specific transcripts were significantly greater than 0, while the mean loadings for delta- and endothelial-cell specific genes were significantly less than 0 (Fig. 8B)
The study does not give p-values for these statements – Fig 8B appears to show boxplots which don't give an indication of significance. There are several places in the MS where boxplots are used as evidence of "significance" without a formal p-value being provided.

> We apologize for this omission. We have added p values to the text for each of these significance claims.

Thank you

(ix) What were the exact criteria for calling eQTL (local and distal) and pQTL? Were different significance thresholds applied for local vs distal? The methods are vague – why does a LOD score threshold of 8 equate to a p-value of 0.05 (and what does this p-value mean – it is genome wide significance?) Surely it would be better to use an FDR-based threshold.

> Local eQTLs were defined as eQTLs within 4Mb of the transcription start site of the encoding gene. We have added this definition to the methods. We used a nominal p value threshold of 0.05 as a permissive and arbitrary cutoff to compare basic stats of local and distal eQTLs. The eQTLs were a marginal component of the manuscript (only presented in supplemental figures), and changing this threshold did not change the general conclusions about eQTLs that we presented. Namely, local eQTLs outnumber distal eQTLs and local eQTLs tend to be shared across tissues whereas distal eQTLs tend to be tissue specific. This study did not include protein data and we therefore do not report any pQTLs.

I don't think a 5% cutoff is at all reasonable. The manuscript does not even report the threshold or the numbers of QTL in the main text – they are buried in supplemental Figure S2B-E, so I think this point still requires attention. By pQTL I meant physiological QTL, not protein QTL, sorry I should have made that clear.

(x) Why was heritability computed from called eQTLs rather than from suitably partitioned genetic relationship matrices?

> The local heritability of each transcript (regardless of whether it had an eQTL) was calculated using the haplotype probabilities for the marker nearest to the transcription start site of the encoding gene. In typical situations, we would include the kinship matrix as a random effect in this model. However, here we

were using the fitted values of the model as the locally encoded transcript level, and we did not want to confound the model by including distal factors. In reviewing these methods we noticed that we mistakenly said that we used the kinship correction. We have now removed this from the text (Line 145 in the Methods with tracked changes).

This does not address my question, but it was a fairly minor point, so I don't insist.

(xi) The human validation material is quite interesting but I am not an expert on this part of the paper. It did seem overly long.

Upon second reading of this section, we agree that it could be shortened. We have reduced the length of this section to highlight the most important results (Section starts on Line 337 in manuscript with tracked changes).

Thank you

(xii) The discussion is incredibly short – there was no attempt to place the findings in a wider context.

We initially included a discussion with a "supplemental discussion" to show that we could reduce the length of the manuscript if needed. We have removed the "supplemental discussion" header to include the full-length discussion, which does discuss the work in a broader context.

Thank you

Methods:

(i) Genotyping (line 25 onwards) Why were haplotypes determined from RNAseq reads and not from the SNP genotypes?

We apologize for the confusing wording of this section. Haplotypes were determined both from GigaMUGA SNPs and by RNA-Seq. Using both methods provides redundancy as a quality control measure. There were several samples in which the two methods disagreed or had poor quality RNA-Seq data. These mice were excluded from the analysis. We have added text to the methods clarifying why both methods were used (Lines 34-36 in Methods with tracked changes.)

Thank you

(ii) It not clear whether the CC-RIX mice were kept in the same animal facility as the DO mice or were from a different experiment. Were they on the same high-fat diet as the DO? - please clarify.

> To clarify the mouse populations, we have added "CC-RIX" to one of the headers in the methods to indicate that this section describes the CC-RIX mice (Line 40 in Methods with tracked changes). The CC-RIX mice were housed at The Jackson Laboratory. The DO mice were part of a previous experiment and were housed at the University of Wisconsin. DO and CC-RIX mice were maintained on different high-fat, high-sugar diets. DO mice received a HF/HS diet (44.6% kcal fat, 34% carbohydrate, and 17.3% protein) from Envigo Teklad (catalog number TD.08811). The CC-RIX mice received a custom-designed high-fat, high-sugar (HF/HS) diet (Research Diets D19070208).
>
> These data should be included in the Methods.

(iii) It is not clear what the “processed data” (Methods line 31) refer to. Are these the CC-RIX genotypes? If so, what are the gene expression data?

> We agree that this heading is confusing. We have changed the heading to "Pre-processed DO data" and added a sentence to the underlying paragraph that these data were part of a previous publication, and we downloaded them directly from Dryad (Line 38 in Methods with tracked changes).

Thank you

(iv) TWAS analysis (methods line 250 onwards. Using just the SNP closest to the TSS for each gene might result in underestimating local genetic effects – it would have been better to have taken the most associated SNP within say 100kb. Not all cis SNPs will be associated with the expression trait, so picking one based solely on location is sub-optimal.

> We agree that this method would be problematic in human data. However, the mice used in this experiment have large haplotype blocks, and the markers within 100kb of any given marker will have identical or nearly identical genotype distributions across the animals.

I don’t agree – it is not true that nearby SNPs will always be surrogates for each other, unless they also share the same strain distribution pattern in the founders.

Reviewer #3 (Remarks on code availability):

The link to the code https://figshare.com/DOI:10.6084/m9.figshare.27066979228 does not open

We apologize for this error. We have fixed the URL in the manuscript, and it should now point to the correct page.
Thank you