Peer Review File

Manuscript Title: Mapping genomic loci implicates genes and synaptic biology in schizophrenia

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1:
Remarks to the Author:
It's been over a decade since the flagship 2009 Nature schizophrenia paper. This paper still fails to include the samples of African ancestry that are available to the PGC (and Latino/Hispanic samples, I believe). This just isn't acceptable. For goodness sake, in 2020, the PGC shouldn't do this, and Nature shouldn't allow it.

That said, this is a tour de force regarding gene prioritization. With the major changes completed, it will make a tremendous contribution to complex trait genetics broadly and schizophrenia research in particular.

MAJOR (#1 and #2 are mandatory):

1) Add in African ancestry and Latino/Hispanic samples.

2) PLEASE don't set the field back by foreclosing too soon on the synapse. I see that you say that pathology extends beyond the synapse, but you also argue that people must focus on synaptic biology. Your more agnostic data (e.g. Table 7) don't even support this conclusion. Please keep things more open so as not to mislead the field into a needlessly specific area of inquiry. Please revise the framing accordingly.

Note: It's fine and good to make use of SynGO, but just be honest about the fact that additional data resources are available for synapse, and so you made use of them.

3) If relevant, redo MAGMA analyses with the correction recently noted:
(https://www.biorxiv.org/content/10.1101/2020.08.20.260224v1)

MINOR

4) Abstract:
"as likely to explain these associations". REWORD so that it does not sound like all 270 signals are explained by these 190 genes. Wording could be improved more generally for this phrase.

"synaptic organization, differentiation and transmission" Use the Oxford comma!! It is eeded for clarity.

"of physiological relevance" -> "relevant"

P3 Arguably 161, 405 should be the N in your abstract

P8 "explained up to 0.077 of variance in liability (using SNPs with GWAS p-value less than 0.05)" This is less than previous papers. WHY? Shouldn't prediction get better?

Extended data figure 5:

"is represented as a bar for each method." I think you mean "is represented with a horizontal black line."

Extended data figure 3: add exact p-values

P11 Color legend is missing.

P12 "These analyses test the importance of the characteristics used to define gene sets for disease aetiology, but they do not speak to the involvement or not of individual genes." This wording is very confusing. How about:
"These analyses test the importance of gene sets, but are not informative about genes individually."

"aetiopathogenesis". Can you say "causal mechanisms" instead? Or something else? It's up to you but the former sounds a bit silly and detracts from the high rigor of this work.

P12 and on: can the MAGMA analyses be re-done with the correction that the Posthuma group is implementing, if needed: https://www.biorxiv.org/content/10.1101/2020.08.20.260224v1

P13 Extended Data figure 4: Overlay points for MAGMA and LDSC separately.

P15: "with prominent enrichment at the synapse". The "with prominent enrichment" is not warranted based on Table 7. I see synapse among the top GO terms, but this is one term among many. If you want to justify your further focus on synaps and SynGO, find another (data-based) way to do it. Or simply say that you looked at SynGO without additional justification.

Synaptic Location and Function of Prioritized Genes
This section should probably go in the supplementary text.

Referee #2:
Remarks to the Author:
The manuscript describes the largest GWAS study of schizophrenia to date with almost 70,000 cases and over 200,000 controls. 270 associated loci are discovered, adding about 100 novel loci to the previous largest effort. There are a very large number of post GWAS analyses presented, including fine-mapping, pathway analyses, tissue and cell enrichment and Mendelian Randomisation. The major conclusions are that there is an overlap between genes identified in this GWAS and in a companion exome sequencing study, suggesting common and rare variant associations are converging on the same genes. There is also a strong enrichment of genes expressed in CNS neutrons and synaptic pathways.

The manuscript contains a huge amount of analyses and results, but despite that is well written and is relatively easy to follow with much of the detail included in extensive supplementary notes. The GWAS and post GWAS analyses appear to be very well done and are well described. All the "state of the art" post GWAS analyses expected in this type of paper are included, although perhaps some of these analyses would be better described in separate papers (e.g. I'm not sure the MR results add a great deal and could have been performed more comprehensively elsewhere). While on its own this might be considered an incremental GWAS, it does make an appropriate companion piece to the exome sequencing study.

I have only a few other comments.
1. The demonstration that common small effect variants and rare large effect variants converge on the same genes is of interest, although it should be acknowledged that this has been seen in

GWAS for many other traits when compared to monogenic genes for a related condition (e.g. for obesity, Type 2 diabetes, lipids, height, etc).

2. Why was a MAF>1% cut-off used for the analysis? With current imputation panels there should be accurate genotype calls down to 0.1% MAF or less and so it would be good to have a justification why these low frequency variants were excluded.

3. The MR analysis section is relatively weak. The criteria for a trait to be included is unclear, and the GWAS studies included seem out of date. For example, lipid GWAS from 2013 and a height GWAS from 2014. Why weren't the latest GWAS for these traits used? And other traits are missing. For example, why weren't large GWAS of insomnia included, that would seem a relevant trait.

Referee #3:
Remarks to the Author:
This manuscript represents an extremely impressive body of work, it will be a key resource for schizophrenia researchers and will act as an excellent template for the modern reporting of large-scale GWAS and flagship downstream analyses that can be performed from their data. The two major areas of advance provided by this paper, beyond the larger sample size of the GWAS, relate to the fine-mapping prioritisation of loci and the gene set enrichment analyses (the Abstract and Discussion highlight the importance of these two areas to the authors). Therefore, my comments below focus on these two areas first and then cover the remaining analyses of the paper together. A minor structural point is that in both the Abstract and Discussion the fine-mapping analyses are discussed first and then the gene-set enrichment analyses, which doesn’t reflect their order in the paper and so perhaps the authors should consider a restructure of the paper or of the Abstract/Discussion.

A chief question that I have about this work is whether the main results and findings represent a substantial advance over what is already known based on the recent literature in this area (much of which the main authors of this study have generated across several key papers since the PGC-2 publication in 2014).

Gene set enrichments section

I think the results of Extended Data Fig.4 and Figure 2, showing how enrichments across brain regions and cell types only become significant or clearly so when using very large GWAS, are extremely nice to see and are testament to the hard collective work done by the PGC over the years. These results really highlight the importance of large GWAS for understanding aetiology even when effect sizes are very small and their combined predictive power is too low for clinical utility. However, these analyses were performed in great depth in Skene et al 2017, Watanabe et al 2019 and Bryois et al 2020, albeit using a less powerful SCZ GWAS than this PGC wave, and so the advances made over those papers here seem limited. Would readers come to any qualitatively different conclusion from the present results than from those previous papers? (e.g. the importance of excitatory neurons, medium spiny neurons etc were highlighted in the previous papers).

While it is reassuring to observe that the brain regions are important in schizophrenia, it would be helpful to give the readers more of a sense of whether the brain regions and cell types that are most/least significant are consistent with expectations from the broad literature on schizophrenia outside of GWAS, e.g. no enrichment is observed for neural stem cells - is this surprising? Does it have important implications if true? Would pyramidal neurons have been hypothesised as important outside of the GWAS field?

The difference between the analyses of Figure 2 and Supplementary Figure 5 are not made clear in the main text - in the former there are 24 classifications, the latter 265, but are these on different mouse cell lines? For readers without a neuroscience background it will be unclear whether the two sets of results are supportive of each other or not (or else why they differ) without some guidance from the authors. Similarly, in discussion of the 265 results, the authors state that the strongest enrichments are for genes expressed in glutamatergic neurons in the cortex, amygdala and hippocampus but that enrichments were seen for inhibitory and excitatory neurons more widely - which will be misunderstood unless readers know that glutamatergic neurons are a common subset of excitatory neurons, and this summary of the results doesn't quite get across what appears to be a rather overwhelming enrichment of excitatory neurons in the cerebral cortex specifically, which probably deserves more explanation (eg. Is this expected? Does this specific enrichment add to our knowledge of SCZ aetiology?).

Fine-mapping/prioritisation section

I'm afraid that I find the exposition of the procedures to reduce the 270 independent loci to a prioritised set of likely-causal genes via fine-mapping and prioritisation of genes extremely challenging to understand as presented. I realise that it's a difficult task to explain but I think the authors should re-think how best to explain and illustrate this so that readers have any chance of understanding what was done. Here are some examples of questions that arose when I read through this: (in abstract) 'we prioritise 19 genes based on protein-coding or UTR variation, and 130 genes in total' -> how were the 130 genes prioritised? Why 130 and not 114 as shown in Figure 3? (in abstract) 'Fine-mapped candidates were enriched' - are you referring to the 130 genes here or some other number? Line 237/8: 'We then focused on loci.. that were predicted to contain 3 or fewer causal variants (N=217)' -> is this 3 or fewer according to being defined within the credible sets of SNPs from the previous sentence? This would seem to make sense, but then how can there be 217 genes with 3 or fewer causal SNPs based on credible sets, yet the following sentence states 'For 32 loci, the 95% credible set contained 5 or fewer SNPs'; these numbers are clearly very far from each other. Also, within Fig.3 caption 'The combined total of 114 protein-coding genes includes 3 that are outside of the 233 fine-mapped regions, which have expression QTLs..' -> were these other criteria that allowed in these other 3 genes applied systematically and if so why don't they sit alongside the other criteria for prioritisation of genes? Line 265: 'We identified a fine-mapped set comprising genes with at least one credible SNP' - I thought the criteria was 1-3 credible SNPs, so why the switch to 'at least 1' (Fig. 3 suggests > 3.5 are excluded). Why are the 19 (NS+UTR) and 58 (single gene) prioritised genes specifically explained in the main text but not the 30 (Hi-C) and 33 (SMR) genes? (NB. I now see that these are explained later in 'Prioritisation by gene expression' but this seems disconnected, coming after 'Common and Rare Variant..'). Wouldn't it make sense to explain that there were 5 criteria (according to Fig.3, more including the 3 bonus genes) used to prioritise genes from among the 270 loci and just explain each of those in turn (1 <= K <= 3; Hi-C, SMR, etc) within the same subsection devoted to explaining the prioritisation. Why aren't K(1-1.5) and K((1.5-3.5) collapsed in Fig.3 and why isn't the former K(0.5-1.5) instead? Why are the 290 Broad genes highlighted if they're not used? Overall, I think substantial reworking of the text and Figure 3 is required to make this clear.

I'm not sure of the criterion for prioritising genes based on forming credible sets that aggregate to > 95% posterior probability is justified. This means that the inclusion of SNPs is dependent on the PP of other nearby SNPs - wouldn't it make more sense to create a criterion based on the posterior of individual SNPs being causal; e.g. SNPs with PP > 20% are considered sufficiently likely causal (since under the criterion based on aggregation of PP, some SNPs will be included as causal that have very low PP, while others will be excluded despite having high PP).

Figure 4 shows evidence for enrichment of the GWAS signal in different classes of genes relating to SCHEMA genes, those of other outcomes etc, but shows no or little significant evidence (it seems?) of additional signal among the genes within GWAS loci and doesn't even test for the enrichment of

the 114 prioritised genes, which is surprising given that this prioritisation is such a major feature of this paper (at least I think the test here isn't on the 114 genes, but is instead on any genes with 'at least one credible SNP'; the authors should state how many genes were tested in the 'within GWAS loci' set in Fig.4b). Overall, it seems that a well thought-through and complex process was followed to reduce the 270 independent loci to a set of 114 prioritised genes, but I see little evidence from this paper that this prioritisation has worked to produce a more noteworthy set of genes from among the ~ 1500 genes (based on Fig.3) than those of the extended list or eg. compared to instead simply selecting the gene nearest to the top/index SNP at each locus (these comparisons should be performed). I think the readers will expect this issue to be investigated here in order that they know whether they should perform follow-up analyses or experiments focused on the 114 prioritised genes, or whether there might be little loss in expanding their investigations to the full list of ~ 1500 genes. The sense I got from the manuscript as presented is that the evidence that the 114 genes are particularly strong candidates for being causal, above those not prioritised, was rather underwhelming (this may be wrong, but it's the impression I got from the paper) but I think the authors need to be clear about this or else perform more tests to clarify how interesting the prioritised genes are as compared to the non-prioritised genes from within GWS loci.

Similarly, for the claimed convergence between common and rare variant associations comparing findings from this study and that of SCHEMA, this appears to be more a qualitative highlighting of a handful of genes that do show overlap rather than a systematic statistical analyses quantifying the significance and level of overall overlap (4 genes are specifically highlighted in the Discussion summarising this comparison, and a wider list of 8 genes with broader neurodevelopment evidence). We know that genes picked at random are likely to show evidence of GWAS signal given polygenicity, but what is the statistical enrichment of the GWAS signal and SCHEMA-identified genes? Did Fig.4a perform a competitive MAGMA analysis, indicating that the enrichment is over-and-above background polygenic signal? As queried above, was Fig.4b performed on the 114 prioritised genes? I think given that the single largest take-home message from these two companion papers is likely to be that there is a convergence between common and rare variant schizophrenia risk/aetiology, then the authors need to do more to quantify whether there is statistically robust evidence for such a convergence and, if so, how strong it is.

Minor comments on the other sections

I feel that the paper by Pardinas et al 2018, which has considerable overlap with this paper and was the largest SCZ GWAS when published just 2 years ago, has been rather overlooked in the presentation of this paper and so think this needs addressing, at the very least by inclusion in Extended Data Figure 2.

From Extended Data Figure 2 I see that there are 248 loci (independent) loci in the full GWAS, before combining a subset of SNPs with deCODE summary stats. Please state this number of independent loci in this PGC-3 GWAS wave around line 50 where the number of SNPs are stated but not the number of independent loci (many researchers will want to make use of the full GWAS results, not the subset with summary stats with deCODE data too).

h2 estimate - as well as the 0.24 estimate (with SE 0.007) from this wave, I think it would be good for readers to know what the estimates were based on the previous waves (ideally analysed in exactly the same way).

The similarity in the results between males and females seems very interesting given the differences in prevalence, onset etc. How does this compare to the same analyses in other psychiatric or physical disorders? Would be good to comment a little more on this finding I think.

For such a powerful study, the difference in GWS PRS R2 of 2.7% and the P<0.05 R2 of 7.7% seems surprising and suggests that the GWS results from this large GWAS are still just the 'tip of

the iceberg'. I'd like to see the authors comment on this point and the potential implications in terms of the importance of their prioritised genes, which are within this tip-of-iceberg (or is the extra R2 signal coming from refinement of prediction from the same regions?).

Is the AUC of 0.71 based on incremental increase with addition of PRS to the model or is that the AUC with PRS combined with other covariates (I assume the former, but this needs stating).

The authors conclude that the MR results indicate that the genetic correlations here are largely due to horizontal pleiotropy, but depression is close to significant and considering the associations of other anxiety related traits it seems that depression/anxiety may be a causal factor for SCZ in itself based on these results, so given the importance of conclusions from these findings I suggest the authors consider whether slight rewording may be justified here.

The authors highlight schizophrenia as a potential causal risk factor for breast cancer (L166) yet this result has a highly significant P_bias, and so if breast cancer is highlighted then several other outcomes should be.

Author Rebuttals to Initial Comments:

Point by point response:

Reviewers

We appreciate the reviewers' many positive comments about the quality and importance of our paper. Our responses to comments raised are outlined below.

Referee #1 (Remarks to the Author):

It's been over a decade since the flagship 2009 Nature schizophrenia paper. This paper still fails to include the samples of African ancestry that are available to the PGC (and Latino/Hispanic samples, I believe). This just isn't acceptable. For goodness sake, in 2020, the PGC shouldn't do this, and Nature shouldn't allow it. That said, this is a tour de force regarding gene prioritization. With the major changes completed, it will make a tremendous contribution to complex trait genetics broadly and schizophrenia research in particular.

1) *Add in African ancestry and Latino/Hispanic samples.*

We agree that there is a diversity deficit in schizophrenia genomics (and in genomics generally) and that it is important to address this. Please note that the PGC is a consortium of investigators who contribute genotyped samples, and that the recruitment and (in almost all cases) genotyping is not funded or directed by us. Accordingly, sample diversity is not under our direct control. PIs have separately sought funds to promote work and recruitment in under-represented ancestries, and there are finally substantial programmes of recruitment in South Asia, Africa, and South and Central America. These collections are in progress and not available for this paper.

The reviewer's understanding that we excluded substantial available samples of African and Latino ancestry is not correct, although we did not include an admixed African-American sample from the Molecular Genetics of Schizophrenia (MGS) study[1] due to its relatively small sample size. We have now been granted access to summary data from a larger study[2] of African-American and Latino participants (which includes the data from MGS). The

combined samples are still small in the context of our study: admixed African-American 6,152 cases/3,918 controls and admixed Latino 1,234 cases/3,090 controls and do not permit robust ancestry specific discovery work. However, to respond positively to reviewer's suggestion, we have included them in a number of analyses, make reference to this in the main text (lines 106-120) and the supplementary note ("**Exploratory analyses of African-American and Latino samples**"). In the light of strong (and expected) evidence for heterogeneity at the LD level and/or for causal alleles, until more data are available to confirm the observations, we have adopted a conservative approach regarding loci that only become genome-wide significant (GWS) when these samples are included.

2) *PLEASE don't set the field back by foreclosing too soon on the synapse. I see that you say that pathology extends beyond the synapse, but you also argue that people must focus on synaptic biology. Your more agnostic data (e.g. Table 7) don't even support this conclusion. Please keep things more open so as not to mislead the field into a needlessly specific area of inquiry. Please revise the framing accordingly. Note: It's fine and good to make use of SynGO, but just be honest about the fact that additional data resources are available for synapse, and so you made use of them.:*

We agree, and did not intend to suggest that future research should exclusively be directed at the synapse. We do, however, believe that the convergent observations of enrichment for genes annotated to the synapse in GWAS data (present paper), the rare coding variant data from the companion SCHEMA paper, and CNV data cited in the main text, point to the importance in schizophrenia of synaptic pathophysiology. Moreover, that rare mutations in these genes can have large effects implies synaptic biology is a key area. We can see that our use of the word "focus" could suggest other avenues of investigation were invalid. We substantially changed the concluding paragraph (lines 329-358) and hope this is now in line with the reviewer's comment. We have also removed the word 'focus'.

3) If relevant, redo MAGMA analyses with the correction recently noted:
All analyses based on MAGMA have been redone using the corrected version of the program (v1.08). We also updated the gene ontology annotation files. The impact of this was minor.

MINOR

4) Abstract:
"as likely to explain these associations". REWORD so that it does not sound like all 270 signals are explained by these 190 genes. Wording could be improved more generally for this phrase.

We agree and have changed this text (line 13-) to "Using fine-mapping and functional genomic data, we identify 124 genes (109 protein-coding) as likely to underpin associations at some of these loci, including 16 genes with credible causal non-synonymous or UTR variation".

"synaptic organization, differentiation and transmission" Use the Oxford comma!! It is eeded for clarity.

We have included the Oxford comma.

"of physiological relevance" -> "relevant"

We have changed to "biological processes relevant to schizophrenia pathophysiology"

P3 Arguably 161, 405 should be the N in your abstract.

We can see the argument, but the word limit for abstracts allows us to report only on the number of associated loci (N=270) which is based on numbers given in that abstract. However, on lines 47-48, we describe the primary PGC GWAS sample, and the sample size.

"The primary GWAS was performed on 90 cohorts including 67,390 cases and 94,015 controls (161,405 individuals; equivalent in power to 73,189 each of cases and controls)".

P8 "explained up to 0.077 of variance in liability (using SNPs with GWAS p-value less than 0.05)" This is less than previous papers. WHY? Shouldn't prediction get better?

We agree with this intuition, which as we show in the next paragraph, is true for a fair test, and we think it is important to clarify this question in the manuscript. To do so, we have provided a four-page section in the supplementary note ("**Heritability, SNP-based heritability, variance explained in out of sample prediction, and variance explained by genome-wide significant SNPs**"). In short, the 0.077 variance in liability explained is the median across all primary GWAS samples, including the substantial Asian samples. As we show in extended data Figure 2, Asian samples have, on average, lower variance explained than European ancestry samples, a result of the predominance of European ancestry samples in the GWAS. There may also be effects related to ascertainment differences (lines 90-94 of main text, the heritability section of the supplementary note, and Supplementary Figure 11). There may also be as yet unknown ascertainment factors that influence how much variance is explained.

To perform a fair test of progression in successive waves of GWAS studies, we use as a test sample the MGS sample that was used by the International Schizophrenia Consortium (ISC) in the first PRS analysis of schizophrenia in 2009[3], and has been used ever since to benchmark variance explained. We have added the results of those analyses to the main text. (lines 80-87) as follows. "To illustrate change in PRS efficacy with increasing GWAS sample sizes, we consider the MGS cohort (2633 case, 2466 controls) used by the International Schizophrenia Consortium (ISC) and in subsequent studies as a benchmarking cohort. The ISC study used the Nagelkerke's R2 and on that scale the results were: ISC (2009): 0.032, PGC+SWE (2013): 0.06, PGC (2014) 0.184, and in the present study, 0.221, the latter corresponding to 0.104 of variance in liability. A fuller discussion of heritability and polygenic prediction is provided in the Supplementary Note."

Thus, in a like-for-like test, the reviewer's expectation is met.

Extended data figure 5:

"is represented as a bar for each method." I think you mean "is represented with a horizontal black line."

Thanks for the correction. This has been changed in what is now "Extended Data Figure 2".

Extended data figure 3: add exact p-values

We have removed the MR section in the light of comments from the other reviewers.

P12 "These analyses test the importance of the characteristics used to define gene sets for disease aetiology, but they do not speak to the involvement or not of individual genes." This wording is very confusing. How about: "These analyses test the importance of gene sets, but are not informative about genes individually." and *aetiopathogenesis. Can you say "causal mechanisms" instead? Or something else?*.

Agreed and we have adjusted this text, largely along the lines suggested. The relevant phrases are now "To inform hypotheses about causal mechanisms" (line 123), and "These analyses test the importance of the characteristics of gene sets, but they do not address the involvement of individual genes" (lines 125-126). We retained the use of "characteristics" as it is the characteristic that is important (e.g.. brain-expressed, neuron-expressed etc etc).

can the MAGMA analyses be re-done with the correction that the Posthuma group is implementing

Thanks for the suggestion; as noted above, we have done this and the effects are minimal.

P13 Extended Data figure 4: Overlay points for MAGMA and LDSC separately.

We experimented with various formats but found the current to be simplest to grasp at a glance (graphs of this style have appeared in two *Nat Genet* papers and in many presentations and are generally effective). P-values are presented in the data file corresponding to the figures for readers who are particularly interested (now "Extended Data Figure 3").

P15: "with prominent enrichment at the synapse". The "with prominent enrichment" is not warranted based on Table 7. I see synapse among the top GO terms, but this is one term among many. If you want to justify your further focus on synaps and SynGO, find another (data-based) way to do it. Or simply say that you looked at SynGO without additional justification. Review

We have revised the text in this section to reflect the reviewers' concern. We now say "All [enrichments] were relevant to neuronal function including development, differentiation, and synaptic transmission, and involved multiple cellular components including ion channels, synapses, and both axon and dendritic annotations. Using the hierarchical structure of the expert-curated ontology of the SynGO consortium we further examined the synaptic signal" (lines 159-163). See also response to this reviewer's major point 2.

Synaptic Location and Function of Prioritized Genes . This section should probably go in the supplementary text.

The reviewer notes this as a minor point so we hope they do not object to us retaining it. If the papers are accepted, we think this section provides a useful discussion regarding specific prioritised genes and related processes that others will find of high interest, particularly in the context of reading the present paper along with SCHEMA.

Referee #2 (Remarks to the Author):

General comment: All the "state of the art" post GWAS analyses expected in this type of paper are included, although perhaps some of these analyses would be better described in separate papers (e.g. I'm not sure the MR results add a great deal and could have been performed more comprehensively elsewhere).

We are pleased the reviewer appreciates the range and quality of our analyses, and have removed the MR section as suggested.

1. *The demonstration that common small effect variants and rare large effect variants converge on the same genes is of interest, although it should be acknowledged that this has been seen in GWAS for many other traits when compared to monogenic genes for a related condition (e.g. for obesity, Type 2 diabetes, lipids, height, etc).*

We agree there are instances where a GWAS signal has been noted over a gene implicated by rare mutations. We have now cited two papers that to our knowledge represent the most systematic approaches to this phenomenon (line 299).

2. *Why was a MAF>1% cut-off used for the analysis? With current imputation panels there should be accurate genotype calls down to 0.1% MAF or less and so it would be good to have a justification why these low frequency variants were excluded.*

We recognise a MAF threshold of 0.1% is now often used in biobank studies. This exploits the fact that large cohorts can provide adequate minor allele counts (MAC) for analyses. For example in the UK Biobank of around 500,000 people, a MAC of 1000 is expected at a variant site where MAF is 0.1%. In contrast, at sites with MAF of 0.1%, many of the samples contributing to our meta-analysis would have a MAC of 0 or 1, and even our largest contributing cohort (N ~5,000) would have a MAC of only 10. Our study is then underpowered for SNPs with low MAF, and this is reflected by the observation that in the present study, we see no index associations where MAF is <2%, and only 8 index associations with MAF < 5%.

To respond more fully to the reviewer, we re-ran the analyses with a MAF cut off of 0.5%, and find one additional GWS association; however, this SNP is poorly imputed (INFO score 0.239) and is in the already well-known MHC region. Thus, lowering the MAF threshold did not add anything meaningful to the study.

3. *The MR analysis section is relatively weak. The criteria for a trait to be included is unclear, and the GWAS studies included seem out of date. For example, lipid GWAS from 2013 and a height GWAS from 2014. Why weren't the latest GWAS for these traits used? And other traits are missing. For example, why weren't large GWAS of insomnia included, that would seem a relevant trait.*

In light of the comments about this section from all reviewers, we have removed this section.

Referee #3 (Remarks to the Author):

This manuscript represents an extremely impressive body of work, it will be a key resource

for schizophrenia researchers and will act as an excellent template for the modern reporting of large-scale GWAS and flagship downstream analyses that can be performed from their data. The two major areas of advance provided by this paper, beyond the larger sample size of the GWAS, relate to the fine-mapping prioritisation of loci and the gene set enrichment analyses (the Abstract and Discussion highlight the importance of these two areas to the authors). Therefore, my comments below focus on these two areas first and then cover the remaining analyses of the paper together. A minor structural point is that in both the Abstract and Discussion the fine-mapping analyses are discussed first and then the gene-set enrichment analyses, which doesn't reflect their order in the paper and so perhaps the authors should consider a restructure of the paper or of the Abstract/Discussion.

As suggested, we have revised the abstract to reflect the order in the paper.

A chief question that I have about this work is whether the main results and findings represent a substantial advance over what is already known based on the recent literature in this area (much of which the main authors of this study have generated across several key papers since the PGC-2 publication in 2014).

We address this comment below.

Gene set enrichments section

I think the results of Extended Data Fig.4 and Figure 2, showing how enrichments across brain regions and cell types only become significant or clearly so when using very large GWAS, are extremely nice to see and are testament to the hard collective work done by the PGC over the years. These results really highlight the importance of large GWAS for understanding aetiology even when effect sizes are very small and their combined predictive power is too low for clinical utility. However, these analyses were performed in great depth in Skene et al 2017, Watanabe et al 2019 and Bryois et al 2020, albeit using a less powerful SCZ GWAS than this PGC wave, and so the advances made over those papers here seem limited. Would readers come to any qualitatively different conclusion from the present results than from those previous papers? (e.g. the importance of excitatory neurons, medium spiny neurons etc were highlighted in the previous papers).

First, please note that we have revised these figures in line with the newer version of MAGMA as suggested. We thank the reviewer for recognizing the work of the PGC that underpins these and other analyses within the paper. The reviewer notes the importance of sample size. We agree and argue it is important to test major previous findings against the present dataset, not merely for replication purposes but because as power of GWAS increases, this might in principle implicate more tissues and cells types, as happened for example between PGC1 versus PGC2 (now Extended Data Figure 3). Importantly, at the tissue level, we show here that this does not happen between PGC2 and PGC3; rather the evidence for previously enriched tissues becomes stronger but there is no general increase in enrichment signals in additional tissues. This consistency is important because it suggests that the enrichments for genes expressed in brain (over other tissues) is unlikely to meaningfully change as sample sizes increase further. Regarding the single cell analyses, similar considerations apply although here, more classes of excitatory and inhibitory neurons show evidence for association, but additional cell types do not. These findings are important as the idea that schizophrenia is essentially a brain disorder of predominantly perturbed neuronal function is not universally accepted even now. Moreover, since the large literature on the

aetiology of schizophrenia is replete with non-replication, we think this increases the value of our results here that both confirm and deepen the earlier findings. We hope the reviewer agrees it is important that we present the findings.

While it is reassuring to observe that the brain regions are important in schizophrenia, it would be helpful to give the readers more of a sense of whether the brain regions and cell types that are most/least significant are consistent with expectations from the broad literature on schizophrenia outside of GWAS, e.g. no enrichment is observed for neural stem cells - is this surprising? Does it have important implications if true? Would pyramidal neurons have been hypothesised as important outside of the GWAS field?

As the reviewer alludes to, there is an extensive literature on schizophrenia neuroimaging, neuropathology, and pathophysiology. As a body, this literature has been used to implicate a large number of brain regions, cortical and non-cortical. Different researchers have also used the data to highlight the importance of grey over white matter, and vice versa. Among data of highly variable quality, probably the best source from which to make anatomical predictions is the recent review from the ENIGMA consortium who described white matter abnormalities in schizophrenia across "the entire white matter skeleton" of the brain[4]**,** thinner cerebral cortex (both hemispheres, globally), and smaller cerebral cortex surface area (both hemispheres, globally). They have also reported **changes in** total intracranial volume, and the **volumes of** hippocampus, amygdala, thalamus, nucleus accumbens, pallidum, and lateral ventricles. Regarding cell types, there is an extensive literature supporting the involvement of almost every type of CNS cell in schizophrenia (which, further underscores the importance of showing the stability of the results as discussed in response to the previous point) but there is as yet no meta-resource that is an authoritative equivalent to ENIGMA. That literature can be argued to highlight roles for pyramidal excitatory neurons and inhibitory neurons of various types (e.g. parvalbumin, chandelier and others), but it can be, and has been, argued that it implicates oligodendrocytes, astrocytes, and microglia. Consequently, we had no strong consensus expectations regarding brain regions and specific cell types. Rather as stated (lines 152-154), we think our data are best interpreted as "Overall, the findings across all the datasets are consistent with the hypothesis that schizophrenia is primarily a disorder of neuronal function, but do not suggest that pathology is restricted to a circumscribed brain region".We think is also broadly consistent with the imaging findings from ENIGMA in pointing to a distributed anatomy and neuropathology of schizophrenia.

Regarding the point about neural stem cells, we had no expectation that gene expression that is *selective* for this cell type *relative to other brain cell types* should be enriched for association. We also do not think there are implications that elevate this absence of association above the many other negative findings in these analyses. We certainly would not interpret this as meaning that early brain development is of no importance in schizophrenia.

The difference between the analyses of Figure 2 and Supplementary Figure 5 are not made clear in the main text - in the former there are 24 classifications, the latter 265, but are these on different mouse cell lines?

We apologise for a lack of clarity. The 24 cell line classifications were those used in Skene 2017, and correspond to a small number of mouse brain regions and cell type enrichments. Since then, Linarsson and colleagues[5] did a more comprehensive survey of the mouse nervous system including 265 cell types analysed here. These are different cell line resources. We make this clear in the main text. This is now flagged up on line 144 where we say

"Supportive results were also obtained using a different dataset of 265 cell types" and note this also in "**Online Methods"**.

For readers without a neuroscience background it will be unclear whether the two sets of results are supportive of each other or not (or else why they differ) without some guidance from the authors.

We have adjusted the text of the relevant section to make it clear the data are supportive as follows.

"Supportive results were also obtained using a different dataset of 265 cell types in the mouse central and peripheral nervous system [24]. Very strong enrichments were again seen for genes expressed in excitatory glutamatergic neurons of the cortex (especially the deep layers) and hippocampus but also the amygdala (**Supplementary Figure 3**). Highly significant enrichments were also seen for other neuronal populations, including as above, inhibitory medium spiny neurones in striatum," (lines 147-149).

Similarly, in discussion of the 265 results, the authors state that the strongest enrichments are for genes expressed in glutamatergic neurons in the cortex, amygdala and hippocampus but that enrichments were seen for inhibitory and excitatory neurons more widely - which will be misunderstood unless readers know that glutamatergic neurons are a common subset of excitatory neurons.

We revised the structure of this paragraph to address the reviewer's previous comment, but note the text now says "we found associations were enriched in genes with high expression in excitatory glutamatergic neurons from cerebral cortex and hippocampus" (line 135-137) and "excitatory glutamatergic pyramidal neurons from the cortex and hippocampus" (line 140).

This summary of the results doesn't quite get across what appears to be a rather overwhelming enrichment of excitatory neurons in the cerebral cortex specifically, which probably deserves more explanation (eg. Is this expected? Does this specific enrichment add to our knowledge of SCZ aetiology?).

We agree there are robust enrichments for genes expressed in cortical neurons, but there are also strong enrichments for neurons in hippocampus and striatum. As noted, in responses above, and in the main text, we think the results are most conservatively interpreted as consistent with a diffuse neuropathology model rather than one that is selective for the cortex. We now say "Very strong enrichments were again seen for genes expressed in excitatory glutamatergic neurons of the cortex (especially the deep layers)" (line 145-146) but given the other findings, we also say "Overall, the findings across all the datasets are consistent with the hypothesis that schizophrenia is primarily a disorder of neuronal function, but do not suggest that pathology is restricted to a circumscribed brain region" (line 152-154).

Fine-mapping/prioritisation section

I'm afraid that I find the exposition of the procedures to reduce the 270 independent loci to a prioritised set of likely-causal genes via fine-mapping and prioritisation of genes extremely challenging to understand as presented. I realise that it's a difficult task to explain but I think

the authors should re-think how best to explain and illustrate this so that readers have any chance of understanding what was done.

We can see from comments below that there places where our explanations were unclear – we apologise and have taken the reviewer's comments on board to improve the manuscript. We have adjusted the text as indicated in the response that follows, and introduced figure 3A. We hope it is now clearer.

'we prioritise 19 genes based on protein-coding or UTR variation, and 130 genes in total' -> how were the 130 genes prioritised? Why 130 and not 114 as shown in Figure 3? (in abstract)

Please note that the numbers have changed slightly as a result of our responses to the review process, and also some improvements to the fine-mapping algorithms that have been introduced since our original submission.

In our original submission, the number 114 in the figure referred to protein coding genes, while the 130 included non-coding genes. We agree the text in the abstract and elsewhere was confusing. For the abstract, we have changed this to "Using fine-mapping and functional genomic data, we identify 124 genes (109 protein-coding) as likely to underpin associations at some of these loci" (lines 13-15). At all other points in the manuscript, we clarify when we are restricting to protein coding genes.

'Fine-mapped candidates were enriched' - are you referring to the 130 genes here or some other number? Line 237/8.

We accept it was not clear what genes we referring to in the legend or the text. Moreover, a clarifying sentence was mistakenly omitted. In response, we have extensively revised the text and the legend. Please see the rewritten text on lines 255-259.

We then focused on loci.. that were predicted to contain 3 or fewer causal variants (N=217)' -> is this 3 or fewer according to being defined within the credible sets of SNPs from the previous sentence? This would seem to make sense, but then how can there be 217 genes with 3 or fewer causal SNPs based on credible sets, yet the following sentence states 'For 32 loci, the 95% credible set contained 5 or fewer SNPs'; these numbers are clearly very far from each other.

Apologies: we can see that the order of sentences in the main text leads to the reviewer's interpretation. In brief, FINEMAP models how the SNPs within a GWAS locus, accounting for LD, generate an association signal matching the observed data. From all the models evaluated, it then derives the best model for the number of causal variants at a locus. In the original submission, we had 217 loci where the best model was 3 or fewer causal variants, and these were the loci we focussed on. Note we have resolved additional loci taking this number to 228. We then take those loci and look in greater detail. For each locus, we derive the smallest number of SNPs where the sum of the posterior probabilities (PP) for being one of the causal variants is 95% of the total PP. This is the 95% credible set which therefore corresponds to the shortest list of SNPs required to achieve 95% confidence of including the causal variants at the locus. Within our data there are then 31 loci where 5 or fewer SNPs together provide 95% confidence of including the causal variants. We have reworded the text (lines 178-183) to say

"we used FINEMAP[26] to infer the most likely number of causal variants. For loci predicted to contain 3 or fewer causal variants **(N=228; Figure 3; Supplementary Tables 10a, 10b),** we extracted the FINEMAP assigned SNP posterior probabilities (PP) of being causal for every SNP across the locus, and constructed credible sets of SNPs that cumulatively capture 95% of the regional PP (**Supplementary Note**)".

We also provide the detail in the supplementary note FINEMAP section (lines 372-410).

Also, within Fig.3 caption 'The combined total of 114 protein-coding genes includes 3 that are outside of the 233 fine-mapped regions, which have expression QTLs..' -> were these other criteria that allowed in these other 3 genes applied systematically and if so why don't they sit alongside the other criteria for prioritisation of genes?

We did not introduce criteria beyond those described in the paper which we systematically applied throughout. SMR combines data from eQTLs and GWAS, and as a result, can identify associations involving SNPs that have small P-values that do not quite reach genome wide significance in the GWAS alone. While this is one of the benefits of SMR (and other forms of transcriptome wide association study), the focus of our paper is mapping significant loci, and so those genes should have been excluded. We have now done so. We are grateful to the reviewer whose comment here alerted us to this issue.

Line 265: 'We identified a fine-mapped set comprising genes with at least one credible SNP' - I thought the criteria was 1-3 credible SNPs, so why the switch to 'at least 1' (Fig. 3 suggests > 3.5 are excluded).

See response to reviewer 3 above where we explain the difference between the estimate of number of causal SNPs at a locus and the number of credible SNPs. We hope it is now clear that the fine-mapped set of genes referred to is that subset of the genes whose boundaries contain at least one credible SNP, with the extra criterion that they map to loci where the expected number of causal variants is 3 or less. Please also note that FINEMAP calculates the probabilities for between 1-5 causal SNPs at a locus. However, how many SNPs are expected to be causal after evaluating these probabilities (K) is a value on a continuous scale while the actual number of causal SNPs (k) must be discrete. We provide the distinction between these two statistics, which stem from the fine-mapping literature, and their mathematical definition in the supplementary note. A score of K=3.5 means models with 3 or fewer causal SNPs are no more probable than a model with 4 causal SNPs. Hence, we use a cut off in figure 3 of $K \leq 3.5$, meaning the best model has a maximum of 3 causal variants.

Why are the 19 (NS+UTR) and 58 (single gene) prioritised genes specifically explained in the main text but not the 30 (Hi-C) and 33 (SMR) genes? (NB. I now see that these are explained later in 'Prioritisation by gene expression' but this seems disconnected, coming after 'Common and Rare Variant'). Wouldn't it make sense to explain that there were 5 criteria (according to Fig.3, more including the 3 bonus genes) used to prioritise genes from among the 270 loci and just explain each of those in turn (1 <= K <= 3; Hi-C, SMR, etc) within the same subsection devoted to explaining the prioritisation

This is a good suggestion and we have implemented it.

Why aren't K(1-1.5) and K((1.5-3.5) collapsed in Fig.3 and why isn't the former K(0.5-1.5) instead?

An association requires a minimum of 1 causal variant at each locus, so the lower boundary is k=1. We divide K into 1-1.5 and 1.5-3.5 to distinguish between loci where our best estimate is a single causal SNP, and others where the best model is more complex. We think this may be relevant for people wishing to dig deeper into individual loci and this distinction has also been made before in display items from other complex disorder fine-mapping studies (diabetes, blood cell traits and others), but are prepared to adjust the figure if this is a requirement for publication.

Why are the 290 Broad genes highlighted if they're not used?

As we discuss in a later response to this review, the broad set is enriched for schizophrenia relevant genes compared with other genes in the loci. For that reason, we think they are worth mentioning and we provide this full list of genes, along with the PP of all SNPs (Supplementary tables 10, 12 and 13). We do so to allow others to adopt less conservative criteria for specific loci or genes that may be of particular interest to them.

Overall, I think substantial reworking of the text and Figure 3 is required to make this clear.

We agree, and hope the revisions, the additional display item, and alteration to figure 4, make things easier to follow.

I'm not sure of the criterion for prioritising genes based on forming credible sets that aggregate to > 95% posterior probability is justified. This means that the inclusion of SNPs is dependent on the PP of other nearby SNPs - wouldn't it make more sense to create a criterion based on the posterior of individual SNPs being causal; e.g. SNPs with PP > 20% are considered sufficiently likely causal (since under the criterion based on aggregation of PP, some SNPs will be included as causal that have very low PP, while others will be excluded despite having high PP).

From the perspective of a researcher interested in modelling high priority variants, the focus would be on individual SNPs with high PP, as suggested by the reviewer. This is why we drew attention to the "*31 loci [for which] the 95% credible set contained 5 or fewer SNPs"* and on other SNPs with very high PP. The purpose of the aggregation criteria is quite different; it is not to prioritize SNPs but to prioritise genes where there is a very high probability that a true causal variant maps within their boundaries (see also this reviewers' comment about "nearest gene" strategy below). This occurs when the aggregate 95% credible set maps within the boundaries of the gene, regardless of whether that probability is focused on a single or on multiple SNPs. Consider the situation where a single SNP with a PP of 20% maps within one gene and where the remaining PP (cumulative PP = 75%) is distributed across a large number of SNPs, all of which map within another gene. The probability that the causal variant maps within the gene containing the single high PP SNP is then much less than the probability that it maps within the gene with the large number of SNPs in the credible set (20% versus 75%). Please also note our analyses discussed below which show the relationship between the PP captured within a gene and mutation intolerance.

Figure 4 shows evidence for enrichment of the GWAS signal in different classes of genes relating to SCHEMA genes, those of other outcomes etc, but shows no or little significant

evidence (it seems?) of additional signal among the genes within GWAS loci and doesn't even test for the enrichment of the 114 prioritised genes, which is surprising given that this prioritisation is such a major feature of this paper (at least I think the test here isn't on the 114 genes, but is instead on any genes with 'at least one credible SNP'; the authors should state how many genes were tested in the 'within GWAS loci' set in Fig.4b).

We are pleased the reviewer considered our approach well thought out, if somewhat complex, and hope to have provided more clarity as noted above, in the revised text, and the supplementary note, with further elaboration below.

First, we apologise for not inserting the gene numbers and have corrected this in the legend. Second, the reviewer is correct that the focus of figure 4 is genes highlighted by FINEMAP. This is because SMR is widely accepted as an approach for prioritising credible causal genes by linking GWAS associations to functional variation, but FINEMAP has not been widely used for this purpose. We therefore sought to provide support for our FINEMAP approach. From earlier work[6], and reconfirmed here in figure 4A, we know that schizophrenia-associated genes are enriched for genes that are intolerant to loss of function mutations (LOF intolerant). We then expect that if the property "containing at least one FINEMAP credible SNP" (the definition of the broad gene set referred to by the reviewer) identifies genes that are more likely to be causally related to schizophrenia than the rest of the genes at the loci, the broad FINEMAP set should be enriched for LOF intolerant genes compared with the remaining genes in the associated loci. We show this in figure 4B.

Overall, it seems that a well thought-through and complex process was followed to reduce the 270 independent loci to a set of 114 prioritised genes, but I see little evidence from this paper that this prioritisation has worked to produce a more noteworthy set of genes from among the ~ 1500 genes (based on Fig.3) than those of the extended list or eg. compared to instead simply selecting the gene nearest to the top/index SNP at each locus (these comparisons should be performed).

As noted in the response to the previous comment, we presented evidence that the broad list of genes referred to by the reviewer and highlighted by FINEMAP is enriched for susceptibility genes compared to other genes at the loci. The purpose of the FINEMAP gene *prioritisation* we undertook was then to identify a higher confidence set to facilitate biological interpretation and time-consuming lab-based follow up studies. If our additional criteria plucks random genes from the extended list as the reviewer speculates, we predict that the prioritised set would not differ from that list with respect to the loss of function metric. This is not what we found. As noted in the "**Mutation intolerance metrics in fine-mapped genes**" section of the Supplementary Note, and Supplementary Table 14ab, 43.5% of the prioritised FINEMAP genes are mutation intolerant compared with 31% of the broad set of genes and 15.3% of the other genes within the loci. Prioritising genes based on exonic variation is fairly standard, but our approach to prioritising a gene based on the amount of posterior probability captured by SNPs mapping within its boundaries is novel. We therefore further explored this property. In the same section of the supplementary note, we show that the mutation intolerance metric for genes in the extended set is correlated with the posterior probability captured by SNPs in that gene. We also show that the amount of PP captured by a gene is a significant predictor (beta 1.5, $P=2.4x10^{-10}$) of the categorical definition of a gene as mutation intolerant (LOEUF≤0.35). Thus our approach of combining the PP of all SNPs within a gene to derive a combined posterior probability metric indexes a known property of schizophrenia associated genes, supporting its use in prioritising genes.

We appreciate that the volume of supplementary material in our paper made it difficult to find these analyses and have now highlighted their existence in the main text. We also include a statement that:
"...These protein-coding genes had a greater than 3-fold enrichment for loss of function intolerance compared with other protein-coding genes within the loci that were not tagged by credible SNPs **(Supplementary Table 14; Supplementary Note),** supporting our strategy to delimit credible causal genes."

The reviewer asks whether our prioritised set is noteworthy compared with simply choosing the nearest genes to an index SNP. Of nearest genes, 31.4% are mutation intolerant compared with 43.6% of the prioritised list. Please also note that the premise that proximity is a property to exploit in identifying credible causal genes for GWAS hits is one we agree with, and underpins our rationale for both the extended and the prioritised FINEMAP lists, but we define proximity in terms of credible causal SNPs rather than index SNPs. We argue our approach has advantages. FINEMAP generates a probabilistic measure of causality for individual SNPs, providing information that can be acted on for modelling variation. It allows for scenarios where credibly causal non-synonymous or UTR variants exist in genes that are not within the nearest gene, and allows us to consider genes where there credible SNPs with functional annotations for example those in Hi-C loops linked to the promoters of genes, an approach supported by a recent pre-review publication[7]. It also allows us to prioritise genes where association signals are concentrated in only one gene over those where signals are distributed across many genes. Moreover, it allows us to do so in a quantitative way. We have chosen to be conservative, giving priority to genes where every credible causal SNP maps within the boundaries of a single gene. but we provide the posterior probabilities for every SNP (supplementary table 10) which will allow others to set criteria at more or less stringent levels, according to their needs.

For the claimed convergence between common and rare variant associations comparing findings from this study and that of SCHEMA, this appears to be more a qualitative highlighting of a handful of genes that do show overlap rather than a systematic statistical analyses quantifying the significance and level of overall overlap (4 genes are specifically highlighted in the Discussion summarising this comparison, and a wider list of 8 genes with broader neurodevelopment evidence). We know that genes picked at random are likely to show evidence of GWAS signal given polygenicity, but what is the statistical enrichment of the GWAS signal and SCHEMA-identified genes? Did Fig.4a perform a competitive MAGMA analysis, indicating that the enrichment is over-and-above background polygenic signal? As queried above, was Fig.4b performed on the 114 prioritised genes?

We agree that genes picked at random are likely to show evidence of GWAS signal given polygenicity and confirm that throughout this manuscript we adopt competitive tests that address this problem. The MAGMA analysis in Figure 4a tests whether the plotted sets of genes (including SCHEMA genes) are more enriched for GWAS signal than the background polygenic signal. In Figure 4b (and the extended analyses in the supplementary material referred to above), the test compares the enrichment for SCHEMA genes in the broad FINEMAP set versus the background of all other protein coding genes (allowing for the polygenic problem), and more stringently, versus other protein coding genes at the associated loci. Convergence cannot therefore be attributed to the background polygenicity of the disorder. It is only after demonstrating the general GWAS and the FINEMAP convergence, that we then highlight the genes at the extreme end of convergence, that is those that are not

merely enriched for association, but that are associated at genome wide significance in both studies. The observation that we identify 20% (N=2) of the SCHEMA GWS genes in the 0.5% (N=109) of all protein coding genes in our priority set is clearly much higher than any null expectation (Friths test OR 114.4, 95% C.I. 7.58-16499; p=0.0009) but the numbers are too small for robust statistical testing or effect size estimation. Please also note that in the companion paper, SCHEMA with access to all the individual mutations, addresses this issue from a different perspective, and shows the burden of rare variation in genes we prioritise is significantly higher in cases than in controls.

I think given that the single largest take-home message from these two companion papers is likely to be that there is a convergence between common and rare variant schizophrenia risk/aetiology, then the authors need to do more to quantify whether there is statistically robust evidence for such a convergence and, if so, how strong it is.

We believe we have done exactly that as described in response to the point above, and hope that the reviewer also agrees.

Minor comments on the other sections

The paper by Pardinas has been rather overlooked.. this needs addressing, at the very least by inclusion in Extended Data Figure 2.

We have removed ED figure 2. We did not repeat most of the analyses presented in the Pardiñas paper, which is why we did not refer extensively to it.

there are 248 loci (independent) loci in the full GWAS, before combining a subset of SNPs with deCODE summary stats. Please state this number of independent loci in this PGC-3 GWAS wave around line 50 where the number of SNPs are stated

We have done as requested (line 53) and all are reported in ST1.

h2 estimate - as well as the 0.24 estimate (with SE 0.007) from this wave, I think it would be good for readers to know what the estimates were based on the previous waves (ideally analysed in exactly the same way).

We agree. Competition for space, and the need to put heritability estimation into the context of developments in methodology, led us to produce a 4 page section of the supplementary note entitled "**Heritability, SNP-based heritability, variance explained in out of sample prediction, and variance explained by genome-wide significant SNPs**". For the reviewer's convenience we extract a section of that discussion: "One method to estimate method to estimate h^2_{SNP} from GWAS summary statistics is SBayesS; when applied to PGC-SCZ data, using the GERA LD reference, the estimates are PGC2 0.21 (s.e. 0.006), PGC3 (EUR) 0.24 (0.007), PGC3 (all ancestries) 0.24 (0.007)". We also include a discussion of alternative methodologies that have been used to estimate SNP-based heritability to allow for a fair comparison of how the field has developed (see later point by this reviewer).

The similarity in the results between males and females seems very interesting given the differences in prevalence, onset etc. How does this compare to the same analyses in other psychiatric or physical disorders? Would be good to comment a little more on this finding I

think.

It is widely believed that schizophrenia has a greater lifetime prevalence and earlier onset in males, but both of these features have been called into question by a recent large study of the global epidemiology of schizophrenia, which also suggests that prevalence may even be less in males than females in some countries, notably China[8]. The main purpose of our analysis was to settle the question about possible sex differences in the common variant architecture of schizophrenia that people frequently ask. Having shown that there is no difference, we do not really have a great deal to add beyond the comment we have made. In contrast, there is little doubt that many other mental disorders such as major depression, anorexia, and ADHD, show marked sex differences in rates of clinical diagnoses, and there is a working group of the PGC trying to look at these issues across all disorders. Currently, similarly well-powered analyses are not available for most psychiatric disorders, and the schizophrenia group does not have access to the data to conduct them. Given this and the questionable nature of sex differences in schizophrenia, we hope this reviewer understands that we chose not to address this issue further.

For such a powerful study, the difference in GWS PRS R2 of 2.7% and the P<0.05 R2 of 7.7% seems surprising and suggests that the GWS results from this large GWAS are still just the 'tip of the iceberg'. I'd like to see the authors comment on this point and the potential implications in terms of the importance of their prioritised genes, which are within this tip-of-iceberg (or is the extra R2 signal coming from refinement of prediction from the same regions?).

It is a given that a GWAS even of the current size is underpowered to capture the majority of common variation at genome wide significance (GWS), much less refine the mapping of the majority of the genes involved. We are therefore not surprised that a minority of the R2 is accessible by GWS SNPs. This expectation was clearly documented in the simulations presented in the seminal paper of the International Schizophrenia Consortium (2009) which introduced PRS to the human genetics community[3]. It follows that since we detect a minority of all the signal possible, and then we prioritise genes at less than 50% of loci, that set of genes must be minority of all causal genes. However, the importance of prioritisation is not that this set of genes explains a great deal of heritability. To clarify what is implied by prioritised, we now say (lines 168-170) "To facilitate biological interpretation and laboratory follow up, we sought to prioritise specific variants and genes most likely to explain associations at the loci using a combination of fine-mapping, transcriptomic analysis, and functional genomic annotations (Figure 3)".

In response to the question of whether increased R2 is coming from refinement of prediction, since PRS is based on vigorous LD pruning, the majority of the gap between the R2 at a genome wide level and at a GWS level is expected to come from independent signals.

Is the AUC of 0.71 based on incremental increase with addition of PRS to the model or is that the AUC with PRS combined with other covariates (I assume the former, but this needs stating).

The AUC quoted in the main text does not take account of the covariates, but as we note in the heritability section of the Supplementary Note, that is not expected to make much

difference in theory or empirically. Full discussion is in the note (lines 262-173), but we extract some key points.

"Alternative evaluation statistics include the AUC statistic which can be interpreted as the probability that a case ranks higher than a control when a randomly selected case and control are compared on their PRS, and hence is not dependent on proportion of cases in the sample. Again, benchmarking against the MGS, we obtain (PGC2: 0.72, PGC3: 0.75)". However, we go on to note that "AUC is calculated without including ancestry principal components in the model, which are included in the models used to calculate all other statistics. However, the impact is likely small; converting the liability scale variance (which *has* been estimated after accounting for the other covariates) to AUC using normal distribution theory generates AUC for MGS of PGC2: 0.71, PGC3:0.74."

The authors conclude that the MR results indicate that the genetic correlations here are largely due to horizontal pleiotropy, but depression is close to significant and considering the associations of other anxiety related traits it seems that depression/anxiety may be a causal factor for SCZ in itself based on these results, so given the importance of conclusions from these findings I suggest the authors consider whether slight rewording may be justified here. The authors highlight schizophrenia as a potential causal risk factor for breast cancer (L166) yet this result has a highly significant P_bias, and so if breast cancer is highlighted then several other outcomes should be.

We have removed the MR section from the manuscript in response to the comments from all reviewers.

References

1. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753–757 (2009).
2. Bigdeli, T. B. *et al.* Contributions of common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry. *Mol. Psychiatry* **25**, 2455–2467 (2020).
3. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
4. Thompson, P. M. *et al.* ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl. Psychiatry* **10**, 1–28 (2020).
5. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).
6. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
7. Forgetta, V. *et al.* An Effector Index to Predict Causal Genes at GWAS Loci. *bioRxiv* 2020.06.28.171561 (2021) doi:10.1101/2020.06.28.171561.
8. Charlson, F. J. *et al.* Global epidemiology and burden of schizophrenia: Findings from the global burden of disease study 2016. *Schizophr. Bull.* **44**, 1195–1203 (2018).

Reviewer Reports on the First Revision:

Referee #1:
Remarks to the Author:
Regarding the revised manuscript, I have only two comments. One substantive and the other minor.

1) I have to say it again: It seems unfortunate that all African-American and Latino samples are omitted from the main analyses.

The new addition of African-American and Latino samples is useful (lines 106-120 and Supplementary Note), but I wish that more had been done. Instead of reporting on, and ideally including the small numbers of individuals with African ancestry (and other ancestries) from the PGC cohorts, the authors relied on recently published GWASs of African-American and Latino samples (Bigdeli 2020). I fully agree with their use of findings from Bigdeli et al, but they still "threw away" the small numbers of underrepresented minorities in the PGC cohorts. They argue that the numbers are too small (and indeed this is widely known in the research community), but it's not a great justification.

Put another way: the authors argue that N=6,152 African-American cases is too small to include in the trans-ancestry meta-analyses. But an even smaller number of cases was deemed worthy when the subjects were white (ISC 2009 Nature case N=3,322). I realize that this might seem an unfair comparison, but in important ways it is not. Why should >6,000 African-American cases, with far better genomic coverage than in 2009, be thrown away? Further, this number would be somewhat larger still if other miscellaneous PGC cases were included. These same arguments apply to other ancestries.

In the beginning of the revised manuscript, the authors describe data availability for European and Asian ancestry samples only. I don't think this sets a good example. At a very minimum, I think they should make it clear how researchers can access all summary stats (i.e. for multiple ancestries and for a full trans-ancestry meta-analysis).

2) Typos on lines:
23: possible "prioritized" instead of "priority"?
186: some issue here: "186 that contains 25 genes but is the only a single credible and maps within ATP2A2."
196 "Figure 3A" is not bold.

Referee #2:
Remarks to the Author:
The authors have addressed my comments well.

Referee #3:
Remarks to the Author:
Thanks to the authors for their detailed responses to my original review. I have further comments below, which are all essentially minor.

I think make clear in the abstract that you also prioritise variants, as well as genes, here. I think your statement at the end of the Introduction "..and analysed the findings to prioritise variants,

genes and biological processes that contribute to pathogenesis" is a nice summary of the impact of the paper and how it differs from past schizophrenia GWAS papers, and so perhaps would be better placed highlighted in the abstract.

I take the point about the replication of previous findings on neuronal function etc and consistency of results. However, I think this consistency is not too surprising given the overlap in GWAS samples between the present PGC-3 and those of recent studies that performed similar analyses, but clearly these are important results for the field of schizophrenia and will act as an important template for similar investigations into other diseases and disorders.

The conclusion regarding SCZ being primarily a disorder of neuronal function, but not restricted to certain brain regions, seems reasonable from these data and an important qualitative finding for the field to build on.

The evidence in the Supplementary Material showing that the prioritised genes have substantially greater mutation intolerance and loss of function than neighbouring non-prioritised genes is important and satisfies my query on this. However, I note that the authors have inserted highlighting these results "These protein-coding genes had a greater than 3-fold enrichment for loss of function intolerance…" on Line 204, in relation to 66 of the prioritised protein-coding genes, rather than the full 109 genes stated on Line 251. Assuming that the enrichment referred to is in relation to the 109 genes (as expected) then this should be described further down.

I am satisfied with all other responses, apart from multiple remaining issues with the exposition of the "Gene prioritisation" section, detailed below.

Overall, the updates to this section, aided by the new Figure 3A, have improved its exposition a lot. I realise now that the authors' stating of 31 loci with 5 or fewer credible SNPs was just a mention of this, and did not represent a step in the gene prioritisation, and among the various criteria for prioritisation I missed that one of those was selection of genes for which the entire credible set fell within the boundaries of one gene, each of which led to some confusion on my part. Thanks for the explanation. However, despite the improvements to this section, I still found a number of residual issues that should be addressed so that readers can understand the prioritisation process clearly:

It's not entirely explicit how the authors go from 270 "distinct loci" to 236 fine-mapped regions, but I presume that the fine-mapping indicated that those 270 distinct loci (containing 329 LD independent SNPs) were not distinct after all and that they instead corresponded to 236 fine-mapped (thus independent) regions?

In Fig.3A, the 294 LD independent "regions" and (combined GWAS) 329 LD independent "regions", should in fact be referring to SNPs (not regions).

Fig.3A refers to 63 genes containing entire credible set, whereas Fig. 3B refers to 58 (if Fig.3B only highlights protein-coding genes then this should be stated and according to the text this number should be 57 not 58. It was 58 in the original version of the paper, so presumably authors didn't update the figure).

In Fig.3B "Single Gene" label would be clearer if replaced by eg. "Full credible set".

In Fig.3B the header "SMR" is confusing (32 genes) considering that on Line 244 it is stated that there are "57 SMR-prioritised genes.. 47 protein-coding" (intersect of these 32 + 29 Hi-C + 17 eQTL > 50% PP genes). Recommend changing "SMR" label to "Single SMR" since there are other components that fall under the SMR prioritisation.

Why don't the 17 genes prioritised due to SMR eQTLs accounting for > 50% PP make it into

Fig.3B?

I believe that Fig.3B highlights protein-coding genes specifically, not those that are not protein-coding. This should be stated explicitly so that the reader can match with the figures elsewhere. Given that these seemingly refer to the protein-coding genes, then why don't the 375 Broad (non-priority) genes and the 109 priority genes add up to the 470 protein-coding genes stated in Fig.3A and line 198?

In the original version of the paper there were 290 broad regions and that has now become 375 - I just wonder what caused such a large change during the revision of the paper?

Note: There are now two figures labelled as Figure 3 in the paper, and Figure 4 should thus be re-labelled Figure 5.

Author Rebuttals to First Revision:

Referee #1 (Remarks to the Author):

Regarding the revised manuscript, I have only two comments. One substantive and the other minor.

1) I have to say it again: It seems unfortunate that all African-American and Latino samples are omitted from the main analyses. The new addition of African-American and Latino samples is useful (lines 106-120 and Supplementary Note), but I wish that more had been done. Instead of reporting on, and ideally including the small numbers of individuals with African ancestry (and other ancestries) from the PGC cohorts, the authors relied on recently published GWASs of African-American and Latino samples (Bigdeli 2020). I fully agree with their use of findings from Bigdeli et al, but they still "threw away" the small numbers of underrepresented minorities in the PGC cohorts. They argue that the numbers are too small (and indeed this is widely known in the research community), but it's not a great justification. Put another way: the authors argue that N=6,152 African-American cases is too small to include in the trans-ancestry meta-analyses. But an even smaller number of cases was deemed worthy when the subjects were white (ISC 2009 Nature case N=3,322). I realize that this might seem an unfair comparison, but in important ways it is not. Why should >6,000 African-American cases, with far better genomic coverage than in 2009, be thrown away? Further, this number would be somewhat larger still if other miscellaneous PGC cases were included. These same arguments apply to other ancestries.

We have now included the *African-American and Latino samples* as reported by Bigdeli in our primary analysis and as discussed in our response to the editors comments. We have also addressed the issue of the samples that we excluded in the above text.

We would also like to highlight two possible sources of additional samples in case these are a particular source of concern. The African American samples from the Molecular Genetics of Schizophrenia consortium (MGS; (maximum 1,286 cases; 973 controls) that were included in the ISC 2009 paper are included in the Bigdeli analysis, and therefore in our analysis. Before we froze the PGC dataset, we canvassed widely and specifically for non-European ancestry samples and it was clear that samples of Bigdeli and colleagues were not going to be available, and were being re-genotyped with more informative arrays. As Bigdeli has included an additional 7,386 cases not available to the PGC, the use of the results of Bigdeli maximises the sample size available for analysis. Another possible source we considered was the UK Biobank, particularly as a source of

ancestrally diverse controls we might use to 'rescue' some cases we excluded. However, the UK Biobank array is a different genotyping platform to those used for the samples we attempted to rescue. Thus, we are, at this time, unable to rescue these cases on the grounds that technical artifacts would assuredly arise from these analyses.

In the beginning of the revised manuscript, the authors describe data availability for European and Asian ancestry samples only. I don't think this sets a good example. At a very minimum, I think they should make it clear how researchers can access all summary stats (i.e. for multiple ancestries and for a full trans-ancestry meta-analysis).

We thank the reviewer for this fair and sensible suggestion, especially in the context of our new revision of the manuscript. We make summary stats available for the full Primary analysis as requested (https://www.med.unc.edu/pgc/download-results/), as well as the original analysis (now referred to as core PGC dataset), and ancestry specific analyses.

2) Typos on lines:

23: *possible "prioritized" instead of "priority*"?. We have changed this as suggested.

186: some issue here: "186 that contains 25 genes but is the only a single credible and maps within ATP2A2.". Thank you. We have changed this. This now reads "We highlight rs4766428 (PP>0.99) which is the only credible SNP in a locus that contains 25 genes and is located within *ATP2A2"*.

196 *"Figure 3A" is not bold.:* Corrected

Referee #2 (Remarks to the Author):

The authors have addressed my comments well.

Thank you.

Referee #3 (Remarks to the Author):

Thanks to the authors for their detailed responses to my original review. I have further comments below, which are all essentially minor.

I think make clear in the abstract that you also prioritise variants, as well as genes, here. I think your statement at the end of the Introduction "..and analysed the findings to prioritise variants, genes and biological processes that contribute to pathogenesis" is a nice summary of the impact of the paper and how it differs from past schizophrenia GWAS papers, and so perhaps would be better placed highlighted in the abstract.

Thanks for the suggestion. The final sentence of the abstract states that we prioritise variants and genes as follows.

"We identify biological processes relevant to schizophrenia pathophysiology, show convergence of common and rare variant associations in schizophrenia and neurodevelopmental disorders, and provide a rich resource of prioritised genes and variants to advance mechanistic studies".

I take the point about the replication of previous findings on neuronal function etc and consistency of results. However, I think this consistency is not too surprising given the overlap in GWAS samples between the present PGC-3 and those of recent studies that performed similar analyses, but clearly these are important results for the field of schizophrenia and will act as an important template for similar investigations into other diseases and disorders.

Thank you.

The conclusion regarding SCZ being primarily a disorder of neuronal function, but not restricted to certain brain regions, seems reasonable from these data and an important qualitative finding for the field to build on.

Thank you.

The evidence in the Supplementary Material showing that the prioritised genes have substantially greater mutation intolerance and loss of function than neighbouring non-prioritised genes is important and satisfies my query on this. However, I note that the authors have inserted highlighting these results "These protein-coding genes had a greater than 3-fold enrichment for loss of function intolerance…" on Line 204, in relation to 66 of the prioritised protein-coding genes, rather than the full 109 genes stated on Line 251. Assuming that the enrichment referred to is in relation to the 109 genes (as expected) then this should be described further down.

The way we have used FINEMAP to prioritise genes is novel, and therefore we were keen to obtain orthogonal support for our approach. The analyses discussed in this section are then based on the protein coding genes from the FINEMAP set, which was the number 66 referred to.

I am satisfied with all other responses, apart from multiple remaining issues with the exposition of the "Gene prioritisation" section, detailed below.

Thank you. We have further clarified these issues as listed below.

It's not entirely explicit how the authors go from 270 "distinct loci" to 236 fine-mapped regions, but I presume that the fine-mapping indicated that those 270 distinct loci (containing 329 LD independent SNPs) were not distinct after all and that they instead corresponded to 236 fine-mapped (thus independent) regions?

We agree this was confusing. The main issue is that we previously reported fine-mapping on that subset of the 270 regions from the extended GWAS (the analysis which included our primary GWAS and deCODE data) which were GWS in the initial core PGC dataset alone (as noted above, our analyses require a strong signal, access to individual level data, require identical QC and imputation procedures, and an accurate LD reference panel).

In the revised manuscript, we take the same approach but to simplify, we now provide the fine-mapping results of all GWS regions in the core PGC dataset (regardless of whether they are GWS in the multi-ancestry Extended analysis). We then use retained significance in the analysis to prioritise loci. Not only is this simpler, but as loss of some signal is expected in multi-ancestry analyses, our increased reporting will allow readers access to useful fine-mapping information on loci that would otherwise be omitted from the paper. We hope it is clearer in the revised manuscript, and to further facilitate transparency, we have also given the precise chromosomal co-ordinates of the fine-mapped regions in a new table **(Supplementary Table 11e)**.

In Fig.3A, the 294 LD independent "regions" and (combined GWAS) 329 LD independent "regions", should in fact be referring to SNPs (not regions).

We agree and have corrected it.

Fig.3A refers to 63 genes containing entire credible set, whereas Fig. 3B refers to 58 (if Fig.3B only highlights protein-coding genes then this should be stated and according to the text this number should be 57 not 58. It was 58 in the original version of the paper, so presumably authors didn't update the figure).

Thank you for spotting that. We have updated the numbers, ensured they match, and highlighted the fact that 3B refers to protein coding genes.

In Fig.3B "Single Gene" label would be clearer if replaced by eg. "Full credible set".

We agree and have adjusted this.

In Fig.3B the header "SMR" is confusing (32 genes) considering that on Line 244 it is stated that there are "57 SMR-prioritised genes.. 47 protein-coding" (intersect of these 32 + 29 Hi-C + 17 eQTL > 50% PP genes). Recommend changing "SMR" label to "Single SMR" since there are other components that fall under the SMR prioritisation.

The SMR genes in Figure 3B refers to the combined set of single SMR genes (ie the only SMR gene at the locus) and those SMR genes with eQTLs that cumulatively have >50% of the finempa posteruoir probability. We define this in the legend as there is no simple shorthand.

Why don't the 17 genes prioritised due to SMR eQTLs accounting for > 50% PP make it into Fig.3B?

As noted above, these are in the SMR set.

I believe that Fig.3B highlights protein-coding genes specifically, not those that are not protein-coding. This should be stated explicitly so that the reader can match with the figures elsewhere. Given that these seemingly refer to the protein-coding genes, then why don't the 375 Broad (non-priority) genes and the 109 priority genes add up to the 470 protein-coding genes stated in Fig.3A and line 198?

The reviewer is correct that 3B highlights protein coding genes. We apologise that this statement was not present in the legend and have now remedied this.

The numbers have now changed, but they did not match in the way the reviewer expected because 470 referred to genes identified only by the FINEMAP route, whereas the priority genes includes additional genes identified through the SMR route. We hope with all the clarifications we made in response to review, this is now clearer.

In the original version of the paper there were 290 broad regions and that has now become 375 - I just wonder what caused such a large change during the revision of the paper?

All the numbers have been revised due to other changes in the manuscript but for the reviewers interest, the change was a consequence of developments in FINEMAP over the last few iterations which reduced the number of regions we could not finemap to K<3.5. Some of these large regions with large credible causal SNPs sets distributed across large numbers of genes.

Note: There are now two figures labelled as Figure 3 in the paper, and Figure 4 should thus be re-labelled Figure 5.

Thank you. We have corrected figure labelling.

Reviewer Reports on the Second Revision:

Referee #1:
Remarks to the Author:
The authors have done a terrific job in this round of revisions. This version of the manuscript is beautifully polished, clear, and highly informative. The authors are to be commended for this tremendous effort.

As leaders in psychiatric genetics, the PGC's decision to include as many ancestries as possible in the primary analysis is both welcome and important. It shows to researchers around the globe that working toward greater representation of non-European ancestry populations is important. Thank you for doing this.

Referee #3:
Remarks to the Author:
I am now satisfied with the revisions made and thank the authors for making the clarifications requested and believe that having the trans-ancestry GWAS as the primary has enhanced the contribution of the study to the field