nature portfolio

Peer Review File

Mapping the Genetic Landscape across 14 Psychiatric Disorders

Corresponding Author: Dr Andrew Grotzinger

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Review
The manuscript by Grotzinger et al describes a combined analyses of psychiatric disorders, to gain insights into shared underlying genetics. They include data from 14 different disorders, representing over 1 Million cases. Their primary approach is genomic SEM. The authors confirm extensive overlap across psychiatric disorders. Their model returns 5 interpretable factors that are aligned with known groupings, such as internalizing or psychotic disorders. Additionally, a p-factor which was strongly linked to the internalizing factor fits the data well. The authors identify specific regions in the genome with pleiotropic associations. They also use enrichment analyses to characterise the different factors and find that the SB and internalizing factors are linked to specific types of cells and other biological annotations compared to the p-factor which instead captures broader biological annotations such as gene expression.
Overall, this manuscript uses state-of-the-art methods to provide novel and important insights about the genetics of psychiatric disorders. This work really delivers on utilising the extensively shared genetic architecture of psychiatric disorders for further insights. It improves our understanding of their genetics and also paves the way for more discoveries with translational potential. Scientific highlights from this work include the 5-factor model, the comparative findings in terms of annotations for the p-factor vs other factors, and the 100 pleiotropic loci this manuscript has identified.
I have a number of comments related the methods and interpretation, however.

Major comments
Diverse ancestry
It is commendable that the authors attempted to include studies for ancestrally diverse participants but the result of this attempt are largely unhelpful and misleading. Firstly, the authors group data from participants of African and African American ancestry and similarly East Asian ancestry from Western countries and participants from East Asian countries. This does not make any sense in the context of the genetic correlation analyses done by the authors. I expect the grouping was done by the authors of primary GWAS publications of these studies but their aim was identification of loci and grouping by genetic ancestry to account for different allele frequencies can indeed by sensible. Here, however, the authors aim to understand sharing of genetic architecture and it is hard to imagine that a study of Black people in the US has more in common with a study in Nigeria than with a study of people of European ancestry in the US.
The authors summarise the key findings as “Of note, cross-ancestry, within-trait rg results were significantly <1, and the genetic correlation between MD and SCZ in EAS genetic ancestry individuals was double that observed in EUR participants (.45 vs .22). This divergence is not fully understood and highlights the importance of investing in genetic efforts for other ancestry groups.”
For this East Asian ancestry GWAS, the high genetic correlation between SCZ and MDD and low genetic correlation between MDD in East Asian and in European ancestry studies is driven by one particular study, CONVERGE, and is not seen in other studies with East Asian ancestry participants, as the source paper for this GWAS describes in detail. So the conclusion of this present paper stands on very shaky legs. Whilst it is entirely possible that ancestry, geography, ethnicity or cultural background play a role and some of the correlational patterns of this paper do not generalize, there is insufficient evidence here to make any clear conclusions in this respect.
A smaller point but referring to European samples (p4) is not right and not what those guidelines they are citing are

suggesting either.
Overall, the ancestry comparisons are based on heterogeneous studies from different global regions, are rather limited in terms of disorders presented and in terms of sample size for most of the comparisons, consequently are disproportionality influenced by specific cohorts and their designs and therefore the very broad conclusions are not justified in their current form.

The influence of data overlap and outcome definition
No consideration has been given outcome definitions which is surprising given the leading role of the some of the coauthors in the scientific discourse on this. This discourse has so far focused on disorders linked to the internalizing cluster, in particular MDD, but I suspect some of the arguments may also extend to other conditions. This has the potential to significantly influence the findings of this manuscript and has not been accounted for during the analyses and not even been mentioned in the discussion as a limitation. Some of the primary GWAS included data from cohorts that were not dedicated case controls studies for psychiatric disorders, for example general cohorts and electronic healthcare resources. Some studies used extensive structured interviews based on ICD/DSM criteria, others short symptom questions or single questions to define outcomes. Previous research has shown that some of these have notable higher overlap with other disorders and are more likely to indicate general psychopathology rather than disorder specific components for MDD (Cai et al). This would be in line with the patterns in the current paper where only 10% of the internalizing factor is specific and the rest overlaps with the p-factor. Consequently, this is vital for the validity of the key conclusions of the current manuscript and I expect an extensive additional analytical effort to provide evidence that the current findings are valid even when strict definitions are applied to patients recruited from mental health clinics.
I am furthermore concerned about the impact of study overlap between disorders. This is for both cases and controls. The authors should work out where a study contributed data to more than one disorder and rerun their analysis after excluding the study from one disorder. As explained above, this may impact the key conclusions of this work.

Selection of disorders
It would be helpful if authors clearly described their rationale for including the disorders they did and reconsidered this choice. Tourette's, for example, seems to be an outlier and has little overlap with any other outcome. This may, however, influence the gSEM solution and other result. Not everyone would classify Tourette's as primarily psychiatric either.

Population stratification
PhWAS: The p-factor is associated with benign neoplasm of the skin. I am worried this suggests that there is population stratification. Skin neoplasm are strongly linked to skin colour and so population stratification within the included "European samples" seems a possible explanation.

Association analysis
The main analysis approach here was based on the gSEM results rather than any of the dedicated multi-trait GWAS methods. It would be helpful if the authors explained their choice. I am wondering whether a multitrait GWAS would be complementary to the existing findings and worth adding or even to replace the case-case GWAS findings which do not seem to add much and their role in this paper is not clear to me.

Minor comments
It would be helpful to clarify early on in the results section about sample sizes/power for the different disorders, e.g. a couple of sentences about table 1 would help. This is mentioned eventually in the discussion but would make interpretation easier if clarified sooner.
There is also a wider consideration of sample size implications missing in the discussion. For example, the sample size for major depression is much larger than for Tourette's. How does that impact the gSEM results, the discovery of pleiotropic loci etc?
Figure 1A weird scale and hard to estimate how high those rg values really are. This figure doesn't indicate which values are actually reliable and statistically significantly different from 0. Maybe they should only show values or add *.
Also, the groupings/SEM results are useful but by no means real distinct categories and some of them doent seem to work as well. I would suggest the authors point that out more explicitly as this will be read by wider audiences and the way the findings and figures are currently presented suggest stronger boundaries between groups than there really are.
Figure 2: It is difficult to follow the Mixer results. In those Venn diagrams, the shared grey area is takes up the vast majority of space for all disorder pairings which does not seem very likely to be correct.
Figure 5b hard to read, I suggest a different colour scheme or different line type to differentiate the Q ones from the others
Tge p-factor myopia association is also interesting.
The authors find 48 novel loci. At the moment the statement is just relative to the specific version of GWAS summary statistics they authors used but these are often just subsets of the data available in the primary GWAS. So the authors should do a literatire search for these loci to see if they are truly novel.
Only 33 loci with significant Qsnp across the factors: I found that surprising. This links back to the discussion on outcome definitions. However, I am also wondering whether there is a power issue?
"eQTL and Hi-C Datasets to Identify Target Genes" is a somewhat misleading title. This section is not about target genes at all. I understand the authors used a simple approach to link loci to genes which were then used for all these enrichment analyses but that is more of a technical detail and I think the title of the section is confusing.
The pathway analysis had no findings for the internalizing factor. This is quite surprising and not further explained.
The link to rare mutations in genes for neurodevelopmental disorders is very interesting. It would be important to clarify whether this is just driven by autism. The papers they are referring to are both specifically on ASD and that was also one of the disorders included here.
For the interpretation of enrichment analyses, especially some of the genomic annotations, it would be very helpful to learn

whether these are findings you would generally see for complex traits or whether any of this is specific to psychiatric outcomes.
Conclusion: “While much remains to be done, cross-disorder genetics continues to fill in critical gaps in our understanding of shared and unique psychiatric risk with implications for constructing a more biologically valid psychiatric nosology.” The paper had not touched on psychiatric nosology much. If the authors belief this is the case, then they need to discuss this more. Should we drop MDD, anxiety etc in favour of a broad group of internalizing conditions now? Else, what is being implied here?

Referee #2

(Remarks to the Author)
Applying genome-wide structural models to these psychiatric datasets has tremendous value because it provides a relatively unbiased view of shared genetic influences across disorders. By adding disorders that could not be meaningfully analyzed in previous iterations due to low statistical power (e.g. PTSD, anorexia), the authors have provided important additional information beyond their prior publications. Indeed, these results can be used to ground the field’s understanding of comorbidity patterns in a firm biological foundation. My primary recommendation for revision is that the work with cell types should be improved, principally with the use of whole brain cell type data (e.g. from Siletti et al 2023). Below are two important comments and other suggestions.

IMPORTANT:
In contrast to the strength of the genomic SEM work, the functional analyses are more crude, and specifically the choice of single cell and single nucleus RNA sequencing datasets is not good enough. There exists one human whole-brain atlas of cell types (Siletti et al 2023 Science) and two whole brain mouse atlases (Yao et al 2023 Nature, superseding Zeisel et al. 2018 Cell). The authors have instead used atlases of cell types that are much more restricted. There is no reason to do this, and doing so may miss important cellular associations. Thus, additional analyses need to be run before making conclusions like “SB factor was substantially enriched in genes expressed in excitatory neurons, whereas the internalizing factor was associated with oligodendrocyte biology.” Further, the cell type analyses focused on rather crude categories of cell types (e.g. only 8), when in fact there are at least hundreds and likely thousands of cell types in the human brain. Moreover, the more granular cell types offer much stronger associations with psychiatric phenotypes, even accounting for additional statistical test conducted on more granular cell types. e.g. it may be best to analyze data at the cluster level (n=461) from Siletti et al. vs supercluster level (n=31; too crude) or subcluster level (n=3,313; too noisy).

Relatedly, this statement oversimplifies results considerably, and it should be improved after additional analyses:
“At the functional level, we identified consistent results across different methods that highlighted the relevance of excitatory neurons, specifically hippocampal CA1 and CA3 neurons, for the SB factor and oligodendrocytes for the Internalizing disorders factor.”

IMPORTANT but not critical for interpretation

Line 1052 ERROR: using LDSC with admixed ancestry (African American datasets here) violates assumptions in an important way. Estimate is also way off (47%!). This has to be removed. It could be replaced with GCTA on the individual level data.

Top suggestions:
ABSTRACT: The abstract says a few things that are, in my understanding, incorrect about historical debates about nosology in psychiatry. For example, the article states: “Genomic methods have shown that even for schizophrenia and bipolar disorder, two disorders long-thought to be etiologically distinct3” But this is very much not true, it has long been debated as to whether or not these were the same, similar, or distinct conditions. Instead, the genetic results here beautifully capture points of greater certainty versus debate in the history of psychiatric nosology. In other words, where debate has endured over the decades, we see that genetic correlations are highest; therefore there is good biological reason for the longstanding debates. In contrast, few clinicians have argued, for example, that ASD and bipolar disorder are the same, and indeed they are more genetically distinct.

FIG 2: Do the authors need to use MIXER? I’ve always found it to have lots of problematic elements... from an expert statistical geneticist, “Mixer relies on an unrealistic model with a mixture of null SNPs (exactly zero effect) and causal SNPs with normally distributed effects. We showed that this type of approach leads to estimates that are sample size dependent - the estimated polygenicity increases rapidly as a function of sample size, as more causal SNPs are detected.” An alternative: S-LD4M

Just a compliment:

FIG 1B: Really love the use of the pie chart for each of the disorders, to show when the fit the factor well (many disorders) as opposed to disorders that map nearly evenly onto two factors (ADHD), or multiple factors but with one primary factor (TS), and when disorder specific factors are important. If complaints are raised about pie charts (per some journal guidelines), I would push back because this is such an informative and clear usage of pie charts.

Referee #3

(Remarks to the Author)
This study presents a thorough investigation of the shared and divergent genetic architecture of 14 psychiatric disorders. The use of multiple, complementary methods provides a rigorous approach to understanding the genetic relationships among these disorders. A major finding is the strong genetic clustering of bipolar disorder (BIP) and schizophrenia (SCZ) into one factor, and major depression (MD), PTSD, and anxiety disorders (ANX) into another. This work contributes to a better delineation of psychiatric nosology and lays the groundwork for future research into the biological underpinnings of these disorders.

Comments:

• It would be helpful to clarify what this study adds beyond Grotzinger et al. (2022, Nature Genetics), which analyzed 11 major psychiatric disorders with similar methodologies and overlapping findings. Are there methodological improvements, additional insights, or novel interpretations that distinguish this study?
• Table 1:
- Ordering the disorders in Table 1 by liability-scale heritability rather than alphabetically would provide a more informative structure.
- The table lacks an explanation for why Nicotine Dependence does not have case-control counts.
- An additional column for effective sample sizes would be valuable for understanding the relative power of each GWAS.
• The p-Factor:
- The p-factor may not solely reflect genetic risk for psychiatric conditions. It could also reflect one's (in)ability to have a healthy emotional reaction to adverse environmental circumstances or perhaps even the chance of being exposed to adverse environments. Could the p-factor be capturing neuroticism and/or SES? A genetic correlation between p and a range of potential outcomes (e.g., educational attainment, income, neuroticism) could help clarify this. If the p-factor was just another trait that is highly correlated with multiple disorders, that would explain the relative high number of QSNP hits.
- A genetic correlation analysis between p and a range of relevant complex traits (e.g., educational attainment, income, neuroticism) would provide useful context.
- The high number of QSNP hits for the p-factor suggests it may capture general behavioral/psychological variation rather than specific psychiatric risk.
• Chromosome 11:
- The pleiotropic hotspot on chromosome 11 raises the question of whether this region is a general genetic hotspot for psychiatric disorders, behavioral outcomes, or complex traits more broadly. Comparing this region's effects with those in GWAS of intelligence, personality, or addiction could clarify its broader relevance.
• Ancestry Differences:
- Given the marked differences between ancestries in genetic correlations between disorders, I think it should be emphasized (probably also in the abstract) that the underlying factors that are highlighted in this paper could very well differ between ancestries, further complicating what we are trying to achieve here, namely getting a better delineation of what the different psychiatric disorders are on a genetic level and how they differ.
- Page 5, lines 215-219: "Of note, previously reported within-disorder, within-ancestry rg's calculated across cohorts are considerably smaller for PTSD (rg = 0.73, SE=0.21)35 and MD (rg = 0.76, SE =0.03)36 than SCZ (rg = 0.95, SE = 0.03)37, suggesting that cross-ancestry rg's for these two disorders reflect a combination of etiological and phenotypic heterogeneity and ancestry-specific signal." PTSD and MD are also disorders that may be more influenced by life events, perhaps worth adding that environmental and cultural differences between populations may influence which underlying traits (and thus genetic effects) are associated with these disorders across populations.
- The discussion states: "First, at a genome-wide level, we confirm pervasive genetic overlap across 14 clinically distinguished psychiatric disorders, as indicated by large pairwise rg across multiple genetic ancestries." However, the results show that genetic correlations differ significantly across ancestries. This statement should probably be toned down to reflect these findings more accurately.
• Shared loci between disorders:
- Page 6, lines 276-279: "Consistent with the multivariate genetic structure identified using Genomic SEM, the pairs of disorders with the greatest number of significant local rg hits were MD and ANX (113 rg loci), MD and PTSD (88 rg loci), and BIP and SCZ (40 rg loci), accounting for just over half of all significant local rg's detected (Fig. 3A)." Why would MDD and ANX/PTSD have more shared regions be significant than BIP and SCZ, while the latter has both higher rg between each other and higher heritabilities?
• More Genetic Correlations:
- I think there is a missed chance to get a better sense of the content of the signal, more so than diving into the biology. These are all behavioral outcomes, which are a composite of lower level outcomes and environmental influences (Abdellaoui & Verweij 2021). Looking at rg patterns with other complex traits could give us a much better idea of the content of the signal. This could clarify whether certain factors (or p-factor) might reflect something else than just psychiatric genetic risk; it's likely that normal dimensions of variation are captured here. The Phewas does not seem sufficient, as that only considers disease traits and doesn't give a quantitatively interpretable unit of genetic overlap, only whether it's significantly associated or not. Adding rg analyses with a wider range of complex traits is a relatively easy analysis to add, much easier than the analyses already done in this paper.
• Power differences:
- How do differences in statistical power across disorders affect the interpretation of shared genetic factors? Are there sensitivity analyses that could address this?

• Clarity:
- Figure 1 is beautiful, well-designed and clear.
- I find Figure 2 a bit more difficult to interpret (and I'm not sure it warrants being a main Figure?).
- The LAVA and Q-SNP results are among the more complex parts of the study. If possible, adding a clearer explanation of their use and significance (especially LAVA) would help readers unfamiliar with these approaches.
• Common Variants:
- It should be noted in a prominent place, I suggest the abstract, intro, and discussion (perhaps even the title) that this study considers *common* genetic variants only.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have fully addressed my comments.
Overall, this work uses state-of-the-art methods to provide novel and important insights about the genetics of psychiatric disorders. This will be of great interest to the research community.

Referee #2

(Remarks to the Author)
Can the authors please use MAGMA and LDSC for the cell type analyses? The method selected may be fine, but I believe much more rigorous testing of MAGMA and LSDC has been conducted. Further, to make findings comparable with prior leading reports examining cellular associations with psychiatric disorders I would recommend prioritizing MAGMA and/or LDSC results.

Referee #3

(Remarks to the Author)
Most of this work is very well done. The reviewers have addressed my comments and those of the other reviewers well. But there is one major issue with the interpretation that I think is important to address.

I'm happy to see that the authors followed my earlier suggestion to look at genetic correlations (rg's) with "normal" complex traits. I believe these results actually support the idea that the latent factors identified here are not "disease" dimensions per se. If you look at the rg's with various traits, many are quite high, and keep in mind that all the signal comes from common variants (meaning they occur a lot because they probably do "normal" things). That makes it much more likely that these factors reflect normal dimensions of genetic variation -> dimensions that increase risk for psychiatric disorder only when someone falls at an extreme end and is exposed to specific environmental stressors.
I strongly disagree with how the QTrait statistic is used to determine which rg's are considered "significant." The QTrait metric is certainly useful for interpretation and should be reported, but it shouldn't drive the main presentation of the rg's. In its current use, it effectively filters out strong and meaningful genetic correlations, giving a misleading impression of what the latent factors are associated with.
For example, on page 7 the authors write: "Daytime napping was the only trait significantly correlated with the p-factor (rg_p = .31, SE = .03)." And on page 12: "However, the p-factor also had more hits for the QSNP heterogeneity metric (117) than all five-factors from the correlated factors model (33) and was significantly correlated with daytime napping only, indicating the p-factor alone is insufficient for explaining cross-disorder risk."
This interpretation is difficult to understand. It appears that a trait is only considered "significantly correlated" with the p-factor if its rg is significant and the QTrait is not significant. But this decision obscures the true pattern of genetic correlations. Supplementary Figure 31 shows many traits with high rg's with the p-factor: neuroticism has an rg > 0.6, income around 0.4, and there are also substantial rg's with traits like self-harm, loneliness, and childhood intelligence. These are all highly informative and clearly indicate that the p-factor reflects a broad dimension of heritable variation that includes a strong neuroticism/SES component. This is also consistent with the hypothesis I raised in my previous review.
By using QTrait significance as a filter, Figure 2 ends up showing only "daytime napping" as associated with the p-factor, which is clearly misleading. This framing underplays one of the paper's most interesting findings. I recommend showing all significant rg's and presenting QTrait separately as a heterogeneity metric, not as a disqualifier.
More generally, I find Figure 2 quite difficult to interpret. The design is unintuitive, and it's not immediately clear what is being shown or why. I would encourage the authors to consider a more traditional and readable way to present these results. Also, selecting only the top rg's to display in Figure 2 is not informative; it obscures the broader pattern of associations and may give a skewed view of the genetic architecture being described.

Finally, I would encourage the authors to make use of Extended Data Figures, especially for results that are central to the interpretation. Here are three specific suggestions:
• Supplementary Figures 20–25 could be merged into a single, large, readable Extended Data Figure that spans a full page.
• Similarly, Supplementary Figures 26–31 could be merged in the same way.

• Supplementary Figure 5 is highly informative and contains very important information for the future of the field of psychiatric genetics. It deserves to be shown as an Extended Data Figure attached to the main pdf.

Version 3:

Reviewer comments:

Referee #2

(Remarks to the Author)
The reviewers have addressed my comments.

Referee #3

(Remarks to the Author)
The reviewers have fully addressed my comments, I am happy with this beautiful piece of work being published in Nature.

Referee #1 (Remarks to the Author):

Review

The manuscript by Grotzinger et al describes a combined analyses of psychiatric disorders, to gain insights into shared underlying genetics. They include data from 14 different disorders, representing over 1 Million cases. Their primary approach is genomic SEM. The authors confirm extensive overlap across psychiatric disorders. Their model returns 5 interpretable factors that are aligned with known groupings, such as internalizing or psychotic disorders. Additionally, a p-factor which was strongly linked to the internalizing factor fits the data well. The authors identify specific regions in the genome with pleiotropic associations. They also use enrichment analyses to characterise the different factors and find that the SB and internalizing factors are linked to specific types of cells and other biological annotations compared to the p-factor which instead captures broader biological annotations such as gene expression.

Overall, this manuscript uses state-of-the-art methods to provide novel and important insights about the genetics of psychiatric disorders. This work really delivers on utilising the extensively shared genetic architecture of psychiatric disorders for further insights. It improves our understanding of their genetics and also paves the way for more discoveries with translational potential. Scientific highlights from this work include the 5-factor model, the comparative findings in terms of annotations for the p-factor vs other factors, and the 100 pleiotropic loci this manuscript has identified. I have a number of comments related the methods and interpretation, however.

Major comments

Diverse ancestry

It is commendable that the authors attempted to include studies for ancestrally diverse participants but the result of this attempt are largely unhelpful and misleading. Firstly, the authors group data from participants of African and African American ancestry and similarly East Asian ancestry from Western countries and participants from East Asian countries. This does not make any sense in the context of the genetic correlation analyses done by the authors. I expect the grouping was done by the authors of primary GWAS publications of these studies but their aim was identification of loci and grouping by genetic ancestry to account for different allele frequencies can indeed by sensible. Here, however, the authors aim to understand sharing of genetic architecture and it is hard to imagine that a study of Black people in the US has more in common with a study in Nigeria than with a study of people of European ancestry in the US.

The authors summarise the key findings as "Of note, cross-ancestry, within-trait rg results were significantly <1, and the genetic correlation between MD and SCZ in EAS genetic ancestry individuals was double that observed in EUR participants (.45 vs .22). This divergence is not fully understood and highlights the importance of investing in genetic efforts for other ancestry groups." For this East Asian ancestry GWAS, the high genetic correlation between SCZ and MDD and low genetic correlation between MDD in East Asian and in European ancestry studies is driven by one particular study, CONVERGE, and is not seen in other studies with East Asian ancestry participants, as the source paper for this GWAS describes in detail. So the conclusion of this present paper stands on very shaky legs.

Whilst it is entirely possible that ancestry, geography, ethnicity or cultural background play a role and some of the correlational patterns of this paper do not generalize, there is insufficient evidence here to make any clear conclusions in this respect.

A smaller point but referring to European samples (p4) is not right and not what those guidelines they are citing are suggesting either.

Overall, the ancestry comparisons are based on heterogeneous studies from different global regions, are rather limited in terms of disorders presented and in terms of sample size for most of the comparisons, consequently are disproportionality influenced by specific cohorts and their designs and therefore the very broad conclusions are not justified in their current form.

We fully agree on the importance of ancestrally inclusive genetic studies, and we also appreciate the reviewer's comments on the need to clearly articulate current limitations and ensure cautious, nuanced interpretation of results, given the current constraints in available methods and data. Based on the reviewer comments we have now tempered language regarding strong conclusions given the current methods and available data and have provided more nuance and explanation of limitations. We thank the reviewer for raising these important points and feel the discussion around these issues has greatly improved. Summary of edits, responses, and rationale are provided below.

We agree that the question of how best to group and analyze data across populations is important, particularly when the aim is to investigate shared genetic architectures across studies. Our decision to stratify analyses by genetic ancestry follows conventions commonly used in the primary GWAS[1,2], which, as the reviewer notes, are typically stratified to account for differences in allele frequencies, linkage disequilibrium patterns, and to mitigate confounding due to population stratification. Post-GWAS analyses, including genetic correlation estimation and Genomic SEM, are impacted by these factors too, which is why we maintained the primary GWAS ancestry-based groupings in this work.

We also note that while environmental and cultural differences may lead to distinct patterns across regional populations—even within the same genetic ancestry (as the reviewer points out)—our analyses are designed to examine underlying genetic architectures (e.g., similar genetic mechanisms, pathways), which we expect to be largely similar across populations. That said, we recognize that the ancestry groupings used here (largely based on similarity to broad reference panel populations) are proxies and do not fully capture the complex interplay between genetic, environmental, and cultural factors. At present, however, non-European-like genetic ancestry datasets are often orders of magnitude smaller in sample size, limiting the power for robust cross-cultural or cross-region analyses. As the reviewer rightly points out, extending these analyses across ancestries, regions, and cultural contexts is an important future direction and should have been included in the original Discussion. We have now added these limitations in the manuscript and highlight the need for future methods development and data collection efforts to enable such work. This update was guided by co-author Dr. Roseann Peterson, who has leading expertise in analyses in ancestrally diverse populations.[2] We now write on p. 12:

"Our study has several limitations. Analyses were restricted primarily to European-like genetic ancestry populations due to the limited availability of GWAS data for other groups and the limitations of methods requiring more genetically homogeneous groups[61]. GWAS for non-EUR-like populations sample sizes are still orders of magnitude smaller and not currently powered for more precise cross-ancestry assessments; this emphasizes the need for future research including the generation of additional ancestrally representative data, which will allow well-powered studies

and the examination of cross-disorder genetic architecture across regional and cultural differences."

Among six phenotypes included in the cross-ancestry analyses, the three substance use traits were largely underpowered for POPCORN analysis (AFR-like; AUD: 3335 cases/2945 controls; CUD: 3838 cases/5897 controls; OUD 5212: cases/26876 controls). However, for SCZ (EAS-like; 22778 cases/35362 controls), MD (EAS-like; 13893 cases/155912 controls), and PTSD (AFR-like; 17947 cases/59199 controls) sample sizes are much larger, and we consider those data to be useful for estimating cross-ancestry genetic-correlation analyses, albeit with limitations which we now reflect in the updated Discussion section highlighted above. In addition, we revised the Results section for our cross-ancestry analysis to further emphasize that results from our work need to be interpreted with caution until large replication datasets are available (p. 5):

*"As the majority of analyses were restricted to participants of EUR-like genetic ancestry, we sought to gauge how generalizable our findings were across ancestral groups. This was achieved using Popcorn[32], which can estimate r_g's for the same trait across different ancestral groups. We specifically calculated the genetic impact correlation (ρ_{gi}), which considers different allele frequencies across populations by calculating the correlation between the population-specific, allele-variance-normalized SNP effect sizes. Results were underpowered for many comparisons, but included a strong EAS-EUR correlation for SCZ (ρ_{gi} = 0.85, SE = 0.04), followed by lower correlations between EAS-like and EUR-like for MD (ρ_{gi} = 0.67, SE = 0.16) and for AFR-like and EUR-like PTSD (ρ_{gi} = 0.59, SE = 0.27; **Suppl. Table 4**). While these results suggest that the findings that follow for EUR-like ancestry groups may generalize better for some disorders (e.g., SCZ) than for others (e.g., PTSD, MD), that conclusion awaits replication in more highly powered analyses."*

The reviewer notes that the East Asian-like genetic ancestry GWAS for SCZ and MD include heterogeneous mixtures of participants from Western and East Asian countries; however, the vast majority of participants are from East Asian countries. Specifically, for the SCZ GWAS most participants were from Mainland China (16759 cases/27753 controls), with remaining 6019 cases/7609 controls from Hong Kong, Indonesia, Japan, Korea, Singapore, and Taiwan regions. For the MD GWAS, we note that summary statistics shared by the primary GWAS did not include 23andMe and Women's Health Initiative. From the cohorts included in the meta-analysis, the majority were participants from mainland China (11984 cases/145333 controls) and Taiwan (1348 cases/8392 controls), with only 561 cases/2187 controls across four relatively small studies from the UK or US. The number of non-EUR-like genetic ancestry participants included in Western countries is thus minimal for these particular instances and also not powered to detect regional and cultural differences.

The AFR-like genetic ancestry PTSD GWAS meta-analyses applied a different strategy, identifying 11560 cases / 39474 controls that admittedly included a more broadly defined African-like ancestry, determined based on global reference panels[3] and genetic similarity computed by SNPWeights[4] method. African ancestry-like group were defined as a union of two groups: (1) individuals with ≥5% African ancestry, <90% European ancestry, <5% East Asian, Native American, Oceanian and Central-South Asian ancestry; and (2) individuals with ≥50% African ancestry, <5% Native American, Oceanian and <1% Asian ancestry. We appreciate that this is a broad definition,

which may impact alignment with available LD reference panels as required by most post-GWAS analysis tools including POPCORN and LDSC. We therefore applied three distinct reference panels *based on 1000 Genomes reference populations (ASW [African Ancestry in the Southwest USA], AFR [aggregated data across multiple African ancestry populations included in the reference panel], and a combined LD matrix comprised of EUR and AFR reference panel populations), estimating genetic correlations between 0.51 (SE 0.24) and 0.78 (SE 0.68), albeit with large uncertainty in these estimates. To facilitate post-GWAS analysis the primary GWAS summary statistics may need to be shared more granularly, i.e. for genetically more similar sub-groups within broadly defined groups.*

We thank the reviewer for highlighting the extended discussion of the CONVERGE study in MD GWAS. As our work did not present further advances in this area we referred to the original publication (p. 5):

"The r_g between MD and SCZ in EAS-like participants (r_g = 0.45; SE = 0.09) was double that observed in EUR-like participants (r_g = 0.22; SE = 0.04), which was shown[28] to be driven by a single cohort of severe and recurrent MD[31]."

We have now also updated the population descriptors as the reviewer rightfully noted needed editing. For example, we now write in the Results section (p. 5):

"Due to uneven representation of ancestral groups, the full set of cross-disorder analyses was restricted to GWAS summary statistics from a single genetic ancestry group—European-like (EUR-like)—defined based on genetic similarity to European descent in global reference panels[27]. We also report bivariate results for MD[28] and SCZ[29] in East Asian-like (EAS-like) genetic ancestry groups and AUD[30], CUD[25], OUD[24], and PTSD[22] in African-like (AFR-like) genetic ancestry groups similarly defined based on reference panels."

The influence of data overlap and outcome definition

No consideration has been given outcome definitions which is surprising given the leading role of the some of the coauthors in the scientific discourse on this. This discourse has so far focused on disorders linked to the internalizing cluster, in particular MDD, but I suspect some of the arguments may also extend to other conditions. This has the potential to significantly influence the findings of this manuscript and has not been accounted for during the analyses and not even been mentioned in the discussion as a limitation. Some of the primary GWAS included data from cohorts that were not dedicated case controls studies for psychiatric disorders, for example general cohorts and electronic healthcare resources. Some studies used extensive structured interviews based on ICD/DSM criteria, others short symptom questions or single questions to define outcomes. Previous research has shown that some of these have notable higher overlap with other disorders and are more likely to indicate general psychopathology rather than disorder specific components for MDD (Cai et al). This would be in line with the patterns in the current paper where only 10% of the internalizing factor is specific and the rest overlaps with the p-factor. Consequently, this is vital for the validity of the key conclusions of the current manuscript and I expect an extensive additional analytical effort to provide evidence that the current findings are valid even when strict definitions are applied to patients recruited from mental health clinics.

We appreciate the reviewer raising this point and agree that this is a critical consideration in psychiatric genetics. In the original GWAS papers for the MD, BIP, PTSD, SCZ, and ANX datasets used here the ascertainment sensitivity analyses reveal very little ascertainment specific signal when splitting the datasets across clinical, community, or EHR datasets. This is with the exception that self-report of prior diagnosis without further assessment of diagnostic criteria often showed more divergent and less disorder-specific signal (self-report of diagnosis typically reflected the 23andMe sample, which we do not use in the current analyses). This is in line with the Cai et al. paper that the reviewer highlights, which found with the least depression specific signal was self-reported prior help seeking for depression, whereas self-report of full MDD diagnostic criteria showed strong genetic overlap with "gold" standard phenotypes of (semi)-structured clinical interviews delivered by trained raters or clinicians. Taken together, this indicates that ascertainment can be an issue, but largely for a specific form of self-report when full diagnostic criteria is not assessed[5].

We now perform an extensive series of sensitivity analyses (including the addition of 5 Supplementary Tables and 10 Supplementary Figures) that examine the genetic correlations, factor structure, and multivariate GWAS results when excluding short-form self-report cohorts from the psychiatric GWAS datasets. We find that results are largely unchanged, including for the Internalizing factor and its corresponding factor loading on the p-factor in the hierarchical model. We now write in the main text (p. 7):

"As some of the underlying data was obtained using brief, self-reported diagnoses, we performed a sensitivity analysis in which those data were excluded (Suppl. Note; Suppl. Tables 7-11; Suppl. Figures 11-19). The genetic correlation matrix was largely unchanged; the five-factor model identified in the full sample continued to provide good fit to the data and produced similar point estimates, and downstream GWAS analyses (detailed below) identified a similar set of loci."

We then provide further details in the Online Supplement (supplemental pages 3-6):

Case Ascertainment Sensitivity Analyses

The impact of shallow versus deep ascertainment methods for psychiatric cases is an increasing topic of concern and discussion[3]. In support of this concern, ascertainment stratified analyses for major depression in the UK Biobank (UKB) sample have found that cases derived using shallow case definitions (e.g., self-report of prior help seeking for nerves, anxiety, or depression) capture broadly pleiotropic pathways that lack depression specific signal[6]. This was as compared to more deeply phenotyped cases from UKB, defined using self-report of complete diagnostic symptoms, that showed higher genetic overlap with MD defined in outside samples by a (semi-)structured clinical interview delivered by a trained rater. Recent findings for other disorders reveal a similar pattern of results, where the most recent PGC GWAS for BIP[6], MD[7], PTSD[8], ANX[9], and OCD[10] show that results stratified by either clinical, community, EHR, or biobank samples all evince high levels of genetic overlap, with minimal evidence for ascertainment-specific genetic signal. This was with the one exception that self-report of diagnosis (typically defined using the 23andMe sample) often showed lower genetic overlap with clinical and community samples. We note that the results in the main text of the current CDG3 project were generated using the univariate GWAS summary statistics that excluded 23andMe for all psychiatric disorders, which should already significantly reduce genetic heterogeneity due to ascertainment.

We investigated whether restricting analyses to even more strictly ascertained cases affects the conclusions from the correlated factors and hierarchical p-factor model at both the genome-wide and locus level of analyses. This involved making the following two exclusions for these disorders. First, we removed cases defined by self-report of diagnosis or short format questionnaires (1-4 questions) that do not assess complete diagnostic criteria. In addition, we excluded cohorts that were initially ascertained for another disorder, as this may upwardly bias genetic correlations with that same disorder. For SCZ, NIC, AN, ASD and TS this did not result in any change to the GWAS datasets. For CUD, power would be insufficient following the second exclusion criteria to include it in the model and results should be interpreted with this in mind. For the remaining six disorders the following exclusions were made.

PTSD. For PTSD, we excluded self-report cohorts that did not administer the clinician administered PTSD scale (CAPS) that covers full diagnostic criteria. In addition, we exclude cohorts initially ascertained for other disorders (e.g., iPSYCH). This resulted in a 27.27% reduction in effective sample size (Supplementary Table 7 for case/control numbers).

MD. For MD, we excluded the self-report questionnaire cohorts that included cohorts with self-report of prior diagnosis (e.g., HELIUS). This resulted in a 16.65% reduction in effective sample size.

ANX. For ANX, we excluded self-report of prior diagnosis cohorts (e.g., UKB) along with cohorts initially ascertained for another disorder (what is referred to as the comorbidity cohorts in the ANX study that included the QIMR and Utah Suicide cohorts). This resulted in a 47.56% reduction in effective sample size.

ADHD. The ADHD analyses excluded the Yale-Penn cohort as this was initially ascertained for substance use disorders. Analyses were otherwise left unchanged, as the 11 remaining cohorts all met the inclusion criteria, with ADHD cases diagnosed by either semi-structured clinical interviews or psychiatrists. The exclusion of the Yale-Penn cohort resulted in an 0.62% reduction in the effective sample size.

OCD. For OCD we excluded only the iPSYCH cohorts as this was initially ascertained for five other psychiatric diagnoses. This resulted in a 18.71% reduction in effective sample size.

BIP. From bipolar we exclude only the UKB cohort as this reflects self-report of prior diagnosis. This resulted in a 5.56% reduction in effective sample size.

Sensitivity Analyses: LDSC and Genomic SEM. The genetic correlation matrix in the sensitivity sample continued to show high levels of genetic overlap across the 14 disorders, along with notable clustering among subsets of disorders (Suppl. Fig. 11; Suppl. Table 8). On average, there were minimal differences in the genetic correlation matrix between the full sample and sensitivity sample LDSC matrix, with genetic correlations often being slightly larger in the full LDSC sample: for genetic correlations where at least one disorder was updated for the sensitivity analyses, the mean difference between the sensitivity and full sample was -.03 (range: -.18: .06; SD= .05).

We went on to perform the same combination of exploratory factor analyses (EFAs) fit to the LDSC correlation matrix estimated in odd autosomes that were used to guide confirmatory factor analysis (CFAs) using LDSC estimated for even autosomes. As in the full sample, we examined the fit for a common factor model and for one to five factor model solutions identified using varimax or promax rotations at the EFA stage (see main text Method for additional details; Suppl. Table 9 for model fit in sensitivity sample). As in the full sample, the best fitting model derived from EFAs performed in the sensitivity sample was a five-factor correlated factors model (Comparative Fix Index [CFI] = .926, Standard Root Mean Square Residual [SRMR] = .088). This model also provided good fit when applied to the LDSC matrix for all autosomes (CFI = .907, SRMR = .086; Suppl. Table 10 for model output). The five factors identified using the EFA in the sensitivity analysis were defined by a similar subset of disorders relative to those identified using the EFA in the full sample. For example, PTSD, MD, and ANX (three of the disorders with the largest change in effective sample size across the full and sensitivity samples) continued to cluster together onto an internalizing factor. The largest shift was for the third factor, which was defined in the sensitivity sample by ASD and NIC, but was defined by ASD, ADHD, and TS in the full sample.

Given the similarity in factor structure identified using the EFA conducted in the sensitivity sample, we also examined whether the factor structure identified in the full sample also provided good fit in the sensitivity sample. We find that the five-factor correlated factors model identified in the full sample continued to provide the best fit to the data in the sensitivity sample when applied to both odd autosomes (CFI = .945, SRMR = .074) and all autosomes (CFI = .953, SRMR = .064). The point estimates for the five-factor correlated factors model applied to the sensitivity data were also highly similar to estimates from the full sample (Suppl. Fig. 12 for path diagram with comparisons of estimates). The hierarchical p-factor model also provided good fit to the sensitivity data in all autosomes (CFI = .937, SRMR = .074), and once again the loadings of the five psychiatric factors on the p-factor were highly similar in the full and sensitivity sample (Suppl. Fig. 12). The correlated factors model and hierarchical models identified in the full dataset were used to estimate multivariate GWAS in the sensitivity dataset as they both provided better fit to the sensitivity data and offer a more direct form of comparison to the full sample results.

Sensitivity Analyses: Multivariate GWAS. Multivariate GWAS results from the correlated factors and hierarchical models in the sensitivity dataset largely mirrored results from the full dataset. This was indicated, in part, by inspection of Miami and QQ-plots that directly compare the sensitivity and full GWAS factor and Q_{SNP} results; these results generally reveal similar overall patterns (Suppl. Figs. 13-18). In addition, sensitivity GWAS hits were largely subsumed by the original GWAS hits for the factors (Suppl. Table 11). We focus here on GWAS results for the Internalizing factor, which was defined by the three disorders with the largest reduction in effective sample size in the sensitivity sample (PTSD, MD, and ANX). Among the 69 Internalizing factor hits from the sensitivity GWAS, 98.5% were also hits in the original GWAS. Among the 150 Internalizing hits from the original GWAS, 100 hits were not captured by the sensitivity GWAS; however, the average p-value among those 100 hits in the sensitivity GWAS was trending significant (average p = 2.37E-5) and still far lower than the average p-value across all SNPs in the sensitivity GWAS (average p = .422). This is visually depicted in Suppl. Fig. 19, which shows that the distribution of p-values among these 100 SNPs is in the outer right-tail of a distribution of

-log10(p)-values. We take this to indicate that the greater number of hits from the original GWAS reflects increased power from the added participant samples in the original univariate GWAS. This as opposed to an alternative explanation that attributes these 100 hits to reflect signal driven by particular forms of case ascertainment being included in the univariate GWAS. For the Q_{SNP} heterogeneity statistic, we identify 17 hits in the original GWAS and no hits in the sensitivity GWAS for this same Internalizing factor. In this instance, these 17 SNPs were generally no more significant for Q_{SNP} (average p = .348) than the rest of the Q_{SNP} p-values for the sensitivity GWAS (average p = .510; Suppl. Fig. 19). This suggests that these SNPs may be picking up on heterogeneity introduced by different forms of case ascertainment being folded into the univariate GWAS of the psychiatric disorders. As Q_{SNP} is used as a quality control on the factor-level results, factor hits presented in the main text (and downstream analyses that use these hits as input) are not affected by this potential heterogeneity.

*Sensitivity Analyses: Summary. At the genome-wide level, genetic correlations and the genomic factor structure in Genomic SEM (including the size of the factor loadings) were largely unchanged when restricting the psychiatric cases to those that were more strictly ascertained. At the locus level, the signal was slightly attenuated for the sensitivity sample, as would be expected given the reduction in sample size and corresponding decrease in statistical power. However, the overall trend in variant level findings was highly concordant across the original and sensitivity GWAS. The concordance in results across the sensitivity and original sample here is in line with the prior results from univariate GWAS of psychiatric disorders noted above that indicate concordance in results across most forms of ascertainment, coupled with the fact that we do not use the short format, self-report 23andMe sample found to be the most divergent in many of these prior analyses[6,7,9,10]. While future analyses should continue to examine the effects of case ascertainment, the current findings indicate that including the full univariate GWAS samples offers increased statistical power with minimal bias introduced by including other forms of ascertainment. To facilitate ascertainment focused questions in future work, we make available the full sensitivity GWAS summary statistics for the factors (see **Data Availability** statement in main text).”*

As noted at the end of the excerpt above from the Online Supplement, we now make the full set of sensitivity factor GWAS results available alongside the GWAS results from the full samples. We also now write in the Discussion (p. 13):

“Another limitation reflects potential inflation in r_g estimates by cross-trait assortative mating[65], diagnostic misclassification[66], or the use of super-normal controls[7]. However, the high genetic overlap observed among subclusters of psychiatric disorders are unlikely to be explained by cross-trait assortment alone[68] and current sensitivity analyses using stricter case definitions suggested that impact of diagnostic misclassification was modest.”

I am furthermore concerned about the impact of study overlap between disorders. This is for both cases and controls. The authors should work out where a study contributed data to more than one disorder and rerun their analysis after excluding the study from one disorder. As explained above, this may impact the key conclusions of this work.

We agree with the reviewer that study overlap between disorders is an important issue that requires more elaboration.

Firstly, we note that case-case and control-control study overlap primarily increases the covariance of the error-terms of the GWAS results[8,9]. This results in an increased intercept in cross-trait LDscore regression, but does not affect the slope of cross-trait LDscore regression used to estimate the genetic correlation (Eq 1 in Bulik-Sullivan 2015[8]). Consequently, GenomicSEM results, building on genetic correlation estimates, are not impacted by this type of study overlap.

Second, case-control study overlap (i.e., cases of disorder B included as controls for disorder A) decreases the covariance of the error term. Again, this is expected to only impact (decrease) the intercept of LDscore regression and not its slope, hence not expected to impact GenomicSEM results.

However, when case-case overlap results from diagnostic misclassification from one of the disorders, this may result in upward bias in the genetic correlation estimate [10]. Additionally, when control-control overlap pertains to a set of super-normal controls this can also result in upward bias. However, we note that these biases would also be present without sample-overlap, and that assessing potential diagnostic misclassification or use of super-normal controls in the original GWAS studies is beyond the scope of this project.

We have made two changes to our revised manuscript. This includes updating the section describing the GenomicSEM method to clarify that sample overlap is not expected to impact GenomicSEM results:

"We note that study overlap between disorders is not expected to impact the findings, because study overlap only impacts the covariance of error-terms of the GWASs resulting in increased intercepts of cross-trait LDSC with no expected impact on the estimates of genetic correlation.[4,43]"

We have also updated the limitation paragraph of the Discussion section to note:

"Another limitation reflects potential inflation in r_g estimates by cross-trait assortative mating[65], diagnostic misclassification[66], or the use of super-normal controls[67]. However, the high genetic overlap observed among subclusters of psychiatric disorders are unlikely to be explained by cross-trait assortment alone[68] and current sensitivity analyses using stricter case definitions suggested that impact of diagnostic misclassification was modest."

Selection of disorders

It would be helpful if authors clearly described their rationale for including the disorders they did and reconsidered this choice. Tourette's, for example, seems to be an outlier and has little overlap with any other outcome. This may, however, influence the gSEM solution and other result. Not everyone would classify Tourette's as primarily psychiatric either.

We agree with the reviewer that the choice of disorders included is important as this could potentially impact the factor structure. We now write in the Results section (p. 4):
"Psychiatric disorders were included if described in a psychiatric diagnostic manual[9,10] and power was sufficient to interpret genetic correlations[8]."

The reviewer is also correct that Tourette's syndrome (TS) appears as an outlier in our results, with the most unique variance (i.e., residual variance not explained by the latent factors) of all disorders considered (87%). We highlight that TS is the most unique among this set of disorders, but still shows higher genetic correlations with the included traits than other neurocognitive disorders. For example, the highest genetic correlation between Alzheimer's disease and the psychiatric disorders included here was .19 with bipolar disorder. By comparison, TS is genetically correlated .41 with OCD and .28 with ANX in the current analyses. In addition, the largest study with comprehensive assessments of comorbid DSM-IV-TR psychiatric disorders in individuals with TS demonstrated that 85% of children with TS had one or more comorbid psychiatric disorders, with the two most prevalent disorders being OCD and ADHD[11] [doi: 10.1001/jamapsychiatry.2014.2650]. Taken together, TS appears to show sizeable genetic and phenotypic overlap with other psychiatric disorders, which lends itself to the cross-disorder framework employed here.

Population stratification

PhWAS: The p-factor is associated with benign neoplasm of the skin. I am worried this suggests that there is population stratification. Skin neoplasm are strongly linked to skin colour and so population stratification within the included "European samples" seems a possible explanation.

We thank the reviewer for their thorough evaluation of our manuscript. We note that population stratification was conservatively controlled for in the PheWAS and GWAS. For the PheWAS, we utilize ancestry stratified PCA and stringent controls for cryptic relatedness (Supplement p. 7):

"Principal components analysis (PCA) was first performed on the HapMap3 samples, and then MCB samples were projected onto these PCs, with the leading 20 PCs reported. RGC applied a kernel density estimator (KDE) approach to ancestry assignment based on the PCA output. KDEs were trained for each of the individual HapMap3 populations, which were then mapped to one of five main genetic ancestral super-populations (AFR-like=African-like; AMR-like=Admixed American-like; EAS-like=East Asian-like; EUR-like=European-like; SAS-like=South Asian-like) based on specific likelihood criteria for the individual populations such that the likelihood for a given ancestry group was greater than 0.3, the sample was assigned to that ancestry group. When two ancestry groups had a likelihood of greater than 0.3, RGC arbitrarily assigned AFR-like over EUR-like, AMR-like over EUR-like, AMR-like over EAS-like, SAS-like over EUR-like, and AMR-like over AFR-like. We restricted all analyses to include only those determined to be of EUR ancestries.

Cryptic relatedness analysis was performed in multiple steps using PLINK[12] and PRIMUS[12]. IBD sharing estimates were calculated among all individuals with the same ancestral superclass assignment based on a threshold of 0.1875 to capture up to second-degree relationships. A second set of IBD estimates was calculated among all samples based on a cutoff of 0.3 to capture 1st degree relationships that may span ancestral superclasses. Samples were then grouped into first-degree family networks based on the IBD estimates, which were processed using PRIMUS under default settings. Finally, samples with cryptic relatedness were removed from the analysis if they had >100 closely related samples (>0.1875) or >25000 related samples (>0.08), which was performed iteratively until no such samples remained. For each pair with an estimated 2nd degree or higher relatedness, we removed the individual with shorter length of EHR record. An additional

76 participants were excluded due to overlap with the univariate disorder GWAS for either AUD or BD used in this study."

We also note that the original GWAS of the psychiatric disorders themselves performed stringent controls for stratification. For the multivariate GWAS in Genomic SEM, we then perform the additional correction to the univariate GWAS standard errors for any residual population stratification using the univariate LDSC intercept. This means that corrections for population stratification here are both robust and conservative relative to field standards. This is detailed in the method section (p. 27):

"The SEs were additionally corrected for uncontrolled confounds by taking the product of SEs and the LDSC univariate intercept when this value was > 1."

We also want to highlight the part of the Method section that reports results from the attenuation ratio that indicates signal for the multivariate GWAS summary statistics for the psychiatric factors is not contaminated by confounds like stratification (p. 26):

"As a quality control check we confirmed that the attenuation ratio[32] was near 0 for all factors ***(Suppl. Table 17),*** *suggesting that the factor signal is not due to uncontrolled confounds (e.g., population stratification)."*

Association analysis

The main analysis approach here was based on the gSEM results rather than any of the dedicated multi-trait GWAS methods. It would be helpful if the authors explained their choice. I am wondering whether a multitrait GWAS would be complementary to the existing findings and worth adding or even to replace the case-case GWAS findings which do not seem to add much and their role in this paper is not clear to me.

We have previously used simulation to compare multivariate GWAS in Genomic SEM to three other dedicated multi-trait GWAS methods: Multi-trait Analysis of GWAS (MTAG), Model Averaging Genome-wide Association Meta-analysis (MA-GWAMA), and N-weighted Multivariate GWAMA (N-GWAMA). Details of these simulations are provided in pgs. 14-18 of the Online Supplement of [13]. These simulations from prior work revealed that, compared to the three alternative multi-trait methods examined, the multivariate GWAS signal from Genomic SEM was not driven by a single trait and most likely to identify significant findings for the factors when the effect was shared across the disorders that defined the factor. Thus, Genomic SEM is functioning as intended for the purposes of examining shared signal. We do not intend to say that Genomic SEM is always preferable, as some methods aim to boost signal for a specific trait. For example, when the true SNP effect on the target trait was zero, MTAG and MA-GWAMA showed more deflated signals than multivariate GWAS in Genomic SEM. As we are not examining a particular target trait, we continue to use the multivariate GWAS results for Genomic SEM. We also retain CC-GWAS as it offers a unique perspective on what differentiates disorders with high levels of genetic overlap and those results lend insight into what differentiates disorders that load on different psychiatric factors within the Genomic SEM model.

Minor comments

It would be helpful to clarify early on in the results section about sample sizes/power for the different disorders, e.g. a couple of sentences about table 1 would help. This is mentioned eventually in the discussion but would make interpretation easier if clarified sooner.

We thank the reviewer for their suggestion, and we have now included in Table 1 an additional column with $N_{effective}$,better reflecting power of the GWASs included in this study than the overall sample sizes of cases and controls [14]. In the first paragraph of the results, we now include the following statement (p. 4):

"The sample sizes, and hence the power of the disorder GWAS, differed ($N_{effective}$ in Table 1)."

There is also a wider consideration of sample size implications missing in the discussion. For example, the sample size for major depression is much larger than for Tourette's. How does that impact the gSEM results, the discovery of pleiotropic loci etc?

We now write in the Discussion (p. 13):

"Wide ranges in sample sizes across the univariate psychiatric GWAS used as input should also be considered when evaluating relative levels of significant findings, particularly for locus discovery."

Figure 1A weird scale and hard to estimate how high those rg values really are. This figure doesn't indicate which values are actually reliable and statistically significantly different from 0. Maybe they should only show values or add *.

*We have updated Figure 1A to include the *s recommended by the reviewer to denote Bonferroni significant genetic correlations. In addition, the break marks in the heatmap legend are now on a more intuitive scale (i.e., in intervals of 0.25) and the Figure note has been updated to point to the relevant supplementary table to obtain exact values:*

*"Cells depicted with a * denote values that were significant at a Bonferroni-corrected threshold. Exact values are reported in Suppl. Table 1."*

Also, the groupings/SEM results are useful but by no means real distinct categories and some of them don't seem to work as well. I would suggest the authors point that out more explicitly as this will be read by wider audiences and the way the findings and figures are currently presented suggest stronger boundaries between groups than there really are.

We appreciate this point and have added this consideration to the beginning of the newly added genetic correlation analyses (p. 7):

*"We estimated genetic correlations between the five correlated factors, hierarchical p-factor, and 31 complex traits (**Suppl. Table 12)** to place shared genetic liability indexed by the factors in a broader clinical context. These factors vary in their utility for capturing shared genetic signal; accordingly, we used the Q_{Trait} heterogeneity statistic to assess this utility at the genome-wide level. When Q_{Trait} is significant, this indicates a trait's genetic correlation deviates from the factor structure. For instance, if trait X is negatively correlated with SCZ but unrelated to BIP, Q_{Trait} would likely be significant, suggesting trait X lies outside the shared signal captured by the factor. Significant factor correlations were defined at a Bonferroni-corrected threshold of $p < 2.68E-4$, while not significant for Q_{Trait} at this same threshold."*

Figure 2: It is difficult to follow the Mixer results. In those Venn diagrams, the shared grey area is takes up the vast majority of space for all disorder pairings which does not seem very likely to be correct.

To improve focus on our main findings we have now moved MiXeR results to the Supplementary Note. However, in terms of veracity of the MiXeR results, the method had been extensively validated in the primary paper[15], and the results show high polygenic overlap between most psychiatric disorders. While being somewhat surprising, this has been previously pointed out by several studies, albeit with less powered GWAS and a smaller set of psychiatric traits[16,17]. We argue that these results are worth including (but agree that Figure 2 is best suited for the Online Supplement for a general audience).

Figure 5b hard to read, I suggest a different colour scheme or different line type to differentiate the Q ones from the others.

We have updated the colour scheme and now use different line types for the Q results.

The p-factor myopia association is also interesting.

We agree and appreciate the reviewer taking the time to go through the results in detail to point this out. This is the 36th most significant PheWAS association for the p-factor, and given the number of results for the PheWAS we have elected not to specifically highlight this observation in the text.

The authors find 48 novel loci. At the moment the statement is just relative to the specific version of GWAS summary statistics they authors used but these are often just subsets of the data available in the primary GWAS. So the authors should do a literature search for these loci to see if they are truly novel.

We appreciate the reviewer pointing this out and have added a supplementary table that cross-references these 48 loci with those currently described in the GWAS catalogue. We now write on p. 9 of the Results section:

"Forty-eight hits were novel relative to the univariate GWAS, of which 38 have been described in prior GWAS for a broad range of outcomes, and 10 are entirely novel (Suppl. Table 37)."

Only 33 loci with significant Qsnp across the factors: I found that surprising. This links back to the discussion on outcome definitions. However, I am also wondering whether there is a power issue?

It is more plausible that the paucity of Qsnp hits for the factors reflects the high levels of shared signal across the disorders that define the factors. Convergent evidence for this comes from the low levels of residual genetic variance for most disorders in the model, high levels of local genetic correlation for disorders that load on the same factor, and patterns of relationships with external traits that generally conform the factor structure. We do not suspect this reflects outcome definitions as we find more Qsnp hits for the full sample relative to the more carefully ascertained subsample from the newly added sensitivity analyses. We also highlight that the 166 Qsnp hits for the p-factor indicate that Qsnp is capable of detecting heterogeneous effects at current sample sizes. We agree, however, that power is an important consideration more

generally for the Compulsive and Substance Use factor, and as we note in response to the comment about power above, we have updated the Results and Discussion sections to highlight this interpretive limitation.

"eQTL and Hi-C Datasets to Identify Target Genes" is a somewhat misleading title. This section is not about target genes at all. I understand the authors used a simple approach to link loci to genes which were then used for all these enrichment analyses but that is more of a technical detail and I think the title of the section is confusing.

We agree that this was not an accurate description of that section of the results and have changed the title to be more descriptive with respect to the findings that follow:

"Enrichment Analyses Implicate Early Neurodevelopmental Processes and Specific Cell Types."

The pathway analysis had no findings for the internalizing factor. This is quite surprising and not further explained.

This is likely due to the fact that we aimed to enhance specificity of results by removing overlapping target genes across the internalizing factor and p-factor. We have edited the text to directly state this important interpretive consideration (p. 10):

"To enhance the specificity of the gene sets, we removed Internalizing disorders and SB target genes that also appeared for the p-factor. SB (minus p-factor) target genes were enriched in more specific terms related to neuron development. No significant results were identified for the Internalizing disorders factor, likely reflecting the large proportion of target genes overlapping with the p-factor."

The link to rare mutations in genes for neurodevelopmental disorders is very interesting. It would be important to clarify whether this is just driven by autism. The papers they are referring to are both specifically on ASD and that was also one of the disorders included here.

We have amended the text to clarify that enrichment was found for both ASD and NDD (p. 10):

*"Results from MAGMA[48] (**Supplemental Method**) provided convergent support for the role of early neurodevelopmental processes in transdiagnostic psychiatric risk. Specifically, genetic signal for the five correlated factors and p-factor showed enrichment in genes identified from rare variant studies of ASD[49–51], neurodevelopmental delay[49], or both (**Suppl. Fig. 40**)."*

For the interpretation of enrichment analyses, especially some of the genomic annotations, it would be very helpful to learn whether these are findings you would generally see for complex traits or whether any of this is specific to psychiatric outcomes.

We have now added results from a recent GWAS of height to evaluate specificity of our findings. As the reviewer may have expected, we found certain annotations to be significantly enriched and show similar patterns across both psychiatric factors and height. This included results for the evolutionarily conserved annotations and the minor allele frequency bins. We have correspondingly updated Figure 6 to remove those results. Supplementary Table 48 now includes

height results for all annotations and the updated Figure 6 includes an additional bar for height enrichment. These results show that many of our neuronal annotations that are significantly enriched for a psychiatric factor (e.g., excitatory neurons for the Schizophrenia/Bipolar factor) are, reassuringly, not significant for height. We now note in the Results section of the main text (p. 11):

*"Enrichment was also calculated for a recent GWAS of height[58] to evaluate specificity of the psychiatric findings. We employed a Bonferroni-corrected significance threshold of $p < 2.8110^{-5}$ (**Method**). We focus here on results for the better-powered SB, Internalizing, and p-factor and do not discuss annotations that lacked psychiatric specificity, as indicated by significant enrichment for height (e.g., evolutionarily conserved annotations)."*

Conclusion: "While much remains to be done, cross-disorder genetics continues to fill in critical gaps in our understanding of shared and unique psychiatric risk with implications for constructing a more biologically valid psychiatric nosology." The paper had not touched on psychiatric nosology much. If the authors belief this is the case, then they need to discuss this more. Should we drop MDD, anxiety etc in favour of a broad group of internalizing conditions now? Else, what is being implied here?

This was admittedly vague and we appreciate the chance to think critically about what it would look like for these kinds of genetic findings to inform nosology. The decision making around the response here was led by co-supervisor on the project Dr. Kenneth Kendler, who served as the chair of the scientific review committee for DSM-5 and is uniquely suited to think through how much these results should inform nosology. The decision was that given that genetics is one of many validators considered in formal nosology, that any formal recommendations for nosology would need to come from a range of empirical sources, alongside expert input outside of genetics. For this reason we have removed this phrase about constructing biologically valid nosology and instead focus on the etiological implications that are more directly linked to our findings.

Referee #2 (Remarks to the Author):

Applying genome-wide structural models to these psychiatric datasets has tremendous value because it provides a relatively unbiased view of shared genetic influences across disorders. By adding disorders that could not be meaningfully analyzed in previous iterations due to low statistical power (e.g. PTSD, anorexia), the authors have provided important additional information beyond their prior publications. Indeed, these results can be used to ground the field's understanding of comorbidity patterns in a firm biological foundation. My primary recommendation for revision is that the work with cell types should be improved, principally with the use of whole brain cell type data (e.g. from Siletti et al 2023). Below are two important comments and other suggestions.

IMPORTANT:

1a. In contrast to the strength of the genomic SEM work, the functional analyses are more crude, and specifically the choice of single cell and single nucleus RNA sequencing datasets is not good enough. There exists one human whole-brain atlas of cell types (Siletti et al 2023 Science) and two whole brain mouse atlases (Yao et al 2023 Nature, superseding Zeisel et al. 2018 Cell). The authors have instead used atlases of cell types that are much more restricted. There is no reason to do this, and doing so may miss important cellular associations. Thus,

additional analyses need to be run before making conclusions like "SB factor was substantially enriched in genes expressed in excitatory neurons, whereas the internalizing factor was associated with oligodendrocyte biology." Further, the cell type analyses focused on rather crude categories of cell types (e.g. only 8), when in fact there are at least hundreds and likely thousands of cell types in the human brain. Moreover, the more granular cell types offer much stronger associations with psychiatric phenotypes, even accounting for additional statistical test conducted on more granular cell types. e.g. it may be best to analyze data at the cluster level (n=461) from Siletti et al. vs supercluster level (n=31; too crude) or subcluster level (n=3,313; too noisy).

Thank you for this helpful suggestion, which allowed us to improve the functional single cell enrichment analyses. We now have incorporated Siletti 2023[18] into both main figure (Figure 5) and supplementary enrichment analyses (Supplementary Figure 41). For the new single cell analyses from the Siletti dataset, we chose to use the supercluster level broken down by tissue type (e.g. cortex, midbrain, hindbrain, hippocampus, etc). This increased granularity has prompted new enrichment associations. For example, the Internalizing disorder factor is enriched in MGE interneurons in the Siletti dataset in the cortex, which was not captured using the previous datasets, perhaps due to the reduced granularity of 'interneurons' compared with distinct categories of interneurons including MGE, CGE, and LAMP5. We have opted not to use the whole brain mouse atlases as we felt that, with the availability of the human whole-brain atlas the reviewer helpfully pointed out, that Siletti 2023 would be the most appropriate. We now write in the main text (p. 10):

"Expression weighted cell type enrichment (EWCE)[52] was estimated to examine enrichment within neuronal cell types in fetal and adult single-cell datasets[53–57]. Genes associated with the SB factor were significantly enriched within excitatory neuron subtypes and interneurons in fetal brain[53,54] and deep-layer excitatory neurons in adult brain[57] (Fig. 5). Internalizing disorder genes were enriched within excitatory maturing neurons in fetal data[53] and medial ganglionic eminence (MGE) interneurons in the adult cortex[56]. Internalizing signal was also enriched for different glial cells, specifically oligodendrocytes in fetal and adult datasets[53,57] and Bergmann glia in adult data[56]. The p-factor was enriched for oligodendrocyte precursor cells in adult data[56]. A significant proportion of these genes are expressed during both fetal and adult stages; cell type enrichment was largely driven by genes that are not expressed in a particular developmental stage (Suppl. Fig. 41). We also tested enrichment for loci specific to MD and SCZ identified from CC-GWAS. MD-specific signal was enriched for cycling and intermediate progenitors in fetal brain. SCZ-specific signal was enriched for endothelial, vascular, and upper rhombic lip cells in adult brain (Suppl. Fig. 41)."

1b. Relatedly, this statement oversimplifies results considerably, and it should be improved after additional analyses: "At the functional level, we identified consistent results across different methods that highlighted the relevance of excitatory neurons, specifically hippocampal CA1 and CA3 neurons, for the SB factor and oligodendrocytes for the Internalizing disorders factor."

We agree that this did not capture the nuances of the functional findings. We continue to focus on findings replicated across methods in order to maintain the brevity of the manuscript, but have updated this section to provide additional details and note the new findings from the added Silleti 2023 data. We now write in the Discussion (p. 11):

"A replicated finding across functional methods reflected enrichment for the SB factor in excitatory neuron annotations, including CA1 and CA3 hippocampal neurons, deep layer neurons from adult data, and maturing, migrating, prefrontal, and visual cortex excitatory neurons in fetal data. The Internalizing factor showed minimal enrichment in fetal datasets and was more consistently enriched in different glial cells in adult data, including oligodendrocytes and their precursor cells and Bergmann glia."

IMPORTANT but not critical for interpretation

2. Line 1052 ERROR: using LDSC with admixed ancestry (African American datasets here) violates assumptions in an important way. Estimate is also way off (47%!). This has to be removed. It could be replaced with GCTA on the individual level data.

We acknowledge that applying LDSC to admixed ancestry, such as African-like genetic ancestry datasets, is suboptimal due to its assumptions. Given these limitations, we have explored two different LD reference panels for LDSC to assess their impact on our results, reflecting African-like ancestry or admixed American ancestry scores provided through Pan UKB for use in LDSC. We have updated the results reported in Table 1 to reflect those produced using the admixed ancestry LD-scores, which produce more reasonable estimates, including an estimate of 22% heritability for AUD. At the reviewer's or editor's discretion, we are willing to omit the African ancestry analyses from the manuscript, but retain them for now for the reasons highlighted in response to reviewer 1 point 1. We revised online methods and acknowledged that our findings must be interpreted with caution, but our approach provides a reasonable alternative in the absence of perfect methods and data (p. 27):

*"We acknowledge that using LDSC with admixed ancestry may violate its assumptions, thus our results for AFR-like ancestry should be interpreted with caution. With this in mind, we performed LDSC for AFR-like datasets using two different LD reference panels for AFR-like ancestry or admixed American ancestry from Pan UKB to assess their impact on results (see **Suppl. Table 4**). Results in Table 1 report liability-scale heritabilities for AFR-like datasets using the admixed LD-scores as these produced more sensible results."*

Top suggestions:

3. ABSTRACT: The abstract says a few things that are, in my understanding, incorrect about historical debates about nosology in psychiatry. For example, the article states: "Genomic methods have shown that even for schizophrenia and bipolar disorder, two disorders long-thought to be etiologically distinct3" But this is very much not true, it has long been debated as to whether or not these were the same, similar, or distinct conditions. Instead, the genetic results here beautifully capture points of greater certainty versus debate in the history of psychiatric nosology. In other words, where debate has endured over the decades, we see that genetic correlations are highest; therefore there is good biological reason for the longstanding debates. In contrast, few clinicians have argued, for example, that ASD and bipolar disorder are the same, and indeed they are more genetically distinct.

We appreciate the reviewer pointing this out and agree that the abstract as previously written did not appropriately characterize historical debates around nosology. The abstract has been rewritten to reflect this:

"Psychiatric disorders display high levels of comorbidity and genetic overlap[1,2], challenging current diagnostic boundaries. For disorders whose diagnostic separation has been most debated, like schizophrenia and bipolar disorder[3], genomic methods have revealed that the majority of genetic signal is shared[4]."

4. FIG 2: Do the authors need to use MIXER? I've always found it to have lots of problematic elements... from an expert statistical geneticist, "Mixer relies on an unrealistic model with a mixture of null SNPs (exactly zero effect) and causal SNPs with normally distributed effects. We showed that this type of approach leads to estimates that are sample size dependent - the estimated polygenicity increases rapidly as a function of sample size, as more causal SNPs are detected." An alternative: S-LD4M.

We appreciate critical comments from the reviewer and provide our detailed response below. We have now moved MiXeR results to the Supplementary Note in order to focus on our main findings. However, we emphasize that MiXeR model yields appropriate results, and reviewer's claim about "estimated polygenicity increases rapidly as a function of sample size" is not justified, as we show below using MiXeR's polygenicity estimates of schizophrenia using two GWAS studies: PGC3[19] and PGC2 [20]. As PGC2 cohorts form a subset of PGC3 cohorts, we also performed custom meta-analysis - PGC3_only- including only new cohorts from PGC3. Despite PGC2 having 2.5 times larger effective sample size than PGC3_only the polygenicity estimate from MiXeR was largely consistent.

GWAS of schizophrenia	N cases / N controls / Neff	nc@p90 polygenicity estimate, mean (std)
PGC3 (EUR cohorts, Trubetskoy el al, 2022)	53,386 / 77,258 / 126,281	11800 (200)
PGC2 (EUR cohorts, Ripke et al, 2014)	33,640 / 43,456 / 75,846	10500 (300)
PGC3_only (new cohorts only, exclude PGC2, exclude Pardiñas et al. 2018[21])	13,033 / 18,391 / 30,510	10700 (1700)

We note that MiXeR does not attempt to identify and then count specific causal variants (as e.g. in fine mapping). Instead, MiXeR postulates a statistical model where polygenicity is represented by a single parameter, and estimates this parameter from GWAS summary statistics. This is similar to how genetic correlation (rg) between two traits is estimated by LDSC, using underlying model which includes rg as a parameter, and bypassing the need of accurately estimating the effect sizes ("beta's") of all SNPs in two traits, and then computing rg as a correlation coefficient between "beta's" across the genome.

We agree that for real traits MiXeR's mathematical model is likely violated to a certain degree, and the genetic architecture of complex traits is more complex than just a mixture of null SNPs with exactly zero effect, and non-null SNPs following normally distributed effects. To address this concern, MiXeR's model was tested under several scenarios of model misspecification[15], including MAF-dependent genetic architectures, and presence of functional enrichments, showing that resulting estimates of polygenicity remain reasonably stable.

To follow up on the reviewer's suggestion of using S-LD4M, we apply the S-LD4M model to estimate polygenicity (Ma). We use the original configuration of the S-LD4M model (which includes a set of functional annotations),and a "baseline" model (LD4M) which only includes L2

and L4 scores for the "base" category (all SNPs), excluding functional annotations. Due to limiting strength of the GWAS signal for OCD, OUD and TS we observed large uncertainty in Ma estimates from S-LD4M. Across other traits, polygenicity estimates from MiXeR and S-LD4M appear to be reasonably consistent with each other (Pearson's correlation coefficient r=0.81).

We also note that S-LD4M method only provides univariate estimates of polygenicity, and does not have an equivalent cross-trait extension (we're aware of S-LD4M's bivariate extension for causal inference[22], but this analysis is different in scope from MiXeR's polygenic overlap estimates).

TRAIT	MiXeR	LD4M	LD4M_S
ADHD	9000 (400)	9900 (800)	16500 (1200)
AN	8600 (500)	9100 (1000)	11000 (1600)
ANX	10600 (200)	12500 (900)	15500 (1600)
ASD	16400 (1400)	11000 (1500)	18700 (4100)
AUD	7100 (1400)	6000 (1300)	3500 (3200)
BIP	11000 (300)	12500 (700)	17500 (900)
CUD	5300 (500)	9200 (1700)	12400 (2500)
MD	13700 (200)	14400 (800)	20000 (700)
NIC	2600 (400)	5800 (1600)	5600 (1500)
OCD	21300 (3200)	23200 (4200)	57400 (19900)
OUD	17700 (11400)	15800 (10600)	86800 (206100)
PTSD	7900 (300)	11100 (1000)	11600 (1000)
SCZ	11800 (200)	11700 (800)	16400 (800)
TS	18900 (9300)	14800 (4000)	117400 (232800)

Rebuttal Letter Fig 1. *Comparison of the polygenicity estimates: nc@p9 from MiXeR, M4 from LD4M and LD4M_S models). LD4M_S is the original model[23]. LD4M is based only on total LD score, without modeling the set of functional annotations.*

Just a compliment:
FIG 1B: Really love the use of the pie chart for each of the disorders, to show when the fit the factor well (many disorders) as opposed to disorders that map nearly evenly onto two factors (ADHD), or multiple factors but with one primary factor (TS), and when disorder specific factors are important. If complaints are raised about pie charts (per some journal guidelines), I would push back because this is such an informative and clear usage of pie charts.

Thank you! We will do our best to keep it in.

Referee #3 (Remarks to the Author):

This study presents a thorough investigation of the shared and divergent genetic architecture of 14 psychiatric disorders. The use of multiple, complementary methods provides a rigorous approach to understanding the genetic relationships among these disorders. A major finding

is the strong genetic clustering of bipolar disorder (BIP) and schizophrenia (SCZ) into one factor, and major depression (MD), PTSD, and anxiety disorders (ANX) into another. This work contributes to a better delineation of psychiatric nosology and lays the groundwork for future research into the biological underpinnings of these disorders.

Comments:

1. It would be helpful to clarify what this study adds beyond Grotzinger et al. (2022, Nature Genetics), which analyzed 11 major psychiatric disorders with similar methodologies and overlapping findings. Are there methodological improvements, additional insights, or novel interpretations that distinguish this study?

We have now updated Supplementary Figure 1 to include sample size comparisons between CDG3 and the Grotzinger et al. 2022 paper. From a sample size perspective alone, there are significant increases for seven out of the 11 disorders from the Grotzinger et al. paper (ADHD, MD, OCD, BIP, ANX, PTSD, and AUD) along with three entirely new disorders (OUD, CUD, NIC). The inclusion of the three new disorders in the substance use space allows for pulling out an SUD factor that maps onto extant models of psychopathology (as opposed to the model from the 11 disorders paper where AUD cross-loaded across multiple factors). These sample size increases also result in an increase in discovery, with greater GWAS hits also facilitating downstream analyses (e.g., EWCE) that require a sufficient level of loci as input to produce interpretable results. This increase in sample sizes and disorders is now noted in the Results (p. 4):

"The three substance use disorders are novel relative to a more recent cross-disorder analysis[26], and sample size increases were significant for previously included disorders (average case increase ~287%)."

The largest distinguishing element is that this reflects a multi-method approach that allows for triangulating across findings to identify the most robust results. In addition, the developers of the utilized methods all ran the corresponding analyses and met regularly to discuss interpretation of results across methods. More specifically, the Grotzinger et al. paper did not include LAVA, MiXeR, CC-GWAS, Popcorn, diverse genetic ancestry analyses, or the extensive functional annotation performed here. We have added the following to the Discussion to help briefly clarify the value of this work relative to prior efforts (p. 13):

"It extends prior cross-disorder psychiatric genetics work[5,26] using updated datasets, new disorders, and triangulation across different approaches to identify robust sets of findings[69]."

Table 1:
2a. Ordering the disorders in Table 1 by liability-scale heritability rather than alphabetically would provide a more informative structure.

Table 1 has been updated to be ordered by liability-scale heritability; we agree that this provides an additional layer of useful information with respect to relative levels of common variant signal.

2b. The table lacks an explanation for why Nicotine Dependence does not have case-control counts.

We have added the following to the table note and appreciate the reviewer pointing out the absence of information here:

"Nicotine dependence includes a single value for the sample size columns as this was the one continuous measure, defined using the Fagerström Test for Nicotine Dependence."

2c. An additional column for effective sample sizes would be valuable for understanding the relative power of each GWAS.

We agree (and this is also in line with comments from Reviewer 1 about making power differences more explicit); the effective sample sizes have been added to Table 1.

The p-Factor:

3. The p-factor may not solely reflect genetic risk for psychiatric conditions. It could also reflect one's (in)ability to have a healthy emotional reaction to adverse environmental circumstances or perhaps even the chance of being exposed to adverse environments. Could the p-factor be capturing neuroticism and/or SES? A genetic correlation between p and a range of potential outcomes (e.g., educational attainment, income, neuroticism) could help clarify this. If the p-factor was just another trait that is highly correlated with multiple disorders, that would explain the relative high number of QSNP hits.

3a. A genetic correlation analysis between p and a range of relevant complex traits (e.g., educational attainment, income, neuroticism) would provide useful context.

We have now added genetic correlations between the p-factor, the five correlated factors, and 31 external traits (including EA, income, and neuroticism). These results revealed genetic correlations with varying degrees of specificity, with some outcomes specific to a single factors (e.g., childhood BMI for the Neurodevelopmental factor), to a pair of factors (e.g., occupational attainment for Internalizing disorders), or shared across all five factors (e.g., loneliness). As the reviewer hypothesizes, neuroticism was transdiagnostically associated with all five factors; however, since the pattern of correlations with neuroticism did not align with those implied by the p-factor model (as indexed by the Qtrait heterogeneity metric) this is not identified as significant for the p-factor. We now write on p.7 of the Results:

*"Genetic Correlations with Factors. We estimated genetic correlations between the five correlated factors, hierarchical p-factor, and 31 complex traits (**Suppl. Table 12)** to place shared genetic liability indexed by the factors in a broader clinical context. These factors vary in their utility for capturing shared genetic signal; accordingly, we used the Q_{Trait} heterogeneity statistic to assess this utility at the genome-wide level. When Q_{Trait} is significant, this indicates a trait's genetic correlation deviates from the factor structure. For instance, if trait X is negatively correlated with SCZ but unrelated to BIP, Q_{Trait} would likely be significant, suggesting trait X lies outside the shared signal captured by the factor. Significant factor correlations were defined at a Bonferroni-corrected threshold of p < 2.68E-4, while not significant for Q_{Trait} at this same threshold.*

Neuroticism, stress sensitivity, self-harm, suicide attempt, and self-reported loneliness were associated with all five factors from the correlated factors model (Fig. 2), but were not identified as significant for the p-factor (due to significant Q_{Trait} heterogeneity estimates). Daytime napping was the only trait significantly correlated with the p-factor (r_{g_p} = .31, SE = .03). The Internalizing disorders factor was the only factor correlated with occupational attainment (rg_Int = -.29, SE = .02). The Internalizing disorders and SUD factors were the only factors associated with household income (r_{g_Int} = -.40, SE = .02; r_{g_SUD}= -.41, SE = .03) and were the most pervasively

*associated with different cognitive outcomes, including childhood intelligence (r_{g_Int} = -.27, SE = .05; r_{g_SUD}= -.40, SE = .07). Only the SUD factor was associated with adult intelligence (r_{g_SUD} = -.40, SE = .03) and verbal numerical reasoning (r_{g_SUD} = -.41, SE = .03). This was compared to more circumscribed cognitive associations for the Compulsive disorders and SB factors, including a large negative correlation with the pairs matching test (potentially indexing memory; $r_{g\_Compulsive}$ = -.33, SE = .03; r_{g_SB} = -.34, SE = .03). The SB and SUD factors were the only ones associated with risk tolerance (r_{g_SB} = .31, SE = .03; r_{g_SUD} = .38, SE = .03). The Neurodevelopmental factor was uniquely associated with childhood BMI (r_{g_Neuro} = .26, SE = .06) and showed high genetic overlap with childhood aggression (r_{g_Neuro} = .94, SE = .10). For the full set of correlations see **Suppl. Table 13**, for comparison across factors **Suppl. Figs. 20-25**, and for comparison across traits within each factor **Suppl. Figs. 26-31**."*

3b. The high number of QSNP hits for the p-factor suggests it may capture general behavioral/psychological variation rather than specific psychiatric risk.

We think this is an important question that asks whether the p-factor works better at some levels (genome-wide) than others (locus level). The current analyses indicate that the p-factor, as specified here, does not capture even these more general processes. We now consider this in the Discussion section (p. 12):

"At the genome-wide level, the p-factor was strongly related to the Internalizing disorders factor. This is consistent with conceptualizations of the p-factor as reflecting a general tendency towards negative emotionality.[60] In support of the p-factor, LAVA identified pleiotropic hotspots characterized by widespread local r_g across disorders and multivariate GWAS yielded 160 hits for this factor alone. However, the p-factor also had more hits for the Q_{SNP} heterogeneity metric (117) than all five-factors from the correlated factors model (33) and was significantly correlated with daytime napping only, indicating the p-factor alone is insufficient for explaining cross-disorder risk. The fact that many traits showed significant associations with pairs of factors suggests that the genetic correlations across factors may often reflect liability processes unique to particular disorder clusters, rather than shared across psychopathology as a whole. The p-factor was largely enriched for broad biological categories, such as gene regulation. These results suggest a conceptual model in which there is a partial, broadly transdiagnostic component of genetic vulnerability to psychiatric disorders that primarily captures internalizing genetic signals, with subsequent levels of more canalized and neurobiologically meaningful subdomains of psychopathology captured by the five factors."

4. The pleiotropic hotspot on chromosome 11 raises the question of whether this region is a general genetic hotspot for psychiatric disorders, behavioral outcomes, or complex traits more broadly. Comparing this region's effects with those in GWAS of intelligence, personality, or addiction could clarify its broader relevance.

We agree with the point raised by reviewer, and have made this more explicit in the relevant results section for LAVA on page 7:

"Notably, this region contains the NCAM1-TTC12-ANKK1-DRD2 gene cluster that has been frequently associated with psychiatric phenotypes,[36–39] and flagged as a likely pleiotropy hotspot for a range of cognitive and behavioural outcomes related to, for example, intelligence, personality, substance use, and sleep[35,40–42]."

5. Given the marked differences between ancestries in genetic correlations between disorders, I think it should be emphasized (probably also in the abstract) that the underlying factors that are highlighted in this paper could very well differ between ancestries, further complicating what we are trying to achieve here, namely getting a better delineation of what the different psychiatric disorders are on a genetic level and how they differ.

We share the reviewer's concern about generalizability of our findings to individuals beyond EUR-like ancestry. At the same time, we emphasize that our cross-ancestry results are still largely inconclusive due to insufficient sample size, and/or broadly defined ancestral groups. We have now revised our manuscript to reflect these limitations of our cross-disorder analyses. We have updated the Results to read (p. 5):

"While these results suggest that the findings that follow for EUR-like ancestry groups may generalize better for some disorders (e.g., SCZ) than for others (e.g., PTSD, MD), that conclusion awaits replication in more highly powered analyses."

We also now write in the Discussion (p. 12):

"Our study has several limitations. Analyses were restricted primarily to European-like genetic ancestry populations due to the limited availability of GWAS data for other groups and the limitations of methods requiring more genetically homogeneous groups[61]. GWAS for non-EUR-like populations sample sizes are still orders of magnitude smaller and not currently powered for more precise cross-ancestry assessments; this emphasizes the need for future research including the generation of additional ancestrally representative data, which will allow well-powered studies and the examination of cross-disorder genetic architecture across regional and cultural differences."

6. Page 5, lines 215-219: "Of note, previously reported within-disorder, within-ancestry rg's calculated across cohorts are considerably smaller for PTSD (rg = 0.73, SE=0.21)35 and MD (rg = 0.76, SE =0.03)36 than SCZ (rg = 0.95, SE = 0.03)37, suggesting that cross-ancestry rg's for these two disorders reflect a combination of etiological and phenotypic heterogeneity and ancestry-specific signal." PTSD and MD are also disorders that may be more influenced by life events, perhaps worth adding that environmental and cultural differences between populations may influence which underlying traits (and thus genetic effects) are associated with these disorders across populations.

We agree that this is one possible interpretation and have updated this text, which we now put in the limitation section of the Discussion (p. 13):

"Cross-ancestry r_g's should be interpreted in light of findings that show considerably smaller within-disorder, within-ancestry r_g's across cohorts for PTSD (r_g = 0.73, SE=0.21)[62] and MD (r_g = 0.76, SE =0.03)[63] relative to SCZ (r_g = 0.95, SE = 0.03)[64]. This suggests cross-ancestry r_g's for PTSD and MD could drop below 1 for reasons independent of ancestry-specific signal, such as environmental moderation of genetic effects or increased phenotypic heterogeneity."

7. The discussion states: "First, at a genome-wide level, we confirm pervasive genetic overlap across 14 clinically distinguished psychiatric disorders, as indicated by large pairwise rg across multiple genetic ancestries." However, the results show that genetic correlations differ

significantly across ancestries. This statement should probably be toned down to reflect these findings more accurately.

We agree, and have revised the sentence in the Discussion to now read (p. 11):

"At a genome-wide level, we confirmed pervasive genetic overlap across 14 clinically distinguished psychiatric disorders, as indicated by large pairwise r_g within the EUR-like genetic ancestry group and even greater overlap when including loci that are shared, but have divergent directional effects."

Shared loci between disorders:
8. Page 6, lines 276-279: "Consistent with the multivariate genetic structure identified using Genomic SEM, the pairs of disorders with the greatest number of significant local rg hits were MD and ANX (113 rg loci), MD and PTSD (88 rg loci), and BIP and SCZ (40 rg loci), accounting for just over half of all significant local rg's detected (Fig. 3A)." Why would MDD and ANX/PTSD have more shared regions be significant than BIP and SCZ, while the latter has both higher rg between each other and higher heritabilities?

We thank the reviewer for their comment, and apologise for any confusion regarding the presentation of our results. We realise that the MiXeR results in the previous Fig 2 (now Supplementary Figure 6) may have given the impression that BIP-SCZ exhibited the greatest degree of genetic overlap; however, we note that MD-PTSD actually have both the highest estimated r_g according to LDSC and number of pleiotropic SNPs according to MiXeR (although the model fit for MD-PTSD was insufficient and are therefore excluded from Supplementary Figure 6). To prevent further confusion, we have made the concordance between the LAVA results and those of LDSC and MiXeR more explicit in the text (p. 8):

*"This is consistent with the genome-wide levels of overlap indicated via the LDSC global r_g (Fig. 1), the polygenic overlap estimated with MiXeR (**Suppl. Figs. 5-10**), and the multivariate genetic structure identified by Genomic SEM."*

More Genetic Correlations:
9. I think there is a missed chance to get a better sense of the content of the signal, more so than diving into the biology. These are all behavioral outcomes, which are a composite of lower level outcomes and environmental influences (Abdellaoui & Verweij 2021). Looking at rg patterns with other complex traits could give us a much better idea of the content of the signal. This could clarify whether certain factors (or p-factor) might reflect something else than just psychiatric genetic risk; it's likely that normal dimensions of variation are captured here. The Phewas does not seem sufficient, as that only considers disease traits and doesn't give a quantitatively interpretable unit of genetic overlap, only whether it's significantly associated or not. Adding rg analyses with a wider range of complex traits is a relatively easy analysis to add, much easier than the analyses already done in this paper.

We refer to our response to reviewer 3 point 3 above for more details about these analyses, but note here that we fully agree that this adds a useful context and specifically note in the added genetic correlation section (p. 7):

*"We estimated genetic correlations between the five correlated factors, hierarchical p-factor, and 31 complex traits (**Suppl. Table 12)** to place shared genetic liability indexed by the factors in a broader clinical context."*

Power differences:

10. How do differences in statistical power across disorders affect the interpretation of shared genetic factors? Are there sensitivity analyses that could address this?

We thank the reviewer for their suggestion, and we have now included in Table 1 an additional column with $N_{effective}$,better reflecting power of the GWASs included in this study than the overall sample sizes of cases and controls[14]. In the first paragraph of the results (p. 4), we now include the following statement (Also see our response to the first minor comment of reviewer #1.):

"The sample sizes, and hence the power of the disorder GWAS, differed ($N_{effective}$ in Table 1)."

In our revised manuscript, we also include sensitivity analyses using stricter case definition with decreasing in sample size ranging 1% - 48% across six disorders (Table SX1). The results are very similar to our main findings, confirming the statement above that genome-wide analyses are well-powered at current psychiatric GWAS sample sizes. (Also see our response to the comment of reviewer #1 on outcome definition).

11. Figure 1 is beautiful, well-designed and clear.

Thank you!

12. I find Figure 2 a bit more difficult to interpret (and I'm not sure it warrants being a main Figure?).

We have now moved Figure 2 to the Supplementary Note and have added additional interpretive guidelines for both LAVA (see response directly below) and MiXeR in the main text to help the reader understand these results. We now write for MiXeR (p. 6):

"Genome-wide r_g's from LDSC indicate shared genetic risk across psychiatric disorders. However, LDSC may underestimate the extent of genetic overlap if shared causal variants reflect a mixture of directionally concordant and discordant associations. We applied bivariate causal mixture modeling (MiXeR) to quantify the degree of genome-wide polygenic overlap reflecting the total number of shared causal variants regardless of magnitude or directionality[33]."

13. The LAVA and Q-SNP results are among the more complex parts of the study. If possible, adding a clearer explanation of their use and significance (especially LAVA) would help readers unfamiliar with these approaches.

We have substantially updated the results section for LAVA (pgs. 6-7) to convey the use and significance of this approach more clearly:

"LAVA Analyses Reveal Regional Hotspots of Genetic Overlap

Global estimates of pleiotropy, such as the genome-wide r_g's from LDSC, provide an average of the degree of shared signal across the genome. However, because genetic overlap is

*unlikely to be constant across genomic regions, we segmented the genome into 1,093 LD-independent regions, and applied local analysis of [co]variant association (LAVA[35]; **Method**) to assess the rg between disorders within these regions. In addition to capturing heterogeneity in genetic overlap and pinpointing relevant regions, LAVA identifies potential rg hotspots shared among several disorders, thereby providing further insight into genetic architecture.*

*We restricted analyses to loci with sufficient SNP-based heritability for the disorders analyzed ($p < 4.610^{-5} = 0.05/1,093$; **Method**). Correcting for the number of bivariate tests performed across all regions and disorder pairs, we detected 458 significant pairwise local r_g's ($p < 2.110^{-6} = 0.05/24,273$). The pairs of disorders with the greatest number of local r_g hits were MD and ANX (113 regions), MD and PTSD (88 regions), and BIP and SCZ (40 regions), accounting for over half of all significant local r_g's detected (**Fig. 3A**). This is consistent with the genome-wide levels of overlap indicated via the LDSC global r_g (Fig. 1), the polygenic overlap estimated with MiXeR (**Suppl. Figs. 5-10**), and the multivariate genetic structure identified by Genomic SEM. Both global and local r_g's tended to be positive, with significant negative r_g's identified in only three instances (**Suppl. Fig. 32**). This indicates that the genetic risk for one disorder typically increases the risk for another (**Suppl. Fig. 33**).*

*We detected 101 regions that harbored significant local r_g's between several disorder pairs, which we call "r_g hotspots" (**Suppl. Tables 14-23** for local r_g across disorders in the top 10 hotspots). The most pleiotropic of these hotspots was on chromosome 11, which contained 17 positive and significant local r_g's involving 8 of the 14 analyzed disorders (**Fig. 3B**). This region also stands out as the most significantly associated with 8 of these 17 disorder pairs, while ranking in the top 25% of associated loci for 12 of them (**Suppl. Fig. 34**). Notably, this region contains the NCAM1-TTC12-ANKK1-DRD2 gene cluster that has been frequently associated with psychiatric phenotypes,[36–39] and flagged as a likely pleiotropy hotspot for a range of cognitive and behavioural outcomes related to, for example, intelligence, personality, substance use, and sleep[35,40–42]."*

For Q, we now write a more detailed explanation for the statistically identical QTrait metric that is now included for the genetic correlation analyses (p. 7):

"These factors vary in their utility for capturing shared genetic signal; accordingly, we used the Q_{Trait} heterogeneity statistic to assess this utility at the genome-wide level. When Q_{Trait} is significant, this indicates a trait's genetic correlation deviates from the factor structure. For instance, if trait X is negatively correlated with SCZ but unrelated to BIP, Q_{Trait} would likely be significant, suggesting trait X lies outside the shared signal captured by the factor."

14. It should be noted in a prominent place, I suggest the abstract, intro, and discussion (perhaps even the title) that this study considers *common* genetic variants only.

We agree that this is critical for interpretation. We now note this in the Abstract:

"Using genetic association data from common variants, we identified and characterized five underlying genomic factors that explained the majority of the genetic variance of the individual disorders (~66% on average) and were associated with 238 pleiotropic loci."

We go on to write in the Introduction (p. 4):

"Here we examined the shared and unique influences of common genetic variants across 14 psychiatric disorders."

And finally note in the first sentence of the Discussion (p. 11):

"Our analyses characterized the landscape of shared and divergent genetic influences of common variants on 14 psychiatric disorders."

Review References

1. Walters, R. K. *et al.* Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656 (2018).
2. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
3. Nievergelt, C. M. *et al.* Genome-wide association analyses identify 95 risk loci and provide insights into the neurobiology of post-traumatic stress disorder. *Nat. Genet.* (2024) doi:10.1038/s41588-024-01707-9.
4. Chen, C.-Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).
5. Cai, N. *et al.* Assessment and ascertainment in psychiatric molecular genetics: challenges and opportunities for cross-disorder research. *Mol. Psychiatry* **30**, 1627–1638 (2025).
6. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.* 1–11 (2020).
7. Kendler, K. S., Chatzinakos, C. & Bacanu, S.-A. The impact on estimations of genetic correlations by the use of super-normal, unscreened, and family-history screened controls in genome wide case-control studies. *Genet. Epidemiol.* **44**, 283–289 (2020).
8. Consortium, ReproGen *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
9. Peyrot, W. J. & Price, A. L. Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. *Nat. Genet.* **53**, 445–454 (2021).
10. Wray, N. R., Lee, S. H. & Kendler, K. S. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *Eur. J. Hum. Genet.* **20**, 668–674 (2012).
11. Hirschtritt, M. E. *et al.* Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome. *JAMA Psychiatry* **72**, 325–333 (2015).
12. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
13. Grotzinger, A. D. *et al.* Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis. *Nat. Genet.* **54**, 548–559 (2022).
14. Grotzinger, A. D., Fuente, J. de la, Privé, F., Nivard, M. G. & Tucker-Drob, E. M. Pervasive Downward Bias in Estimates of Liability-Scale Heritability in GWAS Meta-Analysis: A Simple Solution. *Biol. Psychiatry* (2022) doi:10.1016/j.biopsych.2022.05.029.
15. Frei, O. *et al.* Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat. Commun.* **10**, 1–11 (2019).
16. Hindley, G. *et al.* Charting the landscape of genetic overlap between mental disorders and related traits beyond genetic correlation. *Am. J. Psychiatry* **179**, 833–843 (2022).
17. Smeland, O. B., Frei, O., Dale, A. M. & Andreassen, O. A. The polygenic architecture of schizophrenia - rethinking pathogenesis and nosology. *Nat. Rev. Neurol.* **16**, 366–379 (2020).
18. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).
19. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
20. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421 (2014).
21. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* 1 (2018).

22. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.* **50**, 1728–1734 (2018).
23. O'Connor, L. J. *et al.* Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).

Referee #1 (Remarks to the Author):

The authors have fully addressed my comments. Overall, this work uses state-of-the-art methods to provide novel and important insights about the genetics of psychiatric disorders. This will be of great interest to the research community.

Referee #2 (Remarks to the Author):

1. Can the authors please use MAGMA and LDSC for the cell type analyses? The method selected may be fine, but I believe much more rigorous testing of MAGMA and LSDC has been conducted. Further, to make findings comparable with prior leading reports examining cellular associations with psychiatric disorders I would recommend prioritizing MAGMA and/or LDSC results.

We appreciate the reviewer raising the possibility of including these additional methods, particularly given the multi-method focus of the manuscript. We have updated the single-cell enrichment findings to now include results from MAGMA. In line with the approach taken in the most recent Schizophrenia GWAS (doi: 10.1038/s41586-022-04434-5), the results reported in the main text reflect the mean of the evidence across the new MAGMA results and the prior Expression Weighted Cell Type Enrichment (EWCE) findings. While we appreciate that MAGMA has been more widely used, we highlight that the EWCE approach has also been extensively applied, with the original paper now cited over 350 times (https://doi.org/10.3389/fnins.2016.00016). More importantly, the noted original EWCE paper performs extensive validation and the analytic approach, which compares the enrichment in the target gene set to a bootstrapped sample of 100,000 randomly generated background gene sets, is rigorous.

We believe that this updated set of results, which considers evidence from both methods (EWCE and MAGMA), yields a particularly robust set of findings. Figure 5 has been updated to report these averaged results and Supplementary Tables 48 and 49 have been added to report the individual *p*-values both within MAGMA and EWCE and averaged across the two methods. The Results section has also been updated to now read (p. 10):

"Averaged results across expression weighted cell type enrichment (EWCE)[52] and MAGMA were used to evaluate enrichment within neuronal cell types in fetal and adult single-cell datasets[53–57] ***(Suppl. Tables 48-49)****. Genes associated with the SB factor were significantly enriched in fetal data in interneurons and seven excitatory neuron subtypes, the strongest of which was for excitatory maturing neurons (****Fig. 5****).[53,54] The SB factor was also uniquely enriched for deep-layer excitatory neurons in adult brain.[57] Internalizing disorder genes were enriched within four excitatory neuron subtypes in fetal data,[53] though the signal was not as strong or pervasive as for the SB factor. In adult data, the Internalizing factor was enriched for medial ganglionic eminence (MGE) interneurons[56] and different glial cells, specifically oligodendrocytes and Bergmann glia[56,57]. The p-factor was enriched for five excitatory neuron subtypes in fetal data and oligodendrocyte precursor cells in adult data[56]."*

In addition, we now write in the Method section regarding the application of MAGMA (p. 30):

"MAGMA gene-set enrichment analyses, performed via the 'MAGMA.Celltyping' package in R[81]. Rather than considering only the top associated genes, as done in EWCE, MAGMA relies on the

*genome-wide signals to competitively evaluate enrichment via linear regression[48]. We used the European subset of the 1000 Genomes[78] as LD reference data, and mapped SNPs to genes based on their genomic location (GRCh37/hg19). To allow the inclusion of nearby regulatory variants, we considered all SNPs within a 35kb upstream and 10kb downstream window of the gene transcription region. Since signed effect size estimates are not available for the Q_{SNP} results, these analyses were restricted to the factors. The FDR corrected p-values from MAGMA and EWCE were averaged together to produce the results reported in the main text (but see **Suppl. Tables 48-49** for p-values from the individual methods).”*

We also appreciate the reviewer's suggestion to add in partitioned LDSC to evaluate enrichment. We have updated the text to now highlight that Stratified Genomic SEM reflects a multivariate extension of the partitioned LDSC framework (p. 10):

*“We applied Stratified Genomic SEM[27], a multivariate corollary of partitioned LDSC[58], to characterize the functional signals captured by the psychiatric factors in the five-factor and p-factor models, estimating enrichment for 162 functional annotations that passed QC (**Method**; **Suppl. Table 50**).”*

Referee #3 (Remarks to the Author):

Most of this work is very well done. The reviewers have addressed my comments and those of the other reviewers well. But there is one major issue with the interpretation that I think is important to address.

1. I'm happy to see that the authors followed my earlier suggestion to look at genetic correlations (rg's) with "normal" complex traits. I believe these results actually support the idea that the latent factors identified here are not "disease" dimensions per se. If you look at the rg's with various traits, many are quite high, and keep in mind that all the signal comes from common variants (meaning they occur a lot because they probably do "normal" things). That makes it much more likely that these factors reflect normal dimensions of genetic variation -> dimensions that increase risk for psychiatric disorder only when someone falls at an extreme end and is exposed to specific environmental stressors.

I strongly disagree with how the QTrait statistic is used to determine which rg's are considered "significant." The QTrait metric is certainly useful for interpretation and should be reported, but it shouldn't drive the main presentation of the rg's. In its current use, it effectively filters out strong and meaningful genetic correlations, giving a misleading impression of what the latent factors are associated with.

For example, on page 7 the authors write: "Daytime napping was the only trait significantly correlated with the p-factor (rg_p = .31, SE = .03)." And on page 12: "However, the p-factor also had more hits for the QSNP heterogeneity metric (117) than all five-factors from the correlated factors model (33) and was significantly correlated with daytime napping only, indicating the p-factor alone is insufficient for explaining cross-disorder risk."

This interpretation is difficult to understand. It appears that a trait is only considered "significantly correlated" with the p-factor if its rg is significant and the QTrait is not significant. But this decision obscures the true pattern of genetic correlations.

Supplementary Figure 31 shows many traits with high rg's with the p-factor: neuroticism has an rg > 0.6, income around 0.4, and there are also substantial rg's with traits like self-harm, loneliness, and childhood intelligence. These are all highly informative and clearly indicate that the p-factor reflects a broad dimension of heritable variation that includes a strong neuroticism/SES component. This is also consistent with the hypothesis I raised in my previous review.

By using QTrait significance as a filter, Figure 2 ends up showing only "daytime napping" as associated with the p-factor, which is clearly misleading. This framing underplays one of the paper's most interesting findings. I recommend showing all significant rg's and presenting QTrait separately as a heterogeneity metric, not as a disqualifier.

We appreciate the reviewer raising this important point and agree that reporting daytime napping as the only pervasively associated external trait fails to accurately summarize the pattern of genetic correlations. We have updated the results to expand the definition of significance for the p-factor to still include traits that were significant for Q_{Trait} as long as they were also significant for the majority (3 or more) of the five factors from the correlated factors model. We hope the reviewer will see this as a reasonable middle ground for evaluating significant correlations for the p-factor in that that does not require that the pattern of correlations conform to the factor model (as is assessed by Q_{Trait}), but does require that the trait is transdiagnostically associated with multiple factors. We have updated Figure 2 and Supplementary Table 13 to denote any traits that meet this new significance definition for the p-factor. We now write in the Results section (p. 7):

"Significant correlations were defined at a Bonferroni-corrected threshold of p < 2.68E-4, while not significant for Q_{Trait} at this same threshold. This Q_{Trait} exclusion criteria was relaxed for the p-factor if that trait was significantly associated with the majority ($\geq$ 3) of the five correlated factors, as this indicates the trait is capturing transdiagnostic associations the p-factor is intended to index."

Among these traits that we now consider significant for p are, as the reviewer notes, the broad dimensions that are speculated to underlie this general psychopathology construct, including neuroticism, self-harm, and loneliness.We have further updated the Results section to report some of these top correlations (p. 8):

"As would be expected, the five traits significantly associated with all five correlated factors were also among the top correlations for the p-factor, reflecting: stress sensitivity (r_{g_p} = .50, SE = .02), loneliness (r_{g_p} = .62, SE = .02), neuroticism (r_{g_p} = .64, SE = .02), self-harm (r_{g_p} = .74, SE = .04), and suicide attempts (r_{g_p} = .87, SE = .03)."

We have accordingly updated the Discussion section to denote these changes (and remove mention of daytime napping). We now write on p. 12:

"At the genome-wide level, the p-factor was strongly related to the Internalizing disorders factor and evinced the largest genetic correlations with external traits reflecting broad clinical characteristics, such as neuroticism, stress sensitivity, and loneliness. These results are consistent with conceptualizations of the p-factor as reflecting a general tendency towards negative emotionality.[61]"

2. More generally, I find Figure 2 quite difficult to interpret. The design is unintuitive, and it's not immediately clear what is being shown or why. I would encourage the authors to consider a more traditional and readable way to present these results. Also, selecting only the top rg's to display in Figure 2 is not informative; it obscures the broader pattern of associations and may give a skewed view of the genetic architecture being described.

We have updated Figure 2 to now include a more traditional bar plot of the genetic correlations and agree that this offers a completer and more comprehensible overview of these findings.

3. Finally, I would encourage the authors to make use of Extended Data Figures, especially for results that are central to the interpretation. Here are three specific suggestions: a. Supplementary Figures 20–25 could be merged into a single, large, readable Extended Data Figure that spans a full page.

We appreciate the reviewer bringing up the use of Extended Data Figures. Supplementary Figures 20-25 have now been added as Extended Data Figure 2. These new Extended Data Figures are now referenced in the Results section on p. 8:

*"For the full set of correlations see **Suppl. Table 13**, for comparison across factors **Extended Data Fig. 2**, and for comparison across traits within each factor **Extended Data Fig. 3."***

b. Similarly, Supplementary Figures 26–31 could be merged in the same way.

This has now been added as Extended Data Figure 3.

c. Supplementary Figure 5 is highly informative and contains very important information for the future of the field of psychiatric genetics. It deserves to be shown as an Extended Data Figure attached to the main pdf.

We agree that this figure offers critical information that guides expectations for the future of the field. We have added it as Extended Data Figure 1.