nature portfolio

Peer Review File

Analysis of Biased Allelic Enhancer Activity of Schizophrenia-Linked Common Variants

Corresponding Author: Dr Zhiqiang Li

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Analysis of Biased Allelic Enhancer Activity of Schizophrenia-Linked Common Variants" by Dr Li and colleagues

The authors of this study investigated 351 common variants associated with schizophrenia (SCZ) to identify those with biased allelic enhancer activity using STARR-seq. They synthesized oligonucleotides containing reference and alternative alleles of these variants, cloned them into a STARR-seq vector, and identified SNP alleles that differentially affected enhancer activity. SNPs with biased allelic enhancer effects were further validated using eQTL, Hi-C data, and behavioral assessment of zebrafish with gene knockouts.

The study is well-designed, and the manuscript is well-written and easy to follow. The results are convincing, and the methods are sound. However, there are several areas where the manuscript can be improved:

1. Comparison with previous studies: Several previous studies have used similar high-throughput screening assays to assess the functional impact of variants on enhancers in disease contexts. The authors should compare their results with these previous findings and perform a meta-analysis if possible (ref: https://doi.org/10.1126/science.adh4265, https://doi.org/10.1016/j.xgen.2023.100404).

2. Emphasis on PCCB as a target gene: The manuscript emphasizes the discovery of PCCB as a target gene candidate associated with SCZ risk, but this gene has previously been reported to be linked with multiple neuropsychiatric traits (ref: https://www.nature.com/articles/s41467-023-40861-2, https://doi.org/10.1186/s12920-022-01202-2, https://doi.org/10.1016/j.ajhg.2022.07.011). Authors should, at minimum, cite prior studies and elucidate some of the connections of this work to previous ones.

3. Limitations of using immortalized cell lines: The use of immortalized cell lines may not accurately capture the gene regulation program and disease biology. The authors should consider validating candidate genes in organoids or primary cell cultures.

4. Clarity of Figure 2d: The description of Figure 2d does not match the figure. The specificity and effect size of SNPs can be better visualized. The authors should consider showing a heatmap using both color and size to represent effect size and statistical significance for all SNPs that are significant ($|FC| > 1.5$ & $FDR < 0.05$) in at least one of the cell types. Asterisks can be used to highlight SNPs prioritized as baaSNPs.

Reviewer #2

(Remarks to the Author)
Summary:

The authors aim to bridge the gap between genetically associated and functional roles for variants in SCZ using a STARR-seq screening method. Additionally, they characterized a potential causal target genes in zebrafish models. From their screen, they identified 351 functional SNPs linked to 217 target genes through chromatin interaction assays and eQTL analysis. They also show that a KO zebrafish model, which is linked to one of the identified SNPs, exhibits behavioral

differences concluding that some of the variants identified in their screen are causative of SCZ. It seems that the authors were very thoughtful in their approach to designing the MPRA and how they chose candidate SNPs. Additionally, the use of multiple cell lines enabled them to disambiguate cell type specific effects from more generalized effects. However, my main concern is that the authors continually state they have identified causal SNPs linked to SCZ, yet they have no experiments that directly validate these claims. On the other hand, they may have identified a functionally important gene (PCCB) in SCZ, and much of the data is strong and well executed in the Z fish and MPRA, so there is some good work here to get out into the world. Them may need to dampen some of their claims, which were:

1. They assessed regulatory activity of the variants across 4 cell lines, and identify 46 SNPs with allelic effects (baaSNPs).
2. Bioinformatics of existing epigenetic data linked these to 217 genes.
3. GO links these genes to synapses and GABA.
4. They focus on rs13072690 for a bit of follow up, which had a consistent impact on luciferase, and robustly altered behavior in zebrafish when knocked out as a het.
5. The study provides a comprehensive approach for identifying causative variants

Review:

Overall, this is a useful paper. It makes two main contributions with new data (claims 1 and 4) above. The experiments for those are generally well executed and analyzed in field standard ways, though I had a few minor questions. I think claim 3 could be supported, but they need to rerun the GO analysis with the appropriate background set to be sure. I think claim 5 is overstated. While this is useful work and should be published, claims this is comprehensive are a bit much (other instances noted in minor comments below): MPRAs/STARRseq generally are useful at determining which sequences can have a functional effect and are thus important complements to epigenetic and high-C approaches, which provide biochemical evidence a sequence might have a role, but don't directly measure functional outputs (i.e., transcription). The authors should reduce the language around their claim 5 or remove it entirely.

The other major concerns to the paper are:

1) MPRAs/STARRseq are very cell type dependent, and while these were convenient choices for transfectable cell lines, none of these cell lines really match the neurons (and perhaps microglia) implicated in SCZ in terms of their transcription factor expression, etc. Thus, the lack of alignment with brain eQTL data, etc. they see is fairly unsurprising. I am not sure there is really any way to correct this short of repeating the whole set of experiments in neurons instead. And that feels like too high a bar for a revision. Rather, that would be a whole new paper – and would require porting their library over to some other plasmid system that enables transduction of neurons (e.g. lenti). If they want to publish the current work here (at the editors discretion, of course), they would need to instead really carefully frame what they can interpret from these cells. And they should definitely reduce the claims about the comprehensiveness of the pipeline, establishing causality, etc. They are adding functional evidence for these SNPs, but this work alone is not definitive for establishing causality (see also specific instances of lines to change in minor notes).

2) I do like that they chose to follow up on the variant that seemed to be the most consistent across cell types, and in the luciferase assay as well. It suggests this one might not depend on cell type as much, and gets around my concern in #2 above a bit. However, the one puzzle is that the direction of effect appears to be opposite between Luci/STARRseq on one hand, and eQTL data (and Z-fish follow up) on the other. The authors describe this, but don't directly discuss this. How should the reader think about this difference? Is the STARR-seq in cell lines not predictive, and thus the STARR-seq hitting on this allele a lucky accident? Or would the direction of effect be different in a neuron (or key glia, given figure 5) than in these cells? Repeating the luciferase assay in such cells might be a doable experiment, as some kinds of glia are culturable from human and/or mouse tissues (though it is unclear from their hi-c data which glia type one should consider. If it is not clear from there, the available scRNAseq of the expression of PCCB might help guide a choice).

Minor:

STARRseq designs conflate posttranscriptional effects and transcriptional effects (https://pmc.ncbi.nlm.nih.gov/articles/PMC7727316/) . This is not yet well recognized, but they should discuss this caveat/limitations.

Line 33: Maybe remove the word 'relevant' here as none of these cell lines get schizophrenia. Similar around line 99, and a few other places in the text.

Line 46: This claim is a bit too strong. A truly comprehensive approach would involve CRISPR editing of the SNP in the genome in a brain relevant cell type to confirm the change in target gene expression. That would be beyond the scope of this paper, so I recommend the authors adjust the language a bit here. Something like 'our study presents an approach for functionally annotating putative causative variants'

Related to the major point above, the word causal is used repeatedly when it is sometimes inappropriate (e.g. line 92 – multiomics data can identify risk variants not necessarily causal). This is especially true in the first sentence of the last paragraph of the discussion (lines 414-416), which to me, is very bold statement given the data because they didn't actually show phenotypic data associated with any SNPs.

Lines 503-505 say that statistical tests were used for filtering, but it might be nice to see this data as a supplementary figure (e.g. read count cut-off/distribution of reads and correlation/PCA for counts data) especially since the analysis techniques are not canonical methods.

To me, figure 2C would be way more interpretable by another volcano plot, which provides more data than this circle plot, since I am more interested in effect size and p-values than genomic coordinates.

Fig 1: Fonts on some figure panels are too small to read without 200% zoom (especially 1C, D). Try and use 6 pt minimum. Check other figures as well.

Fig 2D: This is rather small to read. Perhaps it is better as a long strip across the bottom of the figure, instead of a circle.

Methods: the MPRA library had no controls (positive enhancers, repressors, shuffled sequence, minimal promoter only…)

listed. Were there any controls built in? How did they perform?
How did they move from raw sequences to count level data? This part of the methods was quite unclear. Relatedly, for sequencing library construction, were the entire regulatory elements amplified and sequenced with full coverage, and what alignment tool was used (methods line 507)?

There appears to be no code sharing plan, link to git repository, etc..
The figure legend for 3F says rs13072690 despite the text and figure legend saying rs3088186.

Line 457 – minor note: InFusion products are not ligation products. It does not involve ligase.

Line 523: Were they cloned into the 3'UTR (in the STARR seq style) or upstream of a minimal promoter for the luciferase assay? Maybe clarify this in the methods. I think it is STARRseq
Line 543: What gene set did they set as background (a.k.a reference set) for WebGestalt? An appropriate choice might be all genes with eQTLs in their brain dataset. This controls for the general brain-bias that would be expected to be present in their brain eQTL derived list, and that could thus produce spurious enrichments of categories like 'synapse' and 'GABA' by chance alone.
Line 138: please provide a reference for this vector.
The authors clone each SNP into three positions. How do they deal with this data? Is each location of the same SNP treated independently in their model (i.e., each SNP is tested three times)? Does position matter? Are there SNPs that have the same effect regardless of position?
Line 185: the phrasing here should be tweaked in a minor way. The genes came from HEK-293T STARRseq data, rather than the cells themselves.
Figure 3A: It is a bit unclear what the p-value in the figure is here. Is this p-values from the eQTL study? Or are they plotting their own p-values?
One weakness to Hi-C data is that it will have false negatives based on some aspects of the library prep (i.e., loop size depends on how the library was made). The authors should mention this in the discussion. It may also explain why there is little overlap with eQTLs. Though of course, eQTLs also are ambiguous in linking SNPs to genes, because many SNPs nearby the functional SNP will be in LD.
Figure 3E: I don't see the difference here the authors are pointing out. Further, such a claim would need statistical support. Overall, it may be simpler to simply remove the panel and the related sentence: I am not sure this data is central to their claims. But, perhaps I am missing something? Also, line 204 and legend for 3E seem to disagree with each other on direction of effect?
Line 207: Was this the only variant they tested? Or the only one that met their prediction? It seems pretty strange they would pick to focus on this one.
Line 266: which kind of glia cells is the Hi-C from?
Figure 6E/F: the Y-axis label is the same. Should one of these be different?
Sample size in fish studies was good, and I like that they used a het model.
Line 597: Do they have an accession number yet? Should make sure it is up by publication. GEO might be an appropriate archive as well.
Line 371-376: This may need to be revised given comment above.
Line 393/ref 55: The simpler point here is that there is an easier explanation for the lack of overlap between their baaSNPs and eQTLs: these cell lines are not neurons or glia. But, they get at this a little later in the para. Maybe just tweak the language a bit
Line 408: a STARR-seq assay tells you that an allele can be functional. It doesn't tell you it is functional in the brain. This is a different (albeit complementary) type of information than offered by eQTL studies.
Line 413: they should also note the limitation that an influence of the allele post-transcriptionally (i.e., on the RNA via transcript stability) will also conflate their results in a STARR-seq design.
Typos:
Line 75/76: assay -> assays
Line 234: _MSRA_ -> MSRA (note gene names should be italicized throughout manuscript).
Line 417: the last sentence is quite overstated.

Reviewer #3

(Remarks to the Author)
Summary:
Gao and colleagues use STARR-seq in various cell lines to screen through 351 candidate SNPs associated with schizophrenia to identify SNPs with biased allelic enhancer activity. They then used existing chromatin profiling information and eQTLs to identify genes linked to the SNPs, and validate one of their findings in zebrafish.
The manuscript is well organized and well written. The schematic in figure 1 describing the workflow of the computational strategy used to identify the SNPs was particularly useful. The experiments in zebrafish were well designed, and the qPCR analysis of pccb transcript in the heterozygous mutants where they find a ~50% reduction in pccb transcript levels was great validation of their mutation. I have no major notes on any of the zebrafish figures, results, discussion, or methods (just two notes in the minor comments) as everything was well organized and explained sufficiently. I have a few major questions, particularly about the choice of cell lines used, and an additional suggestion, but I do not think any of the major comments would necessitate rejection of the manuscript. The STARR-seq results are fine as is, but the authors may want to consider addressing the comments to potentially broaden the impact of the paper.

Major comments
I am a little curious as to why the authors chose the cell types they selected. The use of HEK-293T and SH-SY5Y seemed appropriate as they are human cell lines, but can the authors elaborate as to why they selected Neuro-2a and PC-12 over other human cell lines? I don't necessarily think the choice of two rodent cell lines is grounds for rejection but given that the authors are looking at SNPs in regulatory elements which may or may not be conserved, using other human cell lines, especially of different cell types (see below comment) could strengthen the paper.

Furthermore, using HEK-293 as a non-neuronal cell type seems like a good choice for a cell line to look at cells not found in the brain to complement the neuronal cell lines, but can the authors elaborate on why they focused only on neuronal cells as opposed to other glial cells found in the brain (astrocytes, microglia, etc.)? Their chromatin interaction analyses suggested roles of rs13072690 interacting with the PCCB promoter in glial cells, and while this result was likely found after the initial STARR-seq experiments, having other cell types to look for similarities or differences in different cell types in the brain would greatly strengthen the impact of the paper.

While the STARR-seq method was a great for going through the hundreds of SNPs they identified in figure 1, the authors note that this approach cannot identify genes influenced by the SNPs, and thus use alternate approaches from existing data sets to identify targets. The authors may want to consider for this manuscript or future manuscripts crispring one or two of their hits (like rs3088186 or rs13072690) into one or two human cell lines to see if their predicted relationships between the genes occurs in human cells in vitro. For example, taking a neuronal and a glial cell line and crispring in the two SNPs, which the authors note have more cell specific roles in the ATAC data, would strengthen their arguments of cell-type specificity.

Minor comments
Figure 5A – the alignment of the amino acids under the trinucleotide sequence is a little unaligned. It might make the most sense to have the amino acid centered under the trinucleotides, as right now, the amino acid skews off to the next set of nucleotides towards the end.
Line 585 – Can the authors add age in months or days of the adult zebrafish used, similar to what they reported about 6 month zebrafish being used for RNA extraction?
Figures 3F and 4C – The authors report p-values in every other graph but these two, so the authors may want to swap the asterisks for the p-values for these graphs to have consistent reporting.

Reviewer #4

(Remarks to the Author)
Summary:
Gao and colleagues use STARR-seq in various cell lines to screen through 351 candidate SNPs associated with schizophrenia to identify SNPs with biased allelic enhancer activity. They then used existing chromatin profiling information and eQTLs to identify genes linked to the SNPs, and validate the effects of one of the genes they found had reduced expression in one SNP in zebrafish.
The manuscript is well organized and well written. The schematic in figure 1 describing the workflow of the computational strategy used to identify the SNPs was particularly useful. The experiments in zebrafish were well designed, and the qPCR analysis of pccb transcript in the heterozygous mutants where they find a ~50% reduction in pccb transcript levels was great validation of their mutation. I have no major notes on any of the zebrafish figures, results, discussion, or methods (just two notes in the minor comments) as everything was well organized and explained sufficiently. I have a few major questions, particularly about the choice of cell lines used, and an additional suggestion, but I do not think any of the major comments would necessitate rejection of the manuscript. The STARR-seq results are fine as is, but the authors may want to consider addressing the comments to potentially broaden the impact of the paper.

Major comments
I am a little curious as to why the authors chose the cell types they selected. The use of HEK-293T and SH-SY5Y seemed appropriate as they are human cell lines, but can the authors elaborate as to why they selected Neuro-2a and PC-12 over other human cell lines? I don't necessarily think the choice of two rodent cell lines is grounds for rejection but given that the authors are looking at SNPs in regulatory elements which may or may not be conserved, using other human cell lines, especially of different cell types (see below comment) could strengthen the paper.

Furthermore, using HEK-293 as a non-neuronal cell type seems like a good choice for a cell line to look at cells not found in the brain to complement the neuronal cell lines, but can the authors elaborate on why they focused only on neuronal cells as opposed to other glial cells found in the brain (astrocytes, microglia, etc.)? Their chromatin interaction analyses suggested roles of rs13072690 interacting with the PCCB promoter in glial cells, and while this result was likely found after the initial STARR-seq experiments, having other cell types to look for similarities or differences in different cell types in the brain would greatly strengthen the impact of the paper.

While the STARR-seq method was a great for going through the hundreds of SNPs they identified in figure 1, the authors note that this approach cannot identify genes influenced by the SNPs, and thus use alternate approaches from existing data sets to identify targets. The authors may want to consider for this manuscript or future manuscripts crispring one or two of their hits (like rs3088186 or rs13072690) into one or two human cell lines to see if their predicted relationships between the genes occurs in human cells in vitro. For example, taking a neuronal and a glial cell line and crispring in the two SNPs,

which the authors note have more cell specific roles in the ATAC data, would strengthen their arguments of cell-type specificity.

Minor comments
Figure 5A – the alignment of the amino acids under the trinucleotide sequence is a little unaligned. It might make the most sense to have the amino acid centered under the trinucleotides, as right now, the amino acid skews off to the next set of nucleotides towards the end.

Line 585 – Can the authors add age in months or days of the adult zebrafish used, similar to what they reported about 6 month zebrafish being used for RNA extraction?

Figures 3F and 4C – The authors report p-values in every other graph but these two, so the authors may want to swap the asterisks for the p-values for these graphs to have consistent reporting.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
1. I understand the candidate variants (n=351) have been prioritized using a different approach than McAfee et al. In fact, it is encouraging to see 67 variants are shared between the two studies. While a full meta-analysis may not be feasible, the authors should expand on the STARR-seq results for these overlapping variants, particularly in relation to the 439 variants showing allelic regulatory effects in MPRA. Were any of these 67 variants functionally validated by both STARR-seq and MPRA? If so, do the predicted directions of effect agree across platforms? Please consider summarizing this comparison clearly in Tables S2–S3.

2. Figure 2 presents valuable data but would benefit from clearer visualization. For Fig. 2c, it is unclear whether this represents a meta-analysis across the four cell lines (e.g., using mpralm). Consider integrating Fig. 2c into Fig. 2d to improve interpretability and to directly indicate which candidate SNPs show biallelic effects.

3. Regarding Fig. 2d, the updated heatmap is a step forward, but interpretability is limited by the current color scale for logFC. The subtle gradation between -2 and +2 makes it difficult to discern the directionality of effects. I recommend revising the heatmap so that the full cell color reflects logFC (rather than p-value), using a diverging color scale centered at 0 (e.g., red for positive effects, blue for negative, white for neutral, similar Red-Blue colors to volcano plots in Fig. 2b), and capping the scale at -4 and +4 to avoid oversaturation from outliers. Statistical significance can be indicated with asterisks (e.g., $p<0.05$, $*p<0.01$, $**p<0.001$), which will allow for a more intuitive and informative display.

<Minor>
1. For all Supplementary Tables listing candidate SNPs, the inclusion of dbSNP “rsID” would be helpful for cross-referencing and clarity.

Reviewer #2

(Remarks to the Author)
Major
The authors appear to have misunderstood comment #17 (in their numbering system), which was also briefly mentioned in comment 1. Specifically, they need to rerun webgestalt not using the whole genome as the background set, but instead a background set that appropriately accounts for the fact that the eQTL data they are using comes from brain. Otherwise, it is highly likely that their GO term findings of things like Synapse and GABA are spurious. If they do find that they were spurious, they should remove that claim.
I recommend they use as background something like ‘all genes that have brain eQTLs’ in their source datasets. This should correct for the brain bias.
If their finding does go away, they should simply remove those claims from the paper. It is otherwise publishable and this is a minor part of the findings overall.

Minor
1. The authors may have missed the nuance of what I was trying to explain about the use of the word causal and over corrected. They also tried to just replace it with the word ‘risk’ which sometimes worked and some times did not. E.g., in line 61 or 64 it would be fine to use the word causal. I don’t argue that there are not causal variants in the genome. Rather, the point I was trying to make is that MPRAs don’t test causality for schizophrenia. Instead, they test functional potential of the variant. Logically, a variant must be functional to be causal, but it is also possible a variant is functional and not causal. So functionality is necessary, but not sufficient, for causality.
Thus, when the are talking about causality in general (like line 61 and 64) that is fine. But when they are claiming the discovered the causal variant, that is what needs to be edited more carefully. In some places, the word ‘risk’ in perfect. In other places, they may want to use the word ‘functional’ instead of risk. ‘Risk’ is usually taken to mean ‘ a variant that has

statistical association from human genetics studies of the disorder' So all the SNPs from the GWAS study are 'risk' alleles. But again, this does not guarantee function or causality.
2. The authors still talk about constraint in the discussion, but I think they might have removed that part of the analysis from the results?

Reviewer #3

(Remarks to the Author)
The authors have address my major concerns, and I have no other comments that need to be addressed.

Reviewer #4

(Remarks to the Author)
The authors have addressed my major concerns and I have no other comments that need to be addressed.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors' responses were satisfactory, and the figure updates have improved the overall readability of the manuscript. I have a few minor comments that should be addressed prior to publication.

For Table S2, I attempted to look up five SNPs overlapped with McAfee et al. but was unable to identify them. The overlap should indicate all SNPs intersected with the prior study, whether they were found within active enhancers or not. Additionally, the subset of SNPs located in silencers warrants further investigation for biased allelic activity (which could be considered for future work).

For Table S3, could the authors clarify the last column labeled "abeSNP"? Should this instead be "baaSNP"?

Lastly, I have reviewed Reviewer #2's comments and the authors' responses. In my opinion, the responses were appropriate, and the manuscript revisions adequately addressed the concerns raised.

Reviewer #1

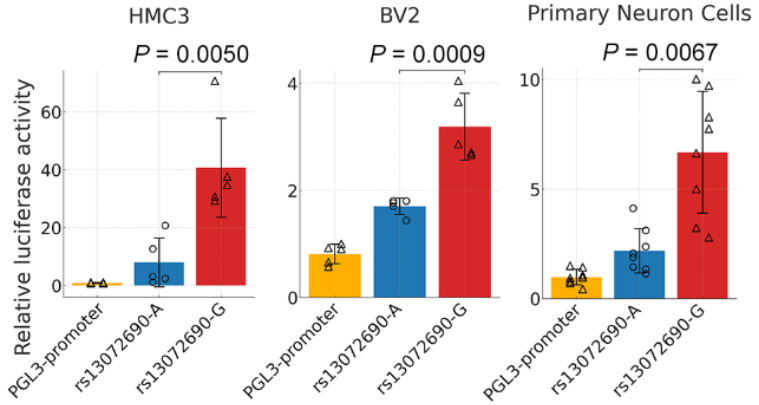
Comment Summary: Analysis of Biased Allelic Enhancer Activity of Schizophrenia-Linked Common Variants" by Dr Li and colleagues. The authors of this study investigated 351 common variants associated with schizophrenia (SCZ) to identify those with biased allelic enhancer activity using STARR-seq. They synthesized oligonucleotides containing reference and alternative alleles of these variants, cloned them into a STARR-seq vector, and identified SNP alleles that differentially affected enhancer activity. SNPs with biased allelic enhancer effects were further validated using eQTL, Hi-C data, and behavioral assessment of zebrafish with gene knockouts. The study is well-designed, and the manuscript is well-written and easy to follow. The results are convincing, and the methods are sound. However, there are several areas where the manuscript can be improved:

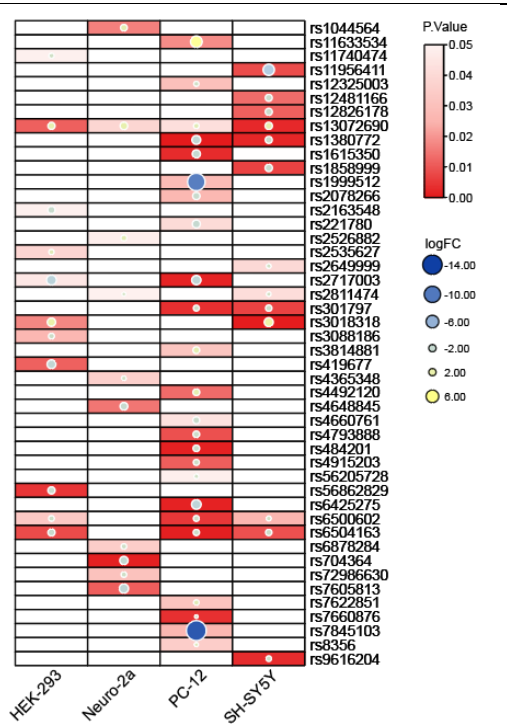
Author Response: We thank the reviewer for the comprehensive summary and constructive feedback. In response, we have revised the manuscript to incorporate the suggested improvements, detailed point by point as below.

	Comment	Author Response	Modification made in manuscript
1	Comparison with previous studies: Several previous studies have used similar high-throughput screening assays to assess the functional impact of variants on enhancers in disease contexts. The authors should compare their results with these previous findings and perform a meta-analysis if possible (ref: https://doi.org/10.1126/science.adh4265, https://doi.org/10.1016/j.xgen.2023.100404).	We greatly appreciate the reviewer's thoughtful comments. In response to your suggestion, we have updated **Table S1** to include a comparison with previous studies that utilized high-throughput screening assays to assess the functional impact of variants on enhancers in disease contexts. McAfee *et al.* (https://doi.org/10.1016/j.xgen.2023.100404) analyzed 5,173 fine-mapped schizophrenia GWAS variants in primary human neural progenitors using the FINEMAP tool to identify allelic regulatory effects, whereas our study employed a tiered multi-omics strategy to predict 351 functional variants, with only 67 variants overlapping between the two studies. This comparison highlights inherent differences in datasets and methodologies that could account for the observed discrepancies. Regarding the study by Zeng *et al.* (https://doi.org/10.1126/science.adh4265),	The modifications were made in **discussion section, pages 20-21 lines 381-394**. We also **updated Table S1** in the revision.

		although they examined 19,893 potential causal variants, the article does not provide a complete list of these variants, preventing a direct comparison. By integrating these prior studies into our manuscript, we offer an integrative perspective that complements these approaches by combining functional screening with multi-layered omics data. We appreciate these valuable suggestions, as this cross-study comparison not only underscores the promise and limitations of current methodologies but also lays the groundwork for future meta-analyses to consolidate findings across diverse datasets.	
2	Emphasis on *PCCB* as a target gene: The manuscript emphasizes the discovery of *PCCB* as a target gene candidate associated with SCZ risk, but this gene has previously been reported to be linked with multiple neuropsychiatric traits (ref: https://www.nature.com/articles/s41467-023-40861-2, https://doi.org/10.1186/s12920-022-01202-2, https://doi.org/10.1016/j.ajhg.2022.07.011). Authors should, at minimum, cite prior studies and elucidate some of the connections of this work to previous ones.	We thank the reviewer for your valuable suggestions. In response, we have expanded the Discussion section of the manuscript to include references to these previous studies and to clarify the connections between our findings and earlier work. As noted in the revised manuscript, prior research has identified *PCCB* as a key gene involved in neuropsychiatric disorders. Notably, Zhang *et al.* (https://www.nature.com/articles/s41467-023-40861-2) demonstrated that downregulation of *PCCB* in human forebrain organoids contributes to schizophrenia-related phenotypes—such as increased neuroactivity and impaired neural synchronization. Additionally, other studies (https://doi.org/10.1186/s12920-022-01202-2, https://doi.org/10.1016/j.ajhg.2022.07.011) have further	The modifications were made in **discussion section, page 22 lines 409-423**.

		connected *PCCB* to neuropsychiatric conditions including schizophrenia and bipolar disorder, reinforcing its potential involvement in these diseases. In our study, we observed that zebrafish with reduced *pccb* expression exhibited hyperlocomotion, heightened anxiety-like behaviors, and social deficits—phenotypes commonly associated with schizophrenia. These convergent findings provide additional support for the involvement of *PCCB* in schizophrenia pathology, particularly in its role in regulating social behavior and anxiety.	
3	Limitations of using immortalized cell lines: The use of immortalized cell lines may not accurately capture the gene regulation program and disease biology. The authors should consider validating candidate genes in organoids or primary cell cultures.	Thank you for your comment. In response, we have expanded our experimental validation by performing dual-luciferase assays in human microglial HMC3 cells, mouse microglial BV2 cells, and mouse cortical primary neurons—cell models that offer a more physiologically relevant context than immortalized lines. We have also revised the Discussion to explicitly acknowledge that while immortalized cell lines serve as useful initial systems, they may not fully capture the complexities of native gene regulation and disease biology. Furthermore, we note that future research will extend validations to organoids or primary cell cultures to further substantiate our findings.	The modifications were made in **page 14 lines 251-264** and **pages 27-28 lines 552-560**. We also **updated Figure 4** in the revision.

			
4	Clarity of Figure 2d: The description of Figure 2d does not match the figure. The specificity and effect size of SNPs can be better visualized. The authors should consider showing a heatmap using both color and size to represent effect size and statistical significance for all SNPs that are significant (|FC| > 1.5 & FDR < 0.05) in at least one of the cell types. Asterisks can be used to highlight SNPs prioritized as baaSNPs.	Thank you for your constructive suggestion regarding Figure 2D. In response, we have revised Figure **2D** to include both statistical significance (p-values) and fold change (FC) information as recommended. Specifically, we used a heatmap to represent the metrics of these baaSNPs, where both color and size are used to visualize the effect size and statistical significance of SNPs in at least one of the cell types. These adjustments improve the clarity of the figure and better align the visual representation with the figure description.	The modifications were made in **page 11 lines 178-182**. We also **updated Figure 2D** in the revision.


rs1044564
rs11633534
rs11740474
rs11956411
rs12325003
rs12481166
rs12826178
rs13072690
rs1380772
rs1615350
rs1858999
rs1999512
rs2078266
rs2163548
rs221780
rs2526882
rs2535627
rs2649999
rs2717003
rs2811474
rs301797
rs3018318
rs3088186
rs3814881
rs419677
rs4365348
rs4492120
rs4648845
rs4660761
rs4793888
rs484201
rs4915203
rs56205728
rs56862829
rs6425275
rs6500602
rs6504163
rs6878284
rs704364
rs72986630
rs7605813
rs7622851
rs7660876
rs7845103
rs8356
rs9616204
HEK-293
Neuro-2a
PC-12
SH-SY5Y
P.Value
0.05
0.04
0.03
0.02
0.01
0.00
logFC
-14.00
-10.00
-6.00
-2.00
2.00
6.00

Reviewer #2

Comment Summary: The authors aim to bridge the gap between genetically associated and functional roles for variants in SCZ using a STARR-seq screening method. Additionally, they characterized a potential causal target genes in zebrafish models. From their screen, they identified 351 functional SNPs linked to 217 target genes through chromatin interaction assays and eQTL analysis. They also show that a KO zebrafish model, which is linked to one of the identified SNPs, exhibits behavioral differences concluding that some of the variants identified in their screen are causative of SCZ. It seems that the authors were very thoughtful in their approach to designing the MPRA and how they chose candidate SNPs. Additionally, the use of multiple cell lines enabled them to disambiguate cell type specific effects from more generalized effects. However, my main concern is that the authors continually state they have identified causal SNPs linked to SCZ, yet they have no experiments that directly validate these claims. On the other hand, they may have identified a functionally important gene (*PCCB*) in SCZ, and much of the data is strong and well executed in the Z fish and MPRA, so there is some good work here to get out into the world. Them may need to dampen some of their claims, which were:

1. They assessed regulatory activity of the variants across 4 cell lines, and identify 46 SNPs with allelic effects (baaSNPs).
2. Bioinformatics of existing epigenetic data linked these to 217 genes.
3. GO links these genes to synapses and GABA.
4. They focus on rs13072690 for a bit of follow up, which had a consistent impact on luciferase, and robustly altered behavior in zebrafish when knocked out as a het.
5. The study provides a comprehensive approach for identifying causative variants

Review: Overall, this is a useful paper. It makes two main contributions with new data (claims 1 and 4) above. The experiments for those are generally well executed and analyzed in field standard ways, though I had a few minor questions.

Author Response: We greatly appreciate the reviewer's constructive feedback and the opportunity to refine our manuscript. In response, we have made several revisions to improve clarity and address potential concerns about our claims, detailed point by point as below.

	Comment	Author Response	Modification made in manuscript
1	I think claim 3 could be supported, but they need to rerun the GO analysis with the appropriate background set to be sure. I think claim 5 is overstated. While this is useful work and should be	We appreciate the reviewer's constructive comments. In response, we have revised the manuscript to ensure that claim 3 is supported by rerunning the Gene Ontology (GO) analysis with	The modifications were made in **page 3 line 46-51** and **page 7 line 108**.

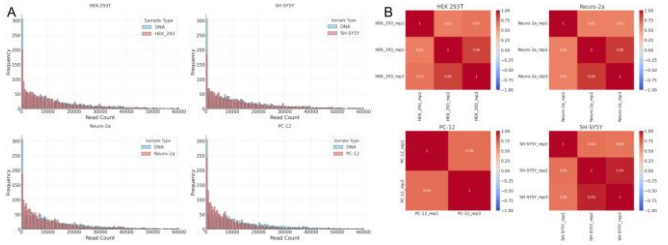
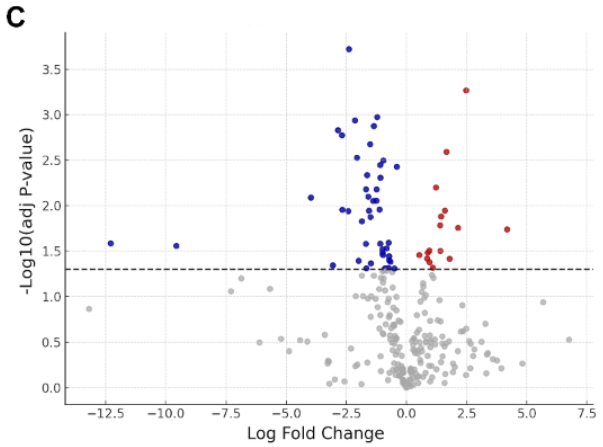
	published, claims this is comprehensive are a bit much (other instances noted in minor comments below): MPRAs/STARRseq generally are useful at determining which sequences can have a functional effect and are thus important complements to epigenetic and high-C approaches, which provide biochemical evidence a sequence might have a role, but don't directly measure functional outputs (i.e., transcription). The authors should reduce the language around their claim 5 or remove it entirely.	the appropriate background set. The updated results now provide a more accurate representation of the enrichment findings, thereby strengthening the evidence for this claim. Regarding claim 5, while we acknowledge that techniques such as MPRAs/STARR-seq are valuable for identifying functional sequences—and serve as important complements to epigenetic and high-throughput chromatin approaches that suggest potential regulatory roles—they do not directly measure functional outputs like transcription. In light of this, we have adjusted our language to avoid overstatement and ensure that our conclusions accurately reflect the scope and limitations of these methods.	
2	MPRAs/STARRseq are very cell type dependent, and while these were convenient choices for transfectable cell lines, none of these cell lines really match the neurons (and perhaps microglia) implicated in SCZ in terms of their transcription factor expression, etc. Thus, the lack of alignment with brain eQTL data, etc. they see is fairly unsurprising. I am not sure there is really any way to correct this short of repeating the whole set of experiments in neurons instead. And that feels like too high a bar for a revision. Rather, that would be	We thank the reviewer for the insightful feedback. In response, we acknowledge that the immortalized cell lines used in our initial STARR-seq screening—Neuro-2a, SH-SY5Y, HEK-293T, and PC-12—do not fully replicate the transcriptional environment of human neurons or microglia implicated in schizophrenia. Consequently, the limited overlap with brain eQTL datasets is not surprising and underscores the importance of interpreting these findings	The modifications were made in **page 14 lines 251-264, page 23 lines 434-445,** and **pages 27-28 lines 552-560**. We also **updated Figure 4** in the revision.

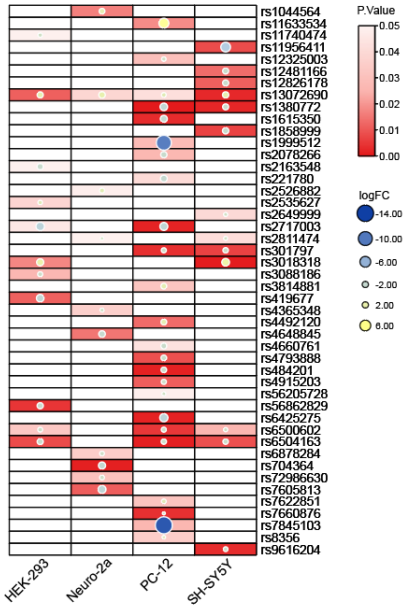
	a whole new paper – and would require porting their library over to some other plasmid system that enables transduction of neurons (e.g. lenti). If they want to publish the current work here (at the editors discretion, of course), they would need to instead really carefully frame what they can interpret from these cells. And they should definitely reduce the claims about the comprehensiveness of the pipeline, establishing causality, etc. They are adding functional evidence for these SNPs, but this work alone is not definitive for establishing causality (see also specific instances of lines to change in minor notes).	cautiously. To address this limitation and strengthen our findings, we conducted additional dual-luciferase assays in more physiologically relevant cellular models: human microglial HMC3 cells, mouse microglial BV2 cells, and mouse cortical primary neurons. The results from these experiments showed consistent biased allelic enhancer activities across all tested cell types, thereby reinforcing our initial observations from the STARR-seq platform and providing further evidence for allele-specific regulatory effects in neuronal and glial contexts. Nevertheless, we have moderated our claims concerning the comprehensiveness of our pipeline and the establishment of causality. We now emphasize that while our approach adds valuable functional evidence for these SNPs, it does not definitively establish causality.	
3	I do like that they chose to follow up on the variant that seemed to be the most consistent across cell types, and in the luciferase assay as well. It suggests this one might not depend on cell type as much, and gets around my concern in #2 above a	We appreciate the reviewer highlighting this important point. We chose to follow up on the variant that showed the most consistent results across cell types and in the luciferase assay because it suggested more robust and potentially	The modifications were made in **page 14 lines 251-264, page 22 lines 423-429,** and **pages 27-28 lines**

	bit. However, the one puzzle is that the direction of effect appears to be opposite between Luci/STARRseq on one hand, and eQTL data (and Z-fish follow up) on the other. The authors describe this, but don't directly discuss this. How should the reader think about this difference? Is the STARR-seq in cell lines not predictive, and thus the STARR-seq hitting on this allele a lucky accident? Or would the direction of effect be different in a neuron (or key glia, given figure 5) than in these cells? Repeating the luciferase assay in such cells might be a doable experiment, as some kinds of glia are culturable from human and/or mouse tissues (though it is unclear from their hi-c data which glia type one should consider. If it is not clear from there, the available scRNAseq of the expression of *PCCB* might help guide a choice).	broader regulatory effects. Indeed, we noted the discrepancy between the direction of effect observed in our STARR-seq and luciferase assays versus the direction inferred from eQTL data. This inconsistency may result from the intrinsic differences between assays: STARR-seq evaluates enhancer activity in an episomal context without the native chromatin structure and long-range genomic interactions that are present in eQTL studies, which reflect endogenous genomic contexts and complex regulatory interactions. Importantly, similar observations of inconsistent directionality between MPRA/STARR-seq assays and eQTL data have been documented previously (e.g., McAfee *et al.* (https://doi.org/10.1016/j.xgen.2023.100404)), underscoring the challenges in interpreting the functional regulatory effects of genetic variants across different platforms. To further clarify this discrepancy, we performed additional dual-luciferase assays in primary cortical neurons and microglial cell lines that confirmed our initial STARR-seq findings, thereby supporting the	**552-560**. We also **updated Figure 4** in the revision.

		robustness of our results. Nevertheless, we recognize that these findings may still be subject to context-dependent effects, and to fully resolve the apparent differences, future studies employing allele-specific knock-in experiments in a physiologically accurate context—guided by single-cell expression data for *PCCB*—will be essential. We have now explicitly discussed this caveat in the revised manuscript and highlighted it as a critical avenue for future research.	
4	STARRseq designs conflate posttranscriptional effects and transcriptional effects (https://pmc.ncbi.nlm.nih.gov/articles/PMC7727316/). This is not yet well recognized, but they should discuss this caveat/limitations.	Thank you for your insightful comment. In response, we have revised the manuscript to include a discussion of the potential limitations of the STARR-seq approach, particularly its tendency to conflate posttranscriptional and transcriptional effects. We have cited the relevant study (https://pmc.ncbi.nlm.nih.gov/articles/PMC7727316/) and acknowledged that this is an important caveat. This discussion now appears in the revised manuscript to ensure that readers are aware of the limitations inherent in this method.	The modifications were made in **page 24 lines 471-481**.
5	Line 33: Maybe remove the word 'relevant' here as	Thank you for your comment. In response, we	The modifications

	none of these cell lines get schizophrenia. Similar around line 99, and a few other places in the text.	have removed the word "relevant" from line 33 as well as from similar instances throughout the manuscript.	were made in **page 3 line 33**, **page 6 line 102,** and **page 27 line 537**.
6	Line 46: This claim is a bit too strong. A truly comprehensive approach would involve CRISPR editing of the SNP in the genome in a brain relevant cell type to confirm the change in target gene expression. That would be beyond the scope of this paper, so I recommend the authors adjust the language a bit here. Something like 'our study presents an approach for functionally annotating putative causative variants'	Thank you for your thoughtful comment. In response, we have revised the language on line 46 to better reflect the scope of our study. We agree that a truly comprehensive approach would involve CRISPR editing of the SNP in a brain-relevant cell type to confirm changes in target gene expression. Therefore, we have adjusted the phrasing to: "our study presents an approach for functionally annotating putative causative variants."	The modifications were made in **page 3 line 46-51**.
7	Related to the major point above, the word causal is used repeatedly when it is sometimes inappropriate (e.g. line 92 – multiomics data can identify risk variants not necessarily causal). This is especially true in the first sentence of the last paragraph of the discussion (lines 414-416), which to me, is very bold statement given the data because they didn't actually show phenotypic data associated with any SNPs.	Thank you for your valuable feedback. In response, we have carefully revised the manuscript and replaced the term "causal variants" with "risk variants" throughout the text. This change is applied in the instance you mentioned as well as in the first sentence of the last paragraph of the Discussion, ensuring that our language accurately reflects the data without overstatement.	The modifications were made in **page 6 line 97** and **page 24 line 462-465**.
8	Lines 503-505 say that statistical tests were used for filtering, but it might be nice to see this data as	Thank you for your valuable suggestion. In response, we have added Supplementary Figure	The modifications were made in **page 40**

	a supplementary figure (e.g. read count cut-off/distribution of reads and correlation/PCA for counts data) especially since the analysis techniques are not canonical methods.	S2, which visualizes the read count distribution and includes a correlation plot for the counts data. 	**lines 924-929**. We also **added Figure S2** in the revision.
9	To me, figure 2C would be way more interpretable by another volcano plot, which provides more data than this circle plot, since I am more interested in effect size and p-values than genomic coordinates.	Thank you for your helpful suggestion. In response, we have revised Figure 2C by replacing the circle plot with a volcano plot. This updated figure now provides a clearer representation of the data, focusing on effect size and p-values rather than genomic coordinates. 	The modifications were made in **page 11 line 175**. We also **updated Figure 2C** in the revision.
10	Fig 1: Fonts on some figure panels are too small to read without 200% zoom (especially 1C, D). Try and use 6 pt minimum. Check other figures as well.	Thank you for your comment. In response, we have increased the font size in Figure 1, particularly in panels 1C and 1D, to ensure legibility without requiring zooming. We have	We have **updated Figure 1** in the revision.

		also reviewed the other figures and adjusted their font sizes as needed, ensuring they meet the minimum recommended size of 6 pt.	
11	Fig 2D: This is rather small to read. Perhaps it is better as a long strip across the bottom of the figure, instead of a circle.	Thank you for your suggestion. In response, we have revised Figure 2D—replacing the circular representation with a long strip heatmap format across the bottom of the figure. This new layout improves readability and provides a clearer visualization of the data. 	We have **updated Figure 2D** in the revision.
12	Methods: the MPRA library had no controls (positive enhancers, repressors, shuffled sequence, minimal promoter only…) listed. Were there any	Thank you for your insightful comment. We have revised the methods section to provide a more detailed explanation of sequencing library	The modifications were made in **page 29 lines 587-594** and

	controls built in? How did they perform? How did they move from raw sequences to count level data? This part of the methods was quite unclear. Relatedly, for sequencing library construction, were the entire regulatory elements amplified and sequenced with full coverage, and what alignment tool was used (methods line 507)?	construction and the processing of raw sequence data into count-level data. During sequencing library construction, the entire regulatory elements were fully amplified and sequenced. Sequencing data were processed using the DADA2 (v1.26) pipeline to generate amplicon sequence variant (ASV) tables at single-nucleotide resolution—in this study, the ASVs represent the entire regulatory elements. The DADA2 algorithm learns error rates directly from the dataset and applies these models to infer true biological variants, even distinguishing single nucleotide differences, resulting in an ASV abundance table that details the count-level abundance of each unique sequence variant across all samples. Regarding controls, the library did not include standard controls such as positive enhancers, repressors, shuffled sequences, or minimal promoter-only constructs. Instead, we relied on technical replicates and cross-validation across multiple cell lines to assess the reproducibility and reliability of the observed regulatory effects. We have acknowledged in the Discussion that the absence of these controls is a limitation of the	**page 24 line 476-481**.

		current design.	
13	There appears to be no code sharing plan, link to git repository, etc..	Thank you for your comment. In response, we have made our code for STARR-seq data analysis available on GitHub. The repository can be accessed at https://github.com/gaochengwen/STARR-seq-Data-Analysis.	The modifications were made in **page 34 lines 689-695**.
14	The figure legend for 3F says rs13072690 despite the text and figure legend saying rs3088186.	Thank you for pointing out the inconsistency. In response, we have corrected the figure legend for Figure 3F to accurately reflect that the data correspond to rs3088186, consistent with both the text and the revised figure legend. We appreciate your attention to detail, as this correction ensures consistency throughout the manuscript.	We have **updated Figure 3F** in the revision.
15	Line 457 – minor note: InFusion products are not ligation products. It does not involve ligase.	Thank you for pointing this out. Based on the reviewer's comment, we have revised the description to accurately reflect the process.	The modification was made in **page 26 line 531**.
16	Line 523: Were they cloned into the 3'UTR (in the STARR seq style) or upstream of a minimal promoter for the luciferase assay? Maybe clarify this in the methods. I think it is STARRseq	We greatly appreciate the reviewer's comment regarding the luciferase assay. In response, we have revised the Methods section to clarify that, in our validation experiments, enhancer sequences were cloned into a KpnI-XhoI linearized pGL3-Promoter luciferase vector—an approach typical of traditional luciferase reporter assays. Previous studies have shown that this method compares	The modification was made in **page 30 lines 611-614**.

		favorably with STARR-seq when validating human enhancer activities (doi:10.1186/s13059-017-1345-5 and doi:10.1016/j.stem.2018.06.014).	
17	Line 543: What gene set did they set as background (a.k.a reference set) for WebGestalt? An appropriate choice might be all genes with eQTLs in their brain dataset. This controls for the general brain-bias that would be expected to be present in their brain eQTL derived list, and that could thus produce spurious enrichments of categories like 'synapse' and 'GABA' by chance alone.	Thank you for your comment. In our WebGestalt analysis, we used the entire genome's gene set as the background. In response, we have updated the Methods section to clarify our background set selection and to discuss its implications.	The modifications were made in **page 31 lines 632-633.**
18	Line 138: please provide a reference for this vector.	Thank you for your comment. In response, we have added the reference (https://www.nature.com/articles/nmeth.4534) for the vector on line 138.	The modification was made in **page 8 line 143**.
19	The authors clone each SNP into three positions. How do they deal with this data? Is each location of the same SNP treated independently in their model (i.e., each SNP is tested three times)? Does position matter? Are there SNPs that have the same effect regardless of position?	Thank you for your thoughtful questions. Each SNP was cloned into three distinct positions (−50 bp, 0 bp, and +50 bp relative to the SNP) to evaluate the impact of positional context on enhancer activity. For the analysis, each position was treated independently, meaning that each SNP was effectively tested three times. Our results, summarized in Table S2, indicate that most SNPs exhibit similar effects regardless of position, suggesting that the positional context does not	The modification was made in **page 26 line 520-522**.

		significantly alter enhancer activity for the majority of the tested SNPs. This design also ensures robust representation of each SNP, reducing the potential impact of missing data from any single fragment.	
20	Line 185: the phrasing here should be tweaked in a minor way. The genes came from HEK-293T STARRseq data, rather than the cells themselves.	Thank you for your comment. In response, we have revised the phrasing to clarify that the genes originated from the HEK-293T STARR-seq data rather than directly from the cells themselves.	The modification was made in **page 11 line 195**.
21	Figure 3A: It is a bit unclear what the p-value in the figure is here. Is this p-values from the eQTL study? Or are they plotting their own p-values?	Thank you for your comment. In response, we have clarified in the figure legend that the p-values shown in Figure 3A are from the eQTL study. The figure legend has been updated to provide this clarification for better understanding.	The modifications were made in **page 13 lines 230-231**.
22	One weakness to Hi-C data is that it will have false negatives based on some aspects of the library prep (i.e., loop size depends on how the library was made). The authors should mention this in the discussion. It may also explain why there is little overlap with eQTLs. Though of course, eQTLs also are ambiguous in linking SNPs to genes, because many SNPs nearby the functional SNP will be in LD.	Thank you for your thoughtful comment. In response, we have added a discussion of the limitations of Hi-C data in the Discussion section. Specifically, we now acknowledge that aspects of the library preparation—such as loop size—can result in false negatives, which may partly explain the limited overlap with eQTLs. We also note the inherent ambiguity in eQTL analyses when linking SNPs to genes, as many SNPs near a functional variant are often in linkage disequilibrium (LD).	The modifications were made in **pages 23-24 lines 456-460**.

23	Figure 3E: I don't see the difference here the authors are pointing out. Further, such a claim would need statistical support. Overall, it may be simpler to simply remove the panel and the related sentence: I am not sure this data is central to their claims. But, perhaps I am missing something? Also, line 204 and legend for 3E seem to disagree with each other on direction of effect?	Thank you for your thoughtful comment. In response, we have removed the results related to Figure 3E, as we agree that this data does not strongly contribute to the overall claims of the paper. Additionally, the inconsistency between line 204 and the legend for Figure 3E regarding the direction of effect has been addressed, as we no longer include this panel in the manuscript	We have **updated Figure 3** in the revision.
24	Line 207: Was this the only variant they tested? Or the only one that met their prediction? It seems pretty strange they would pick to focus on this one.	Thank you for your comment. In response, we clarify that rs3088186 was not the only variant tested; other variants, including rs13072690, were also analyzed. We focused on rs3088186 as a representative example of a gene interacting with a biased allelic SNP, specifically because it is located within the *MSRA* locus—a gene involved in cellular antioxidant defense that has previously been associated with SCZ.	The modifications were made in **page 12 lines 214-215**.
25	Line 266: which kind of glia cells is the Hi-C from?	Thank you for your comment. In response, we have clarified in the manuscript that the Hi-C data is derived from microglia cells. The relevant section of the text has been updated to reflect this.	The modification was made in **page 16 line 284**.
26	Figure 6E/F: the Y-axis label is the same. Should one of these be different?	Thank you for your comment, and we apologize for the confusion. In response, we have corrected the figure legend for Figure 6E/F. The Y-axis labels for these two panels were mistakenly	The modifications were made in **page 20 lines 359-364**.

		swapped. The correct labels are as follows: Figure 6E: Time spent near other fish without direct contact, reflecting social interest without physical engagement. Figure 6F: Time spent in direct body contact with other fish, indicating levels of social interaction.	
27	Sample size in fish studies was good, and I like that they used a het model.	Thank you for your positive feedback. We appreciate your comment on the sample size and the use of a heterozygous (het) model. As noted, we used a het model in our study, which allowed for a more accurate representation of the phenotypic effects and is aligned with the goals of our analysis.	The modification was made.
28	Line 597: Do they have an accession number yet? Should make sure it is up by publication. GEO might be an appropriate archive as well.	Thank you for your comment. In response, we would like to inform you that the raw sequencing data has already been uploaded to GEO, and the accession number is **GSE290050**. This will be available by the time of publication.	The modifications were made in **page 34 lines 689-695**.
29	Line 371-376: This may need to be revised given comment above.	Thank you for your comment. In response to your previous feedback, we have revised the manuscript accordingly.	The modifications were made in **page 22 lines 409-421**.
30	Line 393/ref 55: The simpler point here is that there is an easier explanation for the lack of overlap between their baaSNPs and eQTLs: these cell lines are not neurons or glia. But, they get at	Thank you for your helpful comment. Based on your suggestion, here's a revised version of the paragraph with adjusted language: This limited overlap may be partly due to differences in cell	The modifications were made in **page 23 lines 434-445**.

	this a little later in the para. Maybe just tweak the language a bit.	types and developmental stages between eQTL studies, which often use heterogeneous adult brain tissues, and our experiments, which were conducted in specific cell lines. Since these cell lines do not fully represent the neuronal or glial environments of the brain, the observed lack of overlap could also reflect this difference in cellular context.	
31	Line 408: a STARR-seq assay tells you that an allele can be functional. It doesn't tell you it is functional in the brain. This is a different (albeit complementary) type of information than offered by eQTL studies.	Thank you for your comment. Based on your feedback, here is the revised version of the sentence: "Overall, our study demonstrates that characterizing GWAS variants with STARR-seq can reveal functional regulatory variants that traditional eQTL analyses might miss, offering complementary insights into allele-specific regulatory effects while not confirming their functionality in the brain"	The modifications were made in **page 24 lines 462-465**.
32	Line 413: they should also note the limitation that an influence of the allele post-transcriptionally (i.e., on the RNA via transcript stability) will also conflate their results in a STARR-seq design.	Thank you for your insightful comment. In response, we have revised the manuscript to include a discussion of the potential limitations of the STARR-seq approach, specifically regarding the conflation of posttranscriptional and transcriptional effects. We have cited the relevant study (PMC7727316, https://www.nature.com/articles/s41592-020-	The modifications were made in **page 24 lines 471-481**.

		0965-y) and acknowledged that this is an important caveat, which we now discuss in the revised manuscript.	
33	Line 75/76: assay -> assays	Thank you for your comment. we have made the necessary revision and changed "assay" to "assays".	The modification was made in **page 6 line 79**.
34	Line 234: _MSRA_ -> *MSRA* (note gene names should be italicized throughout manuscript).	Thank you for your comment. we have made the necessary revision and changed "_MSRA_" to "*MSRA*".	The modifications were made in **page 14 lines 244-245**.
35	Line 417: the last sentence is quite overstated.	Thank you for your comment. Based on your suggestion, here is the revised version of the sentence: "The findings delineate potentail candidates warranting further investigation into the disorder's etiological pathways."	The modifications were made in **page 25 lines 482-485**.

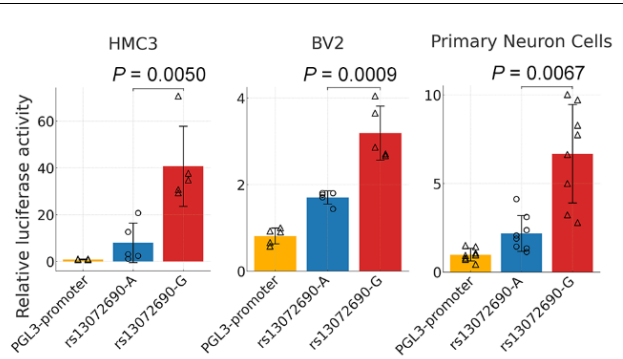
Reviewer #3 and Reviewer #4 (co-reviewer)

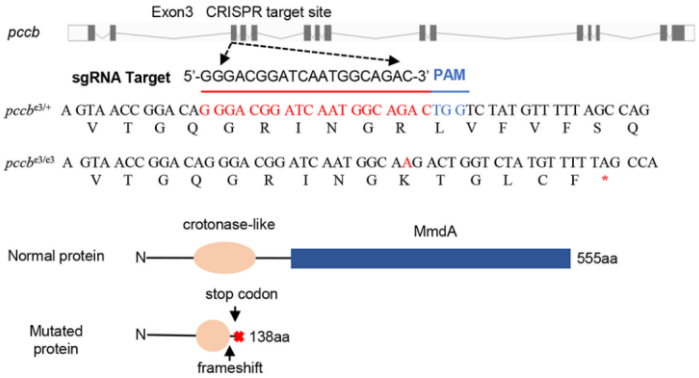
Comment Summary: Gao and colleagues use STARR-seq in various cell lines to screen through 351 candidate SNPs associated with schizophrenia to identify SNPs with biased allelic enhancer activity. They then used existing chromatin profiling information and eQTLs to identify genes linked to the SNPs, and validate one of their findings in zebrafish. The manuscript is well organized and well written. The schematic in figure 1 describing the workflow of the computational strategy used to identify the SNPs was particularly useful. The experiments in zebrafish were well designed, and the qPCR analysis of *pccb* transcript in the heterozygous mutants where they find a ~50% reduction in *pccb* transcript levels was great validation of their mutation. I have no major notes on any of the zebrafish figures, results, discussion, or methods (just two notes in the minor comments) as everything was well organized and explained sufficiently. I have a few major questions, particularly about the choice of cell lines used, and an additional suggestion, but I do not think any of the major comments would necessitate rejection of the manuscript. The STARR-seq results are fine as is, but the authors may want to consider addressing the comments to potentially broaden the impact of the paper.

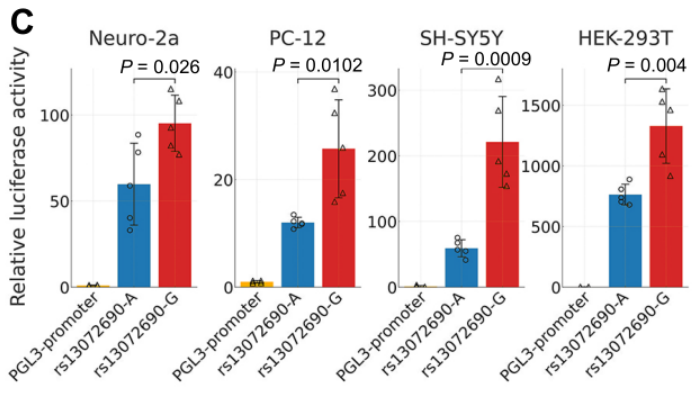
Author Response: We greatly appreciate the reviewer's constructive feedback and the opportunity to refine our manuscript, and in response, we have revised our work to address your concerns as follows.

	Comment	Author Response	Modification made in manuscript
1	I am a little curious as to why the authors chose the cell types they selected. The use of HEK-293T and SH-SY5Y seemed appropriate as they are human cell lines, but can the authors elaborate as to why they selected Neuro-2a and PC-12 over other human cell lines? I don't necessarily think the choice of two rodent cell lines is grounds for rejection but given that the authors are looking at SNPs in regulatory elements which may or may not be conserved, using other human cell lines, especially of different cell types (see below	Thank you for your insightful comment. We selected the cell lines HEK-293T, SH-SY5Y, Neuro-2a, and PC-12 based on their relevance to neuronal biology and experimental utility. Specifically, HEK-293T cells were chosen for their high transfection efficiency and robust reporter assays, SH-SY5Y cells for their neuronal characteristics, Neuro-2a cells for studying neuro developmental signaling https://doi.org/10.1016/j.bbrc.2010.07.112), and PC-12 cells for differentiation and neuroendocrine studies (https://doi.org/10.3390/app11093729). To address the limitation you mentioned, we further	The modifications were made in **page 14 lines 251-264** and **pages 27-28 lines 552-560**. We also **updated Figure 4** in the revision.

	comment) could strengthen the paper.	validated our findings using dual-luciferase assays in **primary neuronal cells and microglial cells**, which are more relevant to the study of regulatory elements in a more physiological context.	
2	Furthermore, using HEK-293 as a non-neuronal cell type seems like a good choice for a cell line to look at cells not found in the brain to complement the neuronal cell lines, but can the authors elaborate on why they focused only on neuronal cells as opposed to other glial cells found in the brain (astrocytes, microglia, etc.)? Their chromatin interaction analyses suggested roles of rs13072690 interacting with the *PCCB* promoter in glial cells, and while this result was likely found after the initial STARR-seq experiments, having other cell types to look for similarities or differences in different cell types in the brain would greatly strengthen the impact of the paper.	We appreciate the reviewer's thoughtful comments. Initially, we selected HEK-293T cells as a non-neuronal control due to their high transfection efficiency and robust performance in reporter assays, complementing our neuronal cell lines. We agree with the reviewer's suggestion regarding the importance of including glial cell types, especially given that chromatin interaction analyses indicated a potential regulatory role for rs13072690 in glial cells. In response to this valuable feedback, we have performed additional dual-luciferase reporter assays in human microglial HMC3 cells and mouse microglial BV2 cells, along with mouse cortical primary neurons. These supplementary experiments demonstrated consistent allele-specific enhancer activity across these glial and neuronal contexts, significantly strengthening the conclusions and broadening the biological relevance of our findings. We have updated the manuscript accordingly to discuss these results, emphasizing the importance of considering diverse brain cell types in the functional interpretation of schizophrenia-associated genetic variants.	The modifications were made in **page 14 lines 251-264, page 23 lines 434-445,** and **pages 27-28 lines 552-560**. We also **updated Figure 4** in the revision.

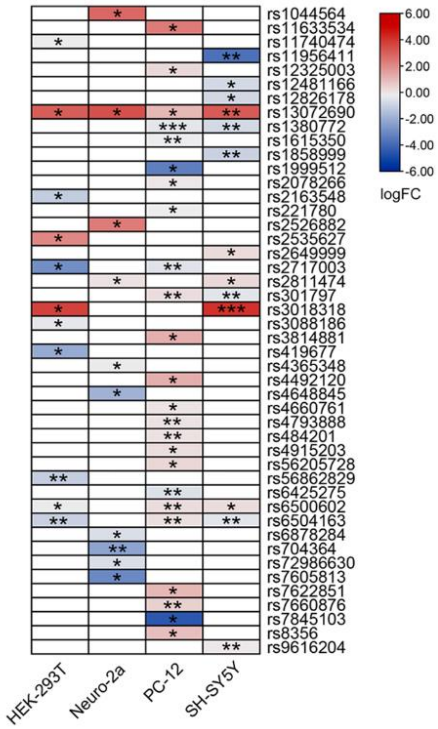
			
3	While the STARR-seq method was a great for going through the hundreds of SNPs they identified in figure 1, the authors note that this approach cannot identify genes influenced by the SNPs, and thus use alternate approaches from existing data sets to identify targets. The authors may want to consider for this manuscript or future manuscripts crispring one or two of their hits (like rs3088186 or rs13072690) into one or two human cell lines to see if their predicted relationships between the genes occurs in human cells in vitro. For example, taking a neuronal and a glial cell line and crispring in the two SNPs, which the authors note have more cell specific roles in the ATAC data, would strengthen their arguments of cell-type specificity.	Thank you for the insightful suggestion. We appreciate your recommendation to incorporate CRISPR experiments in human cell lines, such as neuronal and glial cell lines, to further validate the predicted relationships between genes and SNPs like rs3088186 or rs13072690. We agree that such experiments would provide valuable in vitro evidence to directly test whether the predicted effects of these SNPs on gene regulation are indeed observed in human cells. We have acknowledged this in the revised manuscript **discussion section** as a future direction, emphasizing the potential of CRISPR technology to confirm our findings and further explore the functional implications of these SNPs in relevant human cellular contexts. This experimental validation will be a key direction in our future research.	The modifications were made in **page 24 lines 478-481**.
4	Figure 5A – the alignment of the amino acids under the trinucleotide sequence is a little	Thank you for your comment. In response, we have revised Figure 5A to ensure that the amino acids are	We have **updated Figure 5A** in the

	unaligned. It might make the most sense to have the amino acid centered under the trinucleotides, as right now, the amino acid skews off to the next set of nucleotides towards the end.	properly aligned under the trinucleotide sequence. We have centered the amino acids under the corresponding trinucleotides to improve clarity and avoid any misalignment, as suggested. 	revision.
5	Line 585 – Can the authors add age in months or days of the adult zebrafish used, similar to what they reported about 6 months zebrafish being used for RNA extraction?	Thank you for your comment. In response, we have added the information about the age of the adult zebrafish in the manuscript. We clarified that all zebrafish used in the study were six months old, consistent with the age reported for RNA extraction.	The modification was made in **page 32 line 677**.
6	Figures 3F and 4C – The authors report p-values in every other graph but these two, so the authors may want to swap the asterisks for the p-values for these graphs to have consistent reporting.	Thank you for your comment. In response, we have added the p-values to **Figures 3F** and **4C** to ensure consistency in the reporting across all graphs.	We have **updated Figures 3F** and **4C** in the revision.

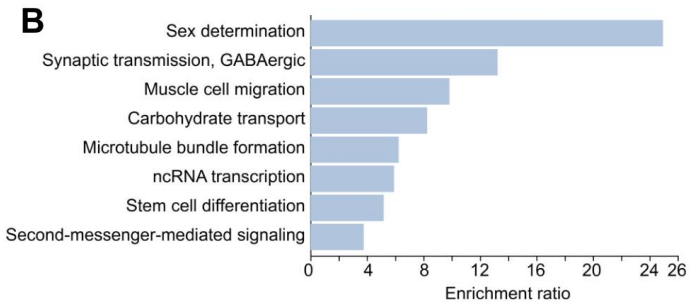

C
Neuro-2a
PC-12
SH-SY5Y
HEK-293T
P = 0.026
P = 0.0102
P = 0.0009
P = 0.004
Relative luciferase activity
PGL3-promoter
rs13072690-A
rs13072690-G

Reviewer #1

	Comment	Author Response	Modification made in manuscript
1	I understand the candidate variants (n=351) have been prioritized using a different approach than McAfee et al. In fact, it is encouraging to see 67 variants are shared between the two studies. While a full meta-analysis may not be feasible, the authors should expand on the STARR-seq results for these overlapping variants, particularly in relation to the 439 variants showing allelic regulatory effects in MPRA. Were any of these 67 variants functionally validated by both STARR-seq and MPRA? If so, do the predicted directions of effect agree across platforms? Please consider summarizing this comparison clearly in Tables S2–S3.	Thank you for your insightful comment. Among the 67 overlapping variants between our study and that of McAfee et al., five SNPs (rs2851447, rs9545047, rs2905426, rs2053079, and rs11713763) were reported by McAfee et al. to exhibit allelic regulatory effects based on MPRA. Of these, only rs2905426 showed significant enhancer activity (classified as an eSNP) in our STARR-seq screen; however, it did not meet our criteria for biased allelic enhancer activity (baaSNP). The remaining four SNPs did not display significant regulatory effects in our assay. To facilitate direct comparison, we have updated Table S2 to include a new column indicating whether each overlapping variant was reported as functionally active by McAfee et al.	The modifications were made **pages 20-21 lines 378-382**. We also **updated Table S2** in the revision.
2	Figure 2 presents valuable data but would benefit from clearer visualization. For Fig. 2c, it is unclear whether this represents a meta-analysis across the four cell lines (e.g., using mpralm). Consider integrating Fig. 2c into Fig. 2d to improve interpretability and to directly indicate which candidate SNPs show biallelic effects.	Thank you for your helpful suggestion. In response, we have revised the figure layout by integrating the original Fig. 2c into Fig. 2d, as recommended. This updated visualization improves interpretability by directly indicating which candidate SNPs exhibit biased allelic effects across the four cell lines. We have also clarified in the figure legend that the data represent results from individual mpralm analyses performed per cell line.	The modifications were made in **page 10 lines 171-182**. We also **updated Figure 2** in the revision.
3	Regarding Fig. 2d, the updated heatmap is a step forward, but interpretability is limited by the current	Thank you for the thoughtful suggestion. In response, we have revised the heatmap in Fig. 2d as recommended. The	The modifications were made in **page 10**

	color scale for logFC. The subtle gradation between -2 and +2 makes it difficult to discern the directionality of effects. I recommend revising the heatmap so that the full cell color reflects logFC (rather than p-value), using a diverging color scale centered at 0 (e.g., red for positive effects, blue for negative, white for neutral, similar Red-Blue colors to volcano plots in Fig. 2b), and capping the scale at -4 and +4 to avoid oversaturation from outliers. Statistical significance can be indicated with asterisks (e.g., p<0.05, *p<0.01, **p<0.001), which will allow for a more intuitive and informative display.	updated version now uses a diverging Red-White-Blue color scale centered at 0 to represent logFC values, with the scale capped at -6 and +6 to improve contrast and prevent oversaturation by outliers. Statistical significance is now indicated using asterisks (*p < 0.05, **p < 0.01, ***p < 0.001), which enhances interpretability by conveying both effect size and significance within a single, intuitive visualization.  	**lines 171-182**. We also **updated Figure 2** in the revision.
4	For all Supplementary Tables listing candidate SNPs, the inclusion of dbSNP “rsID” would be helpful for cross-referencing and clarity.	Thank you for the helpful suggestion. In response, we have updated all relevant Supplementary Tables to include the dbSNP “rsID” for each candidate SNP to facilitate cross-referencing and improve clarity.	The modifications were made in all **Supplementary Tables**.

Reviewer #2

	Comment	Author Response	Modification made in manuscript
1	The authors appear to have misunderstood comment #17 (in their numbering system), which was also briefly mentioned in comment 1. Specifically, they need to rerun webgestalt not using the whole genome as the background set, but instead a background set that appropriately accounts for the fact that the eQTL data they are using comes from brain. Otherwise, it is highly likely that their GO term findings of things like Synapse and GABA are spurious. If they do find that they were spurious, they should remove that claim. I recommend they use as background something like 'all genes that have brain eQTLs' in their source datasets. This should correct for the brain bias. If their finding does go away, they should simply remove those claims from the paper. It is otherwise publishable and this is a minor part of the findings overall.	Thank you for clarifying this important point. We apologize for misunderstanding the reviewer's earlier remark. In response, we have re-performed the Gene Ontology (GO) enrichment analysis using an appropriate background gene set—specifically, all genes with brain eQTLs from our source datasets—to correct for potential brain-specific bias. After this adjustment, the significantly enriched pathways included "sex determination," "synaptic transmission, GABAergic," "muscle cell migration," "carbohydrate transport," "microtubule bundle formation," "ncRNA transcription," "stem cell differentiation," and "second-messenger-mediated signaling." Notably, the "synaptic transmission, GABAergic" pathway remained significant. We have updated the manuscript accordingly, describing this approach and the updated enrichment results. 	The modifications were made in **pages 11 lines 198-203** and **page 30 lines 629-631**. We also **updated Figure 3** in the revision.

2	The authors may have missed the nuance of what I was trying to explain about the use of the word causal and over corrected. They also tried to just replace it with the word 'risk' which sometimes worked and some times did not. E.g., in line 61 or 64 it would be fine to use the word causal. I don't argue that there are not causal variants in the genome. Rather, the point I was trying to make is that MPRAs don't test causality for schizophrenia. Instead, they test functional potential of the variant. Logically, a variant must be functional to be causal, but it is also possible a variant is functional and not causal. So functionality is necessary, but not sufficient, for causality. Thus, when the are talking about causality in general (like line 61 and 64) that is fine. But when they are claiming the discovered the causal variant, that is what needs to be edited more carefully. In some places, the word 'risk' in perfect. In other places, they may want to use the word 'functional' instead of risk. 'Risk' is usually taken to mean ' a variant that has statistical association from human genetics studies of the disorder' So all the SNPs from the GWAS study are 'risk' alleles. But again, this does not guarantee function or causality.	Thank you for the thoughtful clarification. We fully agree with your nuanced distinction between causality, functionality, and genetic risk. In response, we have carefully revised the manuscript to reflect this distinction more accurately. We retained the use of 'causal' in conceptual contexts (e.g., lines 57, 59, and 62) where it refers to the broader goal of identifying biologically meaningful variants. At the same time, we replaced the term 'risk' with 'functional' where appropriate, particularly when describing variants supported by experimental evidence rather than statistical association. These edits help ensure that our language aligns more precisely with the scope and nature of our findings. Thank you again for this important and constructive feedback.	The modifications were made in page **4 lines 57-62** and page **5** lines **86-87**.
3	The authors still talk about constraint in the discussion, but I think they might have removed that part of the analysis from the results?	Thank you for pointing this out. In response to your comment, we have removed the discussion related to selective constraints, as the corresponding analysis was also removed from the results section. This ensures	The modifications were made in **page 23 lines 450-451**.

		consistency between the results and discussion in the revised manuscript.	

Reviewer #1

The authors' responses were satisfactory, and the figure updates have improved the overall readability of the manuscript. I have a few minor comments that should be addressed prior to publication.

	Comment	**Author Response**	**Modification made in manuscript**
1	For Table S2, I attempted to look up five SNPs overlapped with McAfee et al. but was unable to identify them. The overlap should indicate all SNPs intersected with the prior study, whether they were found within active enhancers or not. Additionally, the subset of SNPs located in silencers warrants further investigation for biased allelic activity (which could be considered for future work).	Thank you for your helpful comment, and we apologize for the earlier oversight. We have now updated **Table S2** to include five SNPs that overlapped with McAfee et al., regardless of regulatory classification. We fully agree that this subset warrants further investigation for biased allelic activity. Accordingly, we have now included a statement in the **Discussion** section highlighting this important direction for future work to further refine our understanding of noncoding regulatory mechanisms in schizophrenia.	The modifications were made in page **15** lines **359-361**, we **updated Table S2** in the revision.
2	For Table S3, could the authors clarify the last column labeled "abeSNP"? Should this instead be "baaSNP"?	Thank you for your comment. We appreciate you pointing out the inconsistency. We have corrected the label in **Table S3**—changing "abeSNP" to the appropriate term "baaSNP"—to ensure clarity and consistency with the terminology used throughout the manuscript.	We **updated Table S3** in the revision.