nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits
Corresponding author name(s): Dr Gabriel Hoffman

Editorial notes:

Redactions – unpublished data	Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

Reviewer Comments & Decisions:

Decision Letter, initial version:

5th Mar 2021

Dear Dr Hoffman,

Your Article, "Trans-ethnic eQTL meta-analysis of human brain reveals regulatory architecture and candidate causal variants for brain-related traits" has now been seen by 2 referees. You will see from their comments copied below that while they find your work of considerable potential interest, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be very interested in considering a revised version that addresses these serious concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. As you will see from these comments, reviewer #1 suggests emphasizing how this extends the boundaries of eQTL discovery, analysis, and characterization compared to other recently published studies; reviewer #2 has substantial concerns regarding the robustness of the joint fine-mapping

results and the statistical support, which should be carefully clarified. We hope that you will find the prioritised set of referee points to be useful when revising your study.

If you choose to revise your manuscript taking into account all reviewer and editor comments, please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

You may use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Thank you for the opportunity to review your work.

Sincerely,

Wei

Wei Li, PhD
Senior Editor
Nature Genetics
1 New York Plaza, 47th Fl.
New York, NY 10004, USA
www.nature.com/ng

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Summary: Zeng et al present a large-scale eQTL resource created through meta-analysis of predominantly cortical brain tissue. The authors use this resource to fine map and colocalize GWAS risk variants for a variety of traits. The main results involve a) random effects meta-analysis of multiple eQTL results from different brain regions, termed mmQTL, with corresponding power calculations and application to GTEx b) characterizing the resulting identified primary and conditional eQTLs with regard to cellular specificity, c) fine mapping GWAS variants to show enrichment of both CNS and non-CNS traits, and d) in-depth examples around ZNF823 , THOC7, FURIN, and APH1B . While the resource is large, and the analyses well-conducted, some of the biological results seem counter-intuitive with regard to regional-specificity, and more emphasis could be placed on highlighting how extends the boundaries of eQTL analysis and interpretation.
1) This seems to really extend the sample sizes only of the frontal cortex, which was entirely ROSMAP and predominantly psychENCODE. Only GTEx really contained extensive regional diversity across the brain, by using fewer donors per region. It therefore seems that eQTLs identified by meta-analysis should either be cortically-enriched or broadly region/cell-type associated. I don't fully understand the benefit of including non-cortical tissue from GTEx with regard to power and/or eQTL discovery, since there are three regions (sACC, cortex, frontal cortex) from cortex, and excluding non-cortical regions maybe reduces by ~250 effective sample size (based on Figure 3A), making the total effective sample size ~2750 rather than 2,974.

2) Part of the enthusiasm surrounding GTEx analyses is diminished by the GTEx 2020 eQTL paper (https://science.sciencemag.org/content/369/6509/1318), including the conditional eQTL descriptions. It's unclear why this current paper used the older GTEx V7 - which contains ~30-40% fewer brain samples – rather than the final V8 release (which has been accessible since July 2019).
3) Is this mmQTL framework of meta-analysis is more powerful statistically than mega-analysis, e.g. a single eQTL model for each SNP-gene pair across all samples combined, using the same mixed effects modeling framework?
4) Are differences between previous eQTL lists from psychENCODE and here resulting from different methods for controlling the false discovery rates? It seems like the current manuscript used p < 1e-6 while psychENCODE using the FastQTL permutation procedure. Some comment should maybe by made regarding was marginal p-value that controlled their FDR in the psychENCODE paper – was it more than or less than 1e-6?
5) It seems like psychENCODE – which is already a collection of several datasets/cohorts – is the actual trans-ethnic resource, since the other two datasets added in this paper are just European. Furthermore, beyond simulations, it doesn't really seem like this approach leverages the fact that samples are trans-ethnic, since it was unclear how ancestry information as used in analyses – I didn't see any description of including genomic ancestry PCs from the SNP data, for example. I would therefore remove "trans-ethnic" from the title
6) While the cell type enrichment tests – calculating \pi_1 – do show strong enrichment, I wonder what proportion of cell type-specific eQTLs are directionality consistent (which isn't measured by that metric). Perhaps you could produce some kind of pseudo-\pi_1 where you force directionally-inconsistent eQTLs to have p-values of 1, and then recalculate \pi_1. If all are directionally consistent, you would get the same enrichment statistic, but this would get penalized when eQTLs are not consistent. Or at least you can provide some comment on directional consistency.
7) There's very little in the paper around biological effect sizes, ie log2 fold changes per allele copy, and its unclear how these meta-analysis approaches affect these estimates. Do primary, secondary, and tertiary SNPs have decreasing effect sizes? Reporting effect sizes might be especially important in the cell type enrichment, given the presumable type 2 error rate in using bulk tissue, and moreover, there's an expected effect size in bulk tissue for a cell type-specific eQTL based on the prevalence of that cell type. For example, if there's an eQTL with an effect estimate of 1 in a cell type representing 10% of cells, the effect estimate in bulk tissue should be 0.1. Excitatory neurons in bulk cortex are maybe 15-25% of cells and microglia are maybe 5% of cells – are the effect estimates for their cell type-specific eQTLs attenuated accordingly in the bulk tissue? These effect sizes also seem important for claiming potential molecular mechanisms of candidate variants in GWAS.
8) GWAS colocalization results seem to differ from Schrode 2019 (and Fromer 2016) which identified 5 candidate genes with clear single-gene SCZD colocalization. What happened to the other SCZD coloc genes like SNAP91, TSNARE1, CLCN3, and CNTN4 in the current paper, and what contributed to their decreased emphasis here?

-Andrew Jaffe (please leave my signature in)

Reviewer #2:
Remarks to the Author:
The authors have developed an approach termed multivariate multiple QTL (mmQTL) that allows for identification of trans-ethnic primary and secondary (conditionally independent) eQTLs using a linear mixed model and combination of results across datasets using a random effects meta-analysis that

models the correlation between multiple brain regions from a shared set of individuals.
They apply the mmQTL pipeline to perform a trans-ethnic eQTL meta-analysis of 3,188 human brain RNA-seq samples from 2,029 donors, including 444 non-European individuals. The included datasets are from PsychENCODE, Religious Orders Study and Memory and Aging Project (ROSMAP) and GTEx. The combined effective samples size is 2,974, achieving the largest eQTL analysis of human brain to date.

The large sample size and the novel analysis pipeline increase the analytical power and fine-mapping resolution. In particular, the study substantially increases the number of genes with detected conditional eQTLs.
Furthermore, by combining with available GWAS results (from CAUSALdb), joint fine-mapping identifies candidate causal variants shared between gene expression and GWAS traits, including a total of 301 variant-trait pairs for 23 brain-related traits driven by 189 candidate causal variants for 179 genes.

The paper is generally clear and succinct, presenting a novel analytical pipeline that generates interesting and well-founded results from analyses of a large RNA-seq sample set and integration with available GWAS fine-mapping results.

Particular strengths include the trans-ethnic approach with increased finemapping resolution, detection of secondary and tertiary conditional eQTLs enriched for cell type specific regulatory effects and identification of candidate causal variants implicating potential initial disease mechanisms in schizophrenia, bipolar disorder and Alzheimer's disease.

I have some concerns/questions regarding the claimed identification of novel disease genes and the significance/robustness of the identified candidate causal variants:
• The highlighted top-ranking genes identified in the joint fine-mapping have all been implicated previously in the disorders investigated. Which genes are novel disease genes (as claimed in the abstract) and what is their genome-wide statistical support?
• The authors detect candidate causal variants using a threshold for colocalization posterior probability (CLPP) > 0.01 (citing Hormozdiari et al., 2016). However, it is not clear what is the statistical significance and validity of results surpassing this threshold. Could e.g. the false discovery rate be estimated?

Minor comment:
• There is a typo in the titles of fig 6 and 7: "colocationation" should be colocalization.

Author Rebuttal to Initial comments

We thank the reviewers for their feedback on our manuscript. We have addressed all of their comments and provide our responses below as indented text. Changes to the manuscript are indicated by indented red text. As part of this revision we have now used GTEx v8 for our eQTL meta-analysis. The updated results are used throughout the text and figures in this revision. Figures 1 and 3-7 and basic statistics (i.e. number of eQTLs, etc) have changed slightly with the new version. These changes are highlighted in red in the new manuscript, but we don't refer to every small change in this document, unless it is of particular interest.

Reviewer #1:

Comment 1

This seems to really extend the sample sizes only of the frontal cortex, which was entirely ROSMAP and predominantly psychENCODE. Only GTEx really contained extensive regional diversity across the brain, by using fewer donors per region. It therefore seems that eQTLs identified by meta-analysis should either be cortically-enriched or broadly region/cell-type associated. I don't fully understand the benefit of including non-cortical tissue from GTEx with regard to power and/or eQTL discovery, since there are three regions (sACC, cortex, frontal cortex) from cortex, and excluding non-cortical regions may be reduced by ~250 effective sample size (based on Figure 3A), making the total effective sample size ~2750 rather than 2,974.

Response

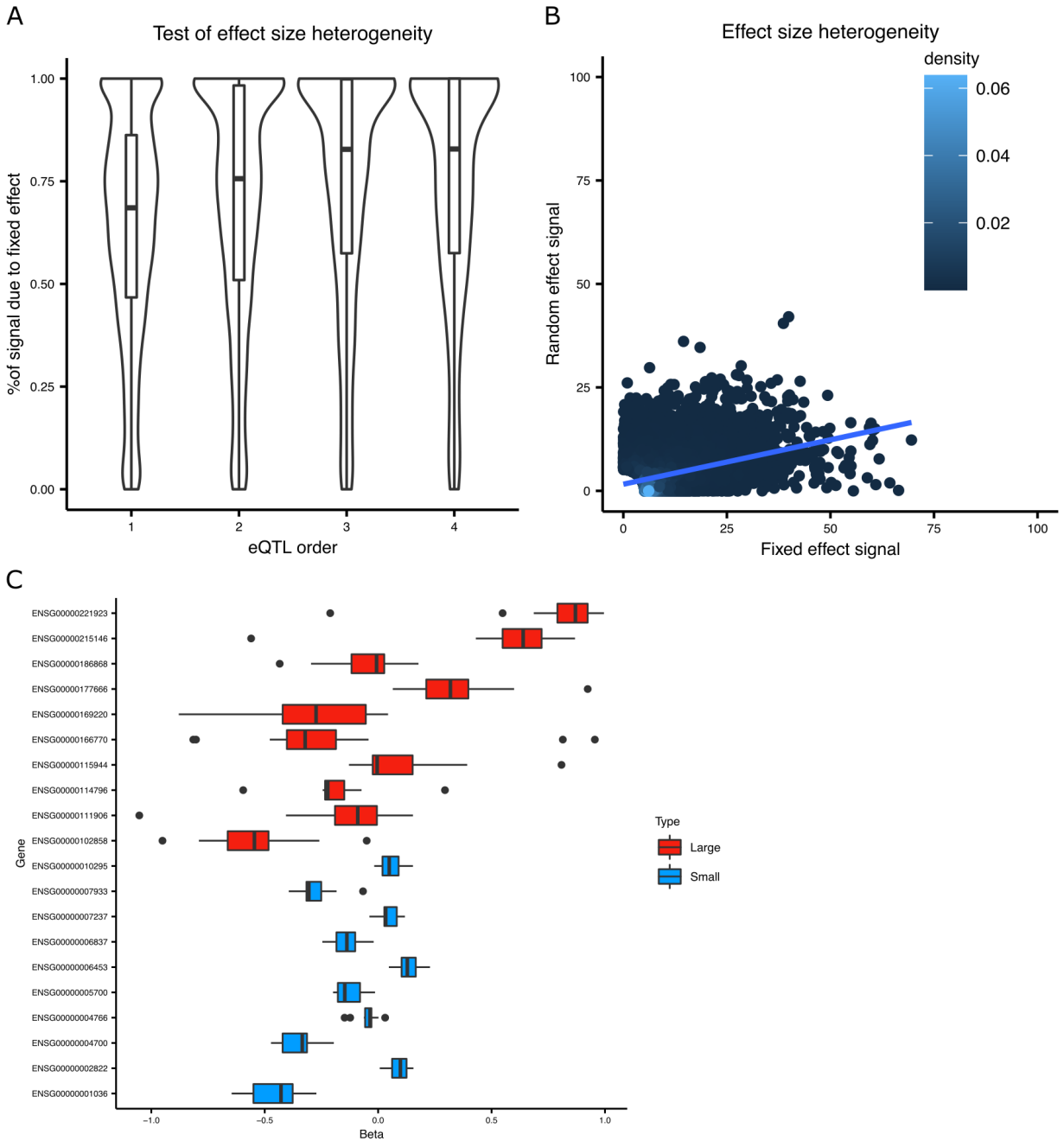
We thank the reviewer for raising this important issue. The power of our analysis is a combination of the data resource and our statistical methods. For example, we take advantage of the large sample size using a meta-analysis approach, and we leverage diverse ancestry using a linear mixed modeling approach. Using the non-cortical brain regions from GTEx increases the statistical power by increasing the effective sample size (GTEx v8 contains 2118 samples for 317 donors) and we leverage this increase by modeling the repeated measures study design. Moreover, using 13 brain regions from GTEx captures the effect size heterogeneity which we leverage using a random effects meta-analysis approach proposed by Han and Eskin (AJHG, 2011, doi: 10.1016/j.ajhg.2011.04.014).

While the standard fixed effect meta-analysis tests whether the mean effect size across all studies (or tissues, groups, etc) is non-zero, a standard random effect meta-analysis can test whether there is significant variation in effect sizes across studies. The Han-Eksin method used here combines these two tests into a single test which evaluates both the mean and variation in the estimated effect sizes across studies. As described in the original paper, the Han-Eksin test statistic can be decomposed into the sum of two statistics with one testing the mean (S_{mean}) and the other testing variance ($S_{variance}$). In order to see the importance of including multiple brain regions we empirically evaluated the contribution of S_{mean} versus $S_{variance}$ in detecting eQTLs.

Based on this feedback, we produced eQTL results from 1) the full dataset and 2) only cortical regions. eQTL summary statistics are available for both analyses at icahn.mssm.edu/brema and https://www.synapse.org/#!Synapse:syn25592266.

We have added the following to the results section:

The test statistic for random effect meta-analysis used here is composed of the sum of statistics testing the mean (S_{mean}) and variance ($S_{variance}$) of the estimated effect sizes across datasets [49]. So statistical power to detect eQTLs depends on both the effect size as well as the effect size heterogeneity across brain regions. In our analysis an average of 72.2% of power for primary eQTL analysis is attributable to effect size, while the rest is attributable to heterogeneity (**Supplementary Figure 3)**. Considering only cortical brain regions reduces effect size heterogeneity and reduces the number of genes with detected eQTLs from 10,769 to 9,431, but, more importantly, reduces the number of genes with conditional eQTLs from 5,336 to 3,533.



Supplementary Figure 3: Impact of effect size heterogeneity. The test statistic from the random effect meta-analysis used here (Han and Eskin, 2011) is the sum of statistics testing the mean (S_{mean}) and variance ($S_{variance}$) of the estimated effect sizes. **A)** The percent of total signal contributed by the fixed effect (i.e. S_{mean} / (S_{mean} + $S_{variance}$)) is shown for the lead eQTL SNP for multiple orders of conditional analysis. **B)** The relationship between the test statistics is visualized by plotting $S_{variance}$ against S_{mean} from the lead eQTL SNP from the primary eQTL analysis. **C)** The estimated effect sizes from the lead eQTL SNP for genes with high and low levels of effect size heterogeneity is shown.

Comment **2**

Part of the enthusiasm surrounding GTEx analysis is diminished by the GTEx 2020 eQTL paper (https://science.sciencemag.org/content/369/6509/1318<https://urldefense.proofpoint.com/v2/url?u=https-3A__science.sciencemag.org_content_369_6509_1318&d=DwMGAg&c=shNJtf5dKgNcPZ6Yh64b-A&r=KdYcmw5SdXylMrTGSuNVkNJulowod64k0PTDC5BHZkk&m=SgD9Aenj0ma6Z5tPuAoM4GH7tyJR7j0_JCJI85OyLu4&s=lJuzTmqbgDXvw7ipslaC1qP6j5tMTkcAC-AMw1usbts&e=>), including the conditional eQTL descriptions. It's unclear why this current paper used the older GTEx V7 - which contains ~30-40% fewer brain samples – rather than the final V8 release (which has been accessible since July 2019).

Response

We thank the reviewer for raising this issue. Our analysis evolved over the last 2 years, and GTEx v7 had been current during our last analysis freeze. We have now included GTEx v8 in our analysis and updated all figures and basic summary statistics (i.e. number of eQTLs) throughout the manuscript.

Comment 3

Is this mmQTL framework of meta-analysis is more powerful statistically than mega-analysis, e.g. a single eQTL model for each SNP-gene pair across all samples combined, using the same mixed effects modeling framework?

Response

Using a standard linear model, meta-analysis retains the same statistical power and control of false positive rate as mega-analysis, while being much simpler to implement for large datasets (Willer, et al, 2010 Bioinformatics, doi: 10.1093/bioinformatics/btq340; Pasaniuc and Price, Nature Reviews Genetics, 2017, doi:10.1038/nrg.2016.142; Lin and Zeng, 2010, Genetic Epidemiology, doi:10.1002/gepi.20435). For this reason, meta-analysis of summary statistics is standard in the GWAS field (Pasaniuc and Price, 2017).

We have added the following text in the Methods:

Statistically, the standard fixed effect mega-analysis combines all data into a single regression model and assumes a fixed effect size across all studies as well as constant error variance across all studies. These assumptions are not satisfied in multi-tissue eQTL analyses due to variation in effect size and variation in error variance across tissues (Sul et al. 2013). Using a random effect meta-analysis addresses both of these issues to retain control of false positive rate while leveraging the effect size heterogeneity to increase power.

Implementing the current analysis model as a mega-analysis would require 1) a random effect over the regression coefficients, 2) a random effect to account for population structure and repeated measures, and 3) heteroskedastic errors to allow study-specific error variances. This approach is impractical to implement, very computationally

depending, and offers little advantage over the meta-analysis framework we have used here.

For these reasons, mmQTL uses a random effect meta-analysis to aggregate results across studies.

Comment 4

Are differences between previous eQTL lists from psychENCODE and here resulting from different methods for controlling the false discovery rates? It seems like the current manuscript used p < 1e-6 while psychENCODE using the FastQTL permutation procedure. Some comment should maybe by made regarding was marginal p-value that controlled their FDR in the psychENCODE paper – was it more than or less than 1e-6?

Response

We thank the reviewer for pointing out the different p-value cutoffs used.

The PsychENCODE analysis in Wang, et al. (2019, Science) uses a FDR 5% cutoff which empirically corresponds to p<8.3e-4 in their analysis. Since Wang, et al performed a mega-analysis of the PsychENCODE data, they used a permutation method implemented in FastQTL in order to compute the FDR and p-value cutoff.

In contrast, using a more conservative cutoff p < 1e-6 here was motivated by several reasons. Applying the permutation approach to our analysis is challenging. The permutation method in FastQTL used by PsychENCODE is very fast for linear models. Yet it is more computationally demanding and challenging to implement for linear mixed models (Joo, et al, 2016 Genome Biology, doi: 10.1186/s13059-016-0903-6). Moreover, use of the random effects meta-analysis after fitting linear mixed models adds another layer of complexity. Developing a permutation approach for this analysis would have been very computationally expensive and was beyond the scope of the current work.

Studies often use more liberal multiple testing cutoffs because of the limited statistical power. Here we have substantially more power than previous resources, and our focus on statistically fine-mapping motivated the use of a strict p-value cutoff. Previous work has find that statistical fine-mapping can perform poorly on genes that only pass a liberal cutoff (Hormozdiari et al., 2016, 2018, doi: 10.1016/j.ajhg.2016.10.003 and 10.1038/s41588-018-0148-2). Since the focus of this work was statistical fine-mapping to identify candidate causal variants, we were concerned that a more liberal p-value cutoff could lead to false positive findings.

We have added the follow text to the Methods:

We note that Wang, et al. (Wang et al. 2018) performed a mega-analysis of the PsychENCODE data and used a permutation method in order to compute the FDR and p-value cutoff. Their FDR 5% cutoff empirically corresponds to p<8.3e-4 in their

analysis. Yet the complexity of applying a permutation approach to linear mixed models (Joo et al. 2016), and the use of a random effects meta-analysis afterwards in this analysis made a computationally efficient permutation approach impractical here.

Comment 5

It seems like psychENCODE – which is already a collection of several datasets/cohorts – is the actual trans-ethnic resource, since the other two datasets added in this paper are just European. Furthermore, beyond simulations, it doesn't really seem like this approach leverages the fact that samples are trans-ethnic, since it was unclear how ancestry information as used in analyses – I didn't see any description of including genomic ancestry PCs from the SNP data, for example. I would therefore remove "trans-ethnic" from the title

Response

We thank the reviewer for allowing us to clarify this issue.

The PsychENCODE resource includes 444 non-European donors and current GTEx v8 includes 30 non-Europeans. Linear mixed models have a long history of being applied to datasets of diverse ancestries in order to retain power while accounting for population structure (Kang, et al. 2010, Nature Genetics, doi: 10.1038/ng.548; Zhou and Stephens, 2012, Nature Genetics; 10.1038/ng.2310; Sul et al, 2018, PloS Genetics, doi:10.1371/journal.pgen.1007309). While genotype PCs can be included to account for some degree of population structure, linear mixed models are more effective for datasets with more genetic diversity (Sul, et al., 2018; Yang, et al. 2014, Nature Genetics, doi: 10.1038/ng.2876).

Our analysis comprised two steps: 1) We applied linear mixed models in analysis within each dataset in order to account for this population structure, and then 2) combined the summary statistics across datasets using a random effects meta-analysis. This first step is where the diverse ancestry is considered.

We have clarified the text in the Results:

We accounted for diverse ancestry (**Figure 1B**) by applying a linear mixed model to the full data within each resource, and then combined summary statistics from these 15 eQTL analyses using a random effects meta-analysis to account for effect size heterogeneity and donor overlap in between brain regions in GTEx (**Figure 1C**).

Comment 6

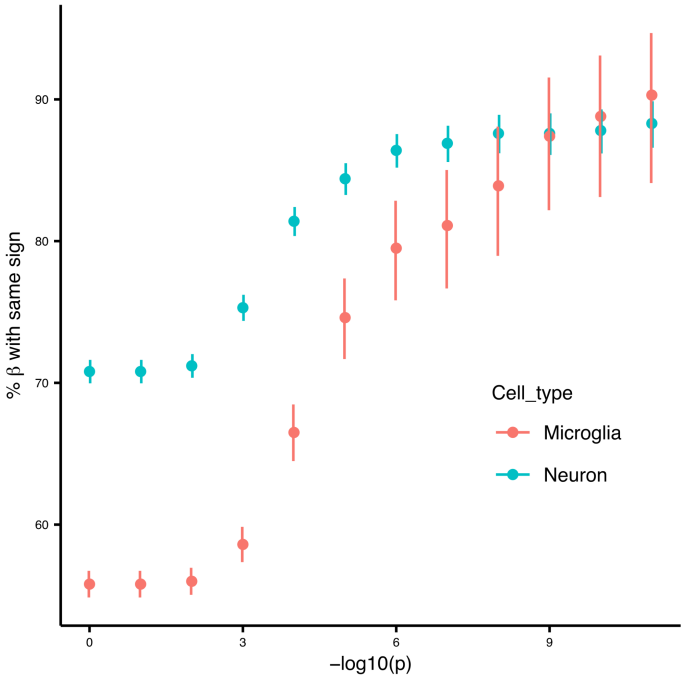
While the cell type enrichment tests – calculating \pi_1 – do show strong enrichment, I wonder what proportion of cell type-specific eQTLs are directionality consistent (which isn't measured by that metric). Perhaps you could produce some kind of pseudo-\pi_1 where you force directionally-inconsistent eQTLs to have p-values of 1, and then recalculate \pi_1. If all are directionally consistent, you would get the same enrichment statistic, but this would get

penalized when eQTLs are not consistent. Or at least you can provide some comment on directional consistency.

Response

We thank the reviewer for raising this issue that we did not address in the previous version. While Storey's π_1 is a metric of replication across datasets, it does not incorporate the sign of the effect. In order to address this, we now measure sign concordance directly across a range of p-value cutoffs.

Moreover, the concordance in the sign of the estimated effect sizes between our meta-analysis and the cell-type specific analyses increased with stricter p-value cutoffs (**Supplementary Figure 2**).



Supplementary Figure 2. Lead eQTL SNP sign concordance. For the lead eQTL SNP of each gene in the meta-analysis, the sign of the mean estimated effect size is compared to the estimated effect sign from neuron and microglia eQTL analyses. The concordance rate increases with the strictness of the p-value cutoff, so a smaller p-value indicates a higher concordance rate. Error bars indicate 95% confidence interval for a binomial proportion.

Comment 7

There's very little in the paper around biological effect sizes, ie log2 fold changes per allele copy, and its unclear how these meta-analysis approaches affect these estimates. Do primary, secondary, and tertiary SNPs have decreasing effect sizes? Reporting effect sizes might be especially important in the cell type enrichment, given the presumable type 2 error rate in using bulk tissue, and moreover, there's an expected effect size in bulk tissue for a cell type-specific eQTL based on the prevalence of that cell type. For example, if there's an eQTL with an effect estimate of 1 in a cell type representing 10% of cells, the effect estimate in bulk tissue should be 0.1. Excitatory neurons in bulk cortex are maybe 15-25% of cells and microglia are maybe 5% of cells – are the effect estimates for their cell type-specific eQTLs attenuated accordingly in the bulk tissue? These effect sizes also seem important for claiming potential molecular mechanisms of candidate variants in GWAS.

Response

We thank the reviewer for raising this important issue. Statistical power to detect an eQTL increases with effect size (beta) and the minor allele frequency (p) according to beta^2(1-p)p. Therefore high power increases with higher effect size and higher MAF. In conditional eQTL analysis, primary eQTLs have higher beta^2(1-p)p than higher order eQTLs, so that the trend is not driven simply by effect size. Analysis of estimated effect size and MAF for the lead eQTL SNP of significant genes in fact shows a similar distribution of effect sizes for increasing conditional analyses, but the MAF shows a marked decrease.

In addition, the reviewer raises a very interesting question about the interpretation of eQTL effect sizes estimated from bulk tissue versus purified cell types. The reviewer's biological intuition, that an eQTL SNP with a large effect size in a rare cell type will be attenuated in the bulk data to produce a smaller effect size, is consistent with our intuition about the underlying biology.

We address these comments in the Results, Methods and new Supplementary Figures:

We have added the following text in Results:

While the distribution of estimated effect sizes is similar for increasing conditional eQTL degree, the minor allele frequency decreases markedly (**Supplementary Figure 5**). Interpretation of the estimated effect sizes from bulk and cell type specific data is challenging and is affected by multiple factors (**Supplementary Figure 6, see Methods**).

We have added the following text in Methods:

Interpretation of estimated effect sizes

In the scenario where a SNP has a large cell type specific effect on gene expression, the true biological effect will be attenuated in bulk data that is composed of multiple cell

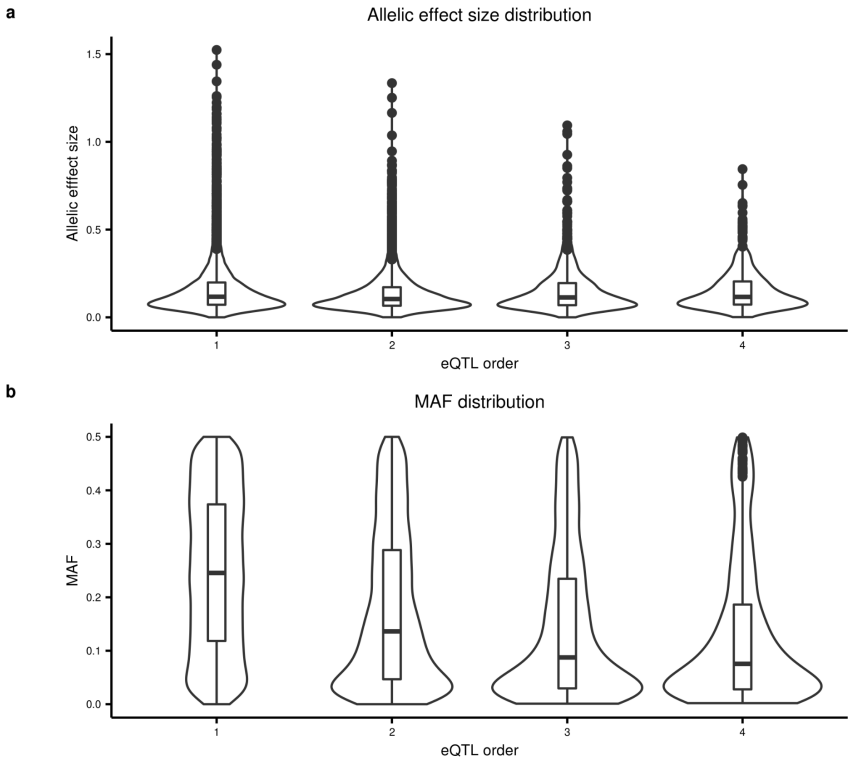
types. Yet testing this biological intuition through eQTL analysis is challenging for a number of technical reasons. Unfortunately, eQTL analysis does not directly estimate the biological effect size because the gene expression is typically $\log_2$ transformed, scaled to have variance 1, and often quantile normalized. In addition, the inclusion of PEER factors, or other covariates can account for cell type heterogeneity across samples in the data. Therefore, the estimated eQTL effect size reflects the association between SNP and (transformed) gene expression after accounting for other variables.

Furthermore, the technical process of obtaining gene expression from bulk tissue versus cell type specific samples is susceptible to different noise profiles based on differing protocols and the biological condition of the physical samples. In fact Young, et al. [26] found that cell-type specific samples from microglia are noisier than those from bulk samples, and that reported estimated effect size in purified cells are attenuated.

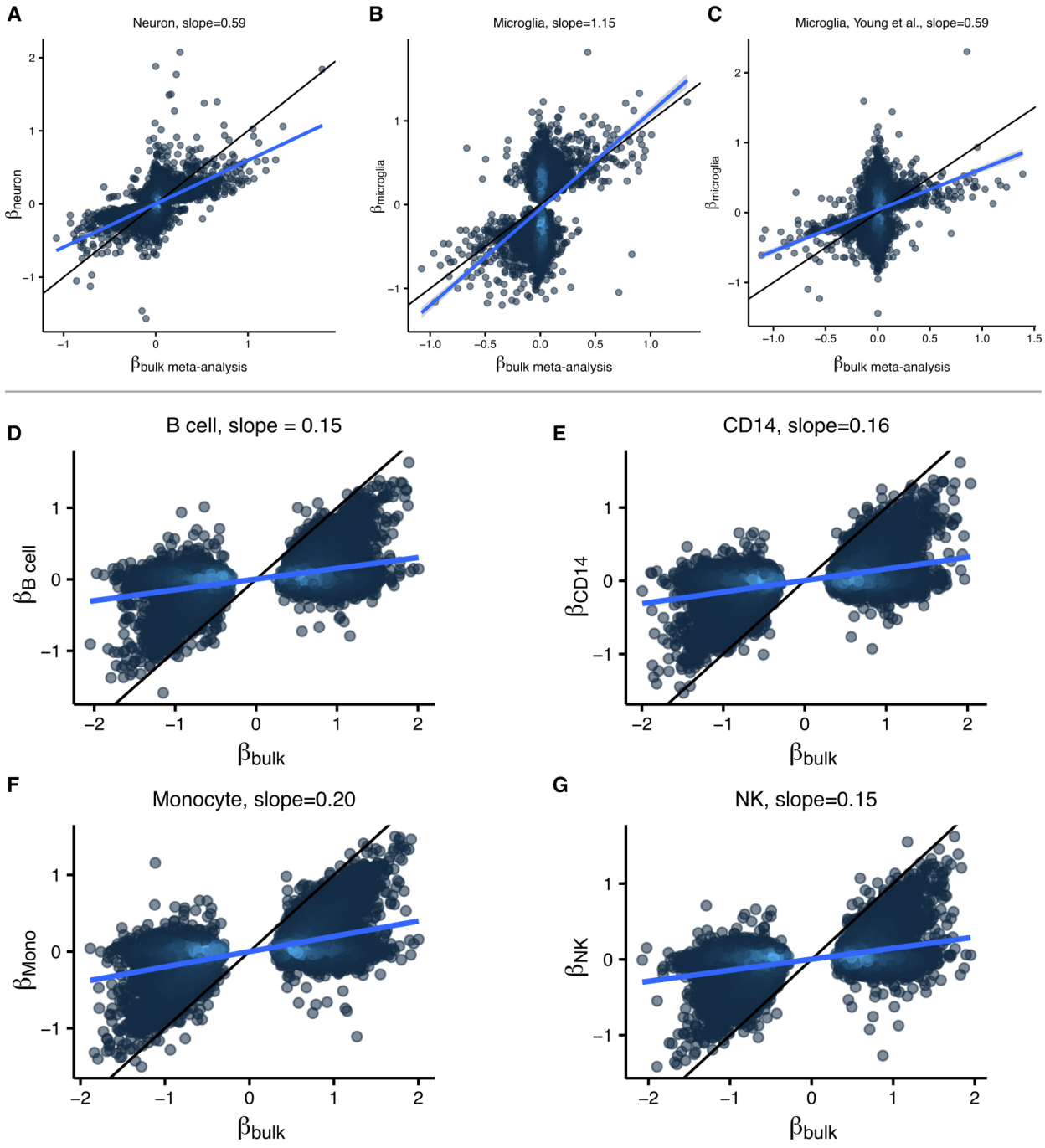
We performed an empirical analysis of the estimated allelic effect sizes from the lead eQTL variants for each gene in our meta-analysis of bulk data compared to estimates from cell type specific data from neurons 27 and microglia from [REDACTED] and Young, et al. 26. The comparison between bulk and neuron-enriched data gives a slope of 0.59, indicating that the slope is on average actually smaller in the cell type data. Comparison to the [REDACTED] and to the Young microglia data gives a slope 0.59. These results are difficult to interpret, especially given the caveats above.

These results are not unexpected given the statistical and technical challenges outlined above. In fact, these findings are not unique to our data. Recently, Ota, et al [71] generated eQTLs four immune cell types and compared the estimated effect sizes to bulk immune data from Ishigaki et al, [72]. Our analysis recapitulates their finding that effect size estimates are actually smaller in the cell type specific data.

Rigorous analysis of effect size estimates is challenging both statistically and due to different noise profiles of bulk and cell type data. Mohammadi et al. [73] developed a method to estimate a more biologically interpretable allelic effect size. Further research on this challenge in the field could yield further insight into cell type specific gene regulation.



Supplementary Figure 5. Estimated effect size and minor allele frequencies from conditional eQTL analysis. The estimated effect size **(A)** and MAF **(B)** are shown for the lead eQTL SNP of significant genes for increasing order to conditional eQTL analysis. **A)** The distribution of estimated effect size is similar for all conditional analyses. **B)** The MAF shows a marked decrease with increasing order of conditional analysis.



Supplementary Figure 6. Comparison of estimated effect size for bulk and cell-type specific data. (A-C) Estimated allelic effect size for eQTL lead in (A) neurons (Jaffe, et etl. 2020), [REDACTED] (C) microglia from Young, et al.(2021) compared to effect size estimates from meta-analysis of bulk data from the current study.(D-G) Estimated allelic effect size for eQTL lead SNP in four immune cell types including (D) B cells, (E) CD14, (F) monocytes, (G) NK cells from Ota, et al. (2021) compared to estimates from bulk samples (Ishigaki, et al. 2017).

Comment 8

GWAS colocalization results seem to differ from Schrode 2019 (and Fromer 2016) which identified 5 candidate genes with clear single-gene SCZD colocalization. What happened to the other SCZD coloc genes like SNAP91, TSNARE1, CLCN3, and CNTN4 in the current paper, and what contributed to their decreased emphasis here?

Response

We thank the reviewer for raising this issue.

There are two general classes of statistical methods to perform eQTL-GWAS integration (see response directly below to Reviewer #2, Comment #1). The first set termed 'gene-level' methods (i.e. TWAS, prediXcan, coloc) focus on identifying *genes* that mediate disease risk. The second approach, termed 'variant-level' methods, uses joint statistical fine-mapping of eQTL and GWAS in order to identify candidate causal variants that drive variation in both gene expression and disease risk. We use the joint fine-mapping approach here in order to better characterize the mechanistic link between DNA sequence and disease risk. Moreover, our data resource and statistical methods were uniquely able to increase the power of variant-level analysis by leveraging the large sample size and the diverse ancestry to break up linkage equilibrium.

The differences in these two complementary approaches mean that the genes discovered by `gene-level` analyses are often a superset of genes identified by `variant-level`. This is due to the fact that identification by a `variant-level` requires 1) a gene to be implicated in disease and 2) low enough linkage disequilibrium so that the set of candidate causal variants is small. This stricter criteria means that `variant-level` analysis often produces fewer findings, but gives a high resolution perspective on disease mechanisms.

The four genes mentioned in the comment were originally proposed as schizophrenia risk genes by Fromer (2016, Nature Neuroscience). This work by the CommonMind Consortium was one of the first large-scale analyses of eQTLs from post mortem brains, and the data in this paper has been widely used, including by this work and PsychENCODE. Since that time, statistical methods for processing RNA-seq data, and performing eQTL analysis has improved substantially. The TSNARE1 finding has not been replicated in subsequent analyses of these data (Huckins, 2019, Nature Genetics; Dobbyn, 2018, AJHG).

We have now included gene-level colocalization analysis of our results with the GWAS for schizophrenia using coloc (Giambartolomei, 2014, PLoS Genetics). While linkage disequilibrium prevented our fine-mapping analysis from identifying a candidate causal variants for SNAP91, CLCN3 or CNTN4, these were identified as contributing to schizophrenia risk through gene-level analysis.

See gene-level coloc results at https://www.synapse.org/#!Synapse:syn25871674

Reviewer #2:

Comment 1

The highlighted top-ranking genes identified in the joint fine-mapping have all been implicated previously in the disorders investigated. Which genes are novel disease genes (as claimed in the abstract) and what is their genome-wide statistical support?

Response

We thank the reviewer for raising this important issue. With the widespread interest in eQTL-GWAS integration over the last few years, two different families of methods have emerged. First, and most widely used, are methods that perform integration at the gene-level. These includes methods to test if the estimated genetic component of gene expression is associated with the GWAS trait (i.e. transcription-wide association studies like FUSION, PrediXcan, SMR), or align p-values from eQTL and GWAS to identify concordance (i.e. coloc, gwas-pw). These methods are useful for identifying specific genes that mediate disease risk, and have been widely used.

The other focus of eQTL-GWAS integration is at the variant-level in order to identify candidate causal variants driving changes in gene expression which then affect the trait of interest. Identifying a candidate causal variant also implicates a specific gene and a mechanistic link from DNA sequence to high-level phenotype. This approach is implemented using statistical fine-mapping of both the gene expression traits and GWAS trait to produce colocalization posterior probabilities (CLPP) which indicate the probability that a variant drives variation in both gene expression and the GWAS trait (Hormozdiari, 2016, AJHG doi: 10.1016/j.ajhg.2016.10.003). These variants can then be validated experimentally using a multiplexed reporter assay that can detect differences in expression driven by changes in allele. These variants can also be followed up in cell culture to identify changes in molecular or other low-level phenotypes driven by changing the allele using CRISPR-mediated allele editing (for example see Schrode, et al, 2019, Nature Genetics, doi: 10.1038/s41588-019-0497-5).

These two goals of eQTL-GWAS integration are complementary, and we focus here on joint statistical fine-mapping to identify candidate causal variants. The data resource and analysis used here is particularly suited for improving fine-mapping due to 1) the substantial increase in sample size compared to previous resources, and 2) the fact that including individuals of non-European ancestry breaks up linkage disequilibrium to reduce the size of candidate causal sets.

The goal of this work is not ‘gene discovery’ *per se*, but rather ‘variant discovery’. Our analysis highlights specific 4 genes, but the emphasis is on identifying candidate causal variants driving their expression:

- THOC7 was previously identified to affect schizophrenia risk (Huckins, et al. 2019, Nature Genetics, doi: 10.1038/s41588-019-0364-4), but we identified a candidate causal variant predicted to drive expression and disease risk.

- A candidate causal variant for FURIN was previously identified and validated experimentally (Schrode, et al. 2019, Nature Genetics), but we identify this variant and gene in other behavioral or psychiatric traits.

- APH1B was previously implicated in Alzheimer's Disease (Jansen, et al, 2019, Nature Genetics, doi: 10.1038/s41588-018-0311-9), but the lead GWAS variant is non-synonymous and Zhang, et al. (2020, HMG, doi: 10.1093/hmg/ddaa017) performed experimental work under the assumption that the disease risk was mediated by a change in the protein sequence. Here we link this candidate casual variant to expression of the APH1B gene and propose that regulation, rather than protein sequence, mediates the disease risk.

- ZNF823 had previously implicated in schizophrenia (Pardiñas, et al. 2018, Nature Genetics, doi: 10.1038/s41588-018-0059-2) but our analysis identifies a candidate causal variant that disrupts a REST binding site.

The contribution of this work is to better understand disease risk at the variant level in order to facilitate design of high- and low-throughput followup experiments.

Comment 2

The authors detect candidate causal variants using a threshold for colocalization posterior probability (CLPP) > 0.01 (citing Hormozdiari et al., 2016). However, it is not clear what is the statistical significance and validity of results surpassing this threshold. Could e.g. the false discovery rate be estimated?

Response

We thank the reviewer for raising this important issue. First, the colocalization posterior probability (CLPP) of a given SNP is the product of the fine-mapping posterior inclusion probability (PIP) for a given gene and GWAS trait. Formally, CLPP = PIP_{gene} * PIP_{trait}. Since each PIP is estimated under a formal statistical model (Hormozdiari, et al, 2016, AJHG), CLIPP is itself a posterior probability of both events (i.e. the SNP influencing variation in the expression and GWAS trait) and is directly interpretable as a posterior probability, assuming the events are independent. Since CLPP is formally defined, and is not simply an arbitrarily defined score, using 1% is a natural cutoff.

Second, this 1% score was proposed in the original eCAVIAR paper (Hormozdiari, et al, 2016, AJHG) and has been widely adopted. For example see:

Klarin et al., 2019, Nature Genetics, doi:10.1038/s41588-019-0519-3
Bonder et al., 2021, Nature Genetics, doi:10.1038/s41588-021-00800-7
Ota et al., 2021, Cell, doi: 10.1016/j.cell.2021.03.056

Lastly, we performed two permutation analyses on our real data to test the empirical performance of the 1% CLPP cutoff. First we considered a scenario where there are no true eQTLs. Using expression of 2,457 genes on chr1 from the PsychENCODE data, we randomly permuted the sample labels and then performed eQTL analysis followed by eCAVIAR colocalization analysis with PGC2 Schizophrenia GWAS summary statistics. No SNP had a CLPP > 1%. Next we considered the scenario where there are true eQTLs in the dataset, but their genome locations are independent of the GWAS associations. We extracted eQTL results for the 920 genes on chr1 that have a lead eQTL variant with p < 1e-6, and performed statistical fine-mapping on each gene. The posterior inclusion probabilities were then shuffled and used for colocalization analysis with eCAVIAR using the same Schizophrenia GWAS data. Only a single SNP passed the 1% CLPP cutoff in this analysis.

These three motivations support our use of a 1% CLPP cutoff.

Minor comment:

- There is a typo in the titles of fig 6 and 7: “colocationation” should be colocalization.

Response

We thank the reviewer for pointing out this typo.

Decision Letter, first revision:

Our ref: NG-A56777R

18th Aug 2021

Dear Dr. Hoffman,

Thank you for submitting your revised manuscript "Trans-ethnic eQTL meta-analysis of human brain reveals regulatory architecture and candidate causal variants for brain-related traits" (NG-A56777R). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Wei

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng

Reviewer #1 (Remarks to the Author):

The authors have addressed my concerns

Reviewer #2 (Remarks to the Author):

The authors have responded satisfactorily to my comments but I do not believe they have changed

the manuscript accordingly.

Re comment 1.
In the response, the authors state that "The goal of this work is not 'gene discovery' per se, but rather 'variant discovery'", and provide details for candidate causal variants impacting four highlighted genes (which have been identified as risk genes/loci in previous studies). This is all fine but I do not think it supports the claim in the abstract: "This integrative analysis identifies novel disease genes…". It would seem more appropriate to state something like "This integrative analysis identifies candidate causal variants…"

Re comment 2.
The response addresses my concern, particularly by adding permutation analyses, and I think it would be informative to include these permutation analyses in the supplementary material.

My comments above are minor and can easily be addressed. Nice paper!

Final Decision Letter:

In reply please quote: NG-A56777R1 Hoffman

17th Nov 2021

Dear Dr. Hoffman,

I am delighted to say that your manuscript "Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits" has been accepted for publication in an upcoming issue of Nature Genetics.

Prior to setting your manuscript, we may make minor changes to enhance the lucidity of the text and with reference to our house style. We therefore ask that you examine the proofs most carefully to ensure that we have not inadvertently altered the sense of your text in any way.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A56777R1) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that Nature Genetics is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

<B>Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Research offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A56777R1). Further information can be found at https://www.nature.com/nprot/.

Sincerely,

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng