nature portfolio

Peer Review File

Genomics yields biological and phenotypic insights into bipolar disorder

Corresponding Author: Dr Kevin O'Connell

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)
In the manuscript, researchers from the Psychiatric Genomics Consortium (PGC) report on a genome-wide association study (GWAS) of bipolar disorder (BD) in an impressively large sample size of ~150k cases and 2.8 million controls. This is a big jump from their previous sample size of ~42k cases and ~370k controls (Mullins et al. Nat Gen 2021). The analysis in the manuscript follows a standard workflow that the PGC has been perfecting for more than a decade now (so there should be little concern about the robustness of the results). In the new GWAS of bipolar disorder, the authors have linked close to 300 genome-wide loci influencing the risk of BD. The authors have thoughtfully investigated the heterogeneity in the results based on disease subtypes (BD1 and BD2) and mode of ascertainment (self-reported, clinical and community cases) and report significant differences between the groups in terms of genetic correlations. The authors find that the self-reported cases show low genetic correlation with clinical cases and also differ slightly in terms of the pattern of genetic correlations with the other psychiatric disorders. The authors argue that this difference is mainly due to differences in the distribution of BD subtypes. As expected, the increase in sample size has increased the explanatory power of the polygenic risk score. In the downstream analysis of the GWAS results, the authors prioritize a list of genes using different commonly used and well-validated statistical approaches and show that the prioritized genes are enriched for synaptic pathways, certain classes of drug targets like antiepileptics and rare deleterious variants observed in an independent cohort of BD and schizophrenia cases.

Overall, this is a well-executed, well-written classic PGC paper that anyone familiar with the psychiatric genetics literature will appreciate. However, looking at the bigger picture of the overall progress in the genetics of BD, compared to previous GWASs of BD (Stahl et al. 2019, Mullins et al. 2021), I am finding it hard to appreciate significant major advancements in our understanding of the genetics of BD. Of course, the PGC has made a huge leap in the sample size, thanks to the impressive team efforts of hundreds of researchers from the consortium. And, yes, there is a slight increase in the representation of non-Europeans. In terms of effective sample size, 17.7% of non-Europeans, which when broken down into individual ancestries, the proportions get much smaller compared to Europeans. I appreciate the increase in non-European samples, but I don't feel that just adding some non-European samples into the meta-analysis justifies the title "diversity fuels gene discovery and biological insight in bipolar disorder". This will only attract unnecessary criticisms from the scientific community and will undermine the very purpose of the title which is to highlight the importance of studying underrepresented populations.

The sample size is the main thing that stood out for me when I started reading the manuscript and it naturally raised the bars high in terms of what to expect from the new results. But to my disappointment, the increase in the sample size didn't appear to bring any major insights into the genetic architecture of BD, in comparison to what has been known already based on previous GWASs of BD (Stahl et al. 2019, Mullins et al. 2021). Compared to previous GWASs, the number of genetic loci for BD has tripled, but I am finding it difficult to appreciate what new major biological insights (emphasis on the word 'biological') these newly discovered loci bring except for the enrichments for brain tissues, cell types, neuronal pathways and certain classes of drug targets, all of which are known from previous studies. This makes me wonder if this is going to be the case when the sample size becomes even higher in future studies. If so, what is the roadmap the authors envision in terms of making meaningful progress in the genetics of BD? I'd have appreciated some honest discussions about this in the paper.

Besides the overall concern I described above, I have a list of comments and suggestions about the paper.

Robustness of the genetic loci
* The authors write that they have identified 298 genome-wide significant loci in the new meta-analysis. But there is no mention of how many novel and known loci. There is no discussion about how many of the previously identified 64 loci continued to stay genome-wide significant, and how many have lost their peaks in the Manhattan plot. It would be informative if the authors could report the genetic correlation between meta-analysis based on new vs old cohorts. Also, a plot comparing the P values of known and novel loci between old and new GWASs.

* Comparing the Manhattan plots from 2019, 2021 and the current BD GWASs, I see some major differences. For example, the tallest tower in chromosome 3 seen in 2019 and 2021 (rs9834970; P=6.6e-19) GWAS results has now lost its first place in the new results. In the new GWAS, which has many fold larger sample size than previous ones, the P value of rs9834970 has become less significant (P=2.2e-14) instead of getting stronger. This is concerning as it suggests that the cohorts are probably not adding much to this signal but rather diluting the signal. If so, then that suggests that makes one question the quality of the phenotype. I wonder if this is the same for other loci as well. Hence, I think it's important the authors present a P value plot visualizing how the P values of 64 previously reported loci have changed in the current analysis and how the P values of newly identified loci look in the previous analysis.

* What about the MHC locus? It is one of the strongest associations for schizophrenia where the causal gene is believed to be C4 (encoding complement factor 4). Despite the compelling demonstrations of the association of C4 copy number with schizophrenia, a recent study by the iPSYCH consortium failed to reproduce the association using genetic predictors of serum C4 levels, challenging the C4 hypothesis (https://www.medrxiv.org/content/10.1101/2022.11.09.22281216v1). Given this background, it would be great if the authors discussed the MHC locus in BD GWAS results. In the 2021 BD GWAS paper, Mullins et al. write that they observed a genome-wide locus in the MHC region, but they did not find a significant association with imputed C4 copy number. Is that the same even in the new data with a substantially increased sample size? Is there a colocalization in the MHC locus between schizophrenia and bipolar disorder? Perhaps, also with SLE? If so, testing the association with the C4 copy number might bring new evidence that might either strengthen or weaken the schizophrenia-C4 hypothesis.

Heritability analysis
* The authors report SNP heritability (SNP-h2) based on genome-wide SNPs, which is around 22% for clinical samples, close to what has been reported for schizophrenia (24%; Trubetskoy et al. Nature 2022). It will be informative to also report on how much of this SNP-h2 is explained by the genome-wide significant variants and if the 298 loci explain more of the SNP-h2 compared to the previously known 64 loci.
Drug repurposing analysis
* I am not a big fan of the "drug repurposing" analysis being presented in the recent GWASs where researchers look at the correlation of drug-induced gene expression patterns from in vitro experiments with gene expression patterns inferred from GWAS results. I am yet to be convinced that such analyses produce any meaningful results. For example, lithium is one of the most effective medicines used in the treatment of bipolar disorder. Did the authors find that lithium gene targets are enriched for bipolar disorder?
Rare variant enrichment analysis
* The statistical evidence for the convergence between common and rare variants is borderline. The authors report that the 80 prioritized genes for BD were enriched for ultra-rare damaging missense and loss of function variants observed in BD and schizophrenia cases. The effect size for enrichment is 1.16 (P=0.002) and 1.21 (P=0.02). How do these numbers compare to the enrichment of rare variants in genes linked to other psychiatric disorders such as major depression, autism, ADHD etc. or just a general list of brain-expressed genes and mutation-intolerant genes? Is the enrichment for rare variants in the bipolar GWAS genes observed in the background of brain-expressed genes or mutation-intolerant genes?
* A previous ExWAS of BD, in combination with schizophrenia, has reported an association with AKAP11, which created excitement in the field as AKAP11 is known to interact with the GSK3B, a known target lithium. Speaking of convergence between common and rare variants, did the authors find any evidence of common variant associations at the AKAP11 locus? Or are there common variant associations at the other lithium target genes? Such lines of investigation will be insightful (irrespective of the positive or negative outcome) rather than the drug repurposing analysis.

Tissue and cell type enrichment analysis
* I expected that the increase in sample size would have powered the tissue and cell-type specific enrichment analysis revealing new insights. In the single-cell enrichment analysis, the authors write that, among the non-brain cell types, they find enrichment for enteroendocrine cells of the large intestine and delta cells of the pancreas. What is the relevance of this finding to bipolar disorder? The authors write that they found enrichment of neuronal populations. Are there differences in the tissue and cell type enrichments between the BD subtypes or between BD and schizophrenia or BD and depression etc.?

Non-European ancestries
* The addition of non-European cohorts to the new meta-analysis is great. However, I am not convinced that the added samples had any major impact on the results except for a slight increase in improvement in fine mapping. The new genome-wide significant locus identified in the East Asian cohort is interesting, but it's not clear if it will replicate. Also, no clear biological link between genes at this locus (TET2, CXXC4) and bipolar disorder.
* In the discussion the author writes that "The multi-ancestry PRS provided the greatest improvement over the EUR-PRS in an AFR target cohort, achieving an over two fold greater increase in variance explained". But the description in the results section does not agree with this statement: "When the multi-ancestry PRS was applied to a clinically ascertained AFR target cohort, … the inclusion of self-report data greatly increased the explained variance (no self-report R2 = 0.010, with self-report R2 = 0.023). A similar pattern was observed for the EUR ancestry PRS (no self-report R2 = 0.010, with self-report R2 = 0.022)" According to the above statement, both EUR PRS and multi ancestry PRS explain similar R2. Only the inclusion of

self-report cases increases the R2 value from 0.01 to 0.02. Am I interpreting that right?

Less technical and more biological discussions needed
* The manuscript overall reads more like a technical paper, spending more time on describing the differences between self-report vs clinical case cohort. I'd rather be interested more in discussions on interesting loci and the nearby genes and their associated pathways and their biological links to bipolar disorder.

Referee #2

(Remarks to the Author)
Review of the manuscript: "Diversity fuels gene discovery and biological insight in bipolar disorder".

The manuscript presents results from a large multi-ancestry GWAS of bipolar disorder (BD) including 158,036 BD cases and 2.8 million controls, identifying 298 genome-wide significant loci. This is a considerable increase in number of samples analyzed and identified loci compared to the latest published GWAS which included 41,917 BD cases, 371,549 controls and identified 64 loci. The major increase in sample size is primarily due to a large contribution from 23andMe of data on self-reported BD (90,088 cases and 1,928,789 controls).
The study includes a careful evaluation of how ascertainment ("clinical", "community" and "self-report") affects the results. Besides the multi-ancestry GWAS, the authors also report results from ancestry specific GWAS of individuals with EUR, EAS, AFR and LAT ancestries, the latter three groups compose 17,7% (proportion of Neff) of the samples.
Enrichment analyses highlight potential affected tissues and cell types, and 47 BD risk genes are highlighted based on convergent evidence from seven different methods that link risk variants to genes by integration with functional genomics data.
The study contributes to our understanding of the genetics underlying BD and represents the next step towards identification of the biological underpinnings of BD, not only due to the current results but also from the wave of future studies that will be based on the new GWAS results.

I have some comments and questions:

1) The analyzed samples are collected using three overall approaches which the authors refer to as "clinical", "community" and "self-report". For most of the analyses the authors present results from analyses both with and without the self-reported samples. It would have made the manuscript and messages clearer if there had been one main phenotype/GWAS and the other analyses e.g. "excluding the self-reported data" were presented as sensitivity analyses in the supplementary. However, I understand the authors choice of reporting both versions of the results, due to the low genetic correlation between self-reported and clinical samples (rg=0.47), and the observation that only around 50% of self-reported influencing common variants are shared and concordant with variants influencing clinically ascertained BD. Self-reported data demonstrates considerable high genetic correlation with the community samples which justifies the inclusion in the meta-analysis. However, I think it would be relevant to investigate the heterogeneity between the clinical and self-reported data bit further (as sensitivity analyses in the supplementary). What are the genetic correlations of the self-reported data alone with other psychiatric disorders compared to the genetic correlations of the clinical samples alone with other psychiatric disorders?

2) Related to my point above – when reading the "Multi-ancestry meta-analysis" section it would be helpful to know the reason why results from both with- and without self-reported BD are reported. I would therefore suggest that the section that evaluates the genetic heterogeneities between ascertainment approaches comes before the GWAS results. I know that this does not follow the traditional structure in GWAS papers, but it will give a better understanding of the reporting of GWAS results with and without the self-reported data.

3) What is the proportion of BDI and BDII among the clinical and community ascertained individuals? I could not locate that information.

4) On page 3 the authors write: "There was minimal evidence of residual population stratification (LDSC intercept = 1.052 (se = 0.016), attenuation ratio=0.071 (s.e. = 0.013))." As far as I have understood the result is based on the multi-ancestry GWAS, so I find the wording a bit strange. Since the analysis includes samples with different ancestries, we would expect population stratification, I assume.

5) Figure 1, I think it should be added to the figure legend that the locus marked with a star is the genome-wide significant locus in EAS.

6) Most supplementary tables miss explanation of the column headers. All headers should be explained not only those that the authors think are essential.

7) What is the genetic correlation of BD between the ancestries? – this could be evaluated with e.g., POPCORN. I think this would be informative also for future GWAS. It would probably make most sense to evaluate this using the self-reported individuals, due to lack of clinical samples with non-European ancestry.

8) In relation to my point above, what is the SNP-h2 of the other ancestries? I assume there should be enough power to calculate this using LDSC, at least for individuals with LAT ancestry.

9) Page 10, the authors write “Within loci associated with BD in the multi-ancestry meta-analysis, genes (N = 80) implicated by putatively causal fine-mapped SNPs (Supplementary Table 31)”. In supplementary Table 31 there are much more than 80 genes, so please highly the 80 genes included in the rare-variant analysis.

10) Add sample sizes to Supplementary Table 9 (the SNP-heritability table).

11) Page 5, add sample sizes to the sentence: “We further stratified Clinical BD samples into BDI and BDII subtypes and performed similar analyses”.

12) Page 5 and Supplementary Table 10, MiXeR: I do not see strong support for the “best” model vs the “max”/”min” models in many of the analyses (i.e., negative AIC/BIC values). I think this needs to be addressed in the manuscript, and can the authors comment on why they think the analyses are valid.

13) Supplementary Table 10, tri-variate MiXeR, it is not clear to me what column one represents, e.g. what does “bdCommunity_bdclinical_bdselfreport_12.json” mean? please use another term or explain in the header.

14) Page 5: the authors could consider to include block-jack-knife estimated p-values for differences in rg to quantify their statements in the sentences “In contrast, the genetic correlation between BDI and Self-report BD (rg = 0.42 (s.e. = 0.02)) was considerably lower than between BDII and Self-report BD (rg = 0.76 (s.e. = 0.05))”, and “… the genetic correlation for Self-report BD was considerably greater with Community samples (rg = 0.79 (s.e. = 0.02)) compared to Clinical samples (rg = 0.47 (s.e. = 0.02))..”.

15) On page 7 the authors write: “The variance explained by the multi-ancestry GWAS without the self-report data was significantly greater than that explained by the multi-ancestry GWAS including self-report data (R2 = 0.058, s.e. = 0.017, P = 2.72 × 10–4), and the EUR ancestry GWAS excluding the self-report data (R2 = 0.084, s.e. = 0.018, P = 5.62 × 10–3) (Supplementary Table 19).” Could the authors add information about the variance explained (i.e. the R2 estimate) by the multi-ancestry GWAS without the self-report data, to the text.

16) On page 9 the authors write: “…indicating involvement of neuronal populations from different brain regions, including hippocampal pyramidal neurons and interneurons of the prefrontal cortex and hippocampus”. It would be informative to know more about the brain data in the result section (e.g. pre- or post-natal tissues). In addition to this could the authors discuss the potential brain developmental stages affected (if they have the results to do so), it would be very interesting to know if genes affected by common risk variants play a role in early brain development or later in life.

17) Supplementary Figure 4, I think the figure needs a better explanation. It is not clear to me how the CS-values/the colors relate to the PIP threshold (or if there is a relation).

18) Page 10, was the VEP annotation of causal variants based on position alone?

19) On page 14 the authors write “Thus, the previous BDI and BDII GWAS summary statistics data were used for BDI and BDII analyses in this study.” Add information about sample sizes.

20) Have the authors evaluated the genetic correlation between males and females based on sex-stratified BD GWAS?

21) Page 18, the authors write “TWAS p-values were Bonferroni-corrected for the number of genes included in the respective analysis.” Please state the p-value threshold for declaring significance and the number of genes corrected for.

22) Page 18, isoTWAS results: the authors write “We then performed probabilistic fine-mapping of the significant associations, …” how did the authors define significant associations and did they apply Bonferroni correction?

23) Very little information is provided about the exome-sequencing data. Even though the authors refer to papers explaining these data it would help the reader to have more information. E.g. I could not locate the sample size of the individuals included in the sequencing studies in the manuscript and how damaging missense and protein-truncating variants are defined.

24) In the enrichment analysis of exome-sequencing data the authors use a Fisher´s exact test. I understand the wish to apply Fisher´s exact test to increase power but I think it would be more appropriate to apply logistic regression which allows for the inclusion of covariates to correct for technical issues related to the sequencing data and potential remaining population stratification. Could the authors comment on their choice of test?

Referee #3

(Remarks to the Author)
O’Connell et al produce the largest GWAS to date of BD and identify many new loci. This is an impressive sample size, but I have several comments particularly regarding the frame and readability for the authors.

1. The frame of the manuscript as written is oriented around the multi-ancestry analysis of BD. However, from Table S2,

69/79 cohorts are of European descent (83% of total participants) with 131,969 cases and 2,322,416 controls, versus a combined total of 26,067 cases and 474,083 controls from all other ancestries. Further, there is only one genome-wide significant association in any non-European ancestry group. The vast majority of the non-European ancestry participants are self-report cases from 23andMe. A few questions/comments arise from this:
a. Is the multi-ancestry frame for this paper the strongest way to introduce this work, given that it consists of participants who are by a large majority of European descent? As framed, I'm surprised there is not more emphasis on the multi-ancestry fine-mapping results, although Figure S4 suggests CS size identified in the meta-analysis closely reflects the European only results, which is a bit underwhelming. Using the test statistics and LD information within vs across populations, some simulation or downsampling/projection work showing how many loci in the multi-ancestry meta-analysis were discovered that wouldn't have been discovered from a similar sample size of European ancestry group participants alone could be interesting.
b. There seem to be significant limitations in polygenic cross-ancestry comparisons of genetic architecture in this study due to differences in phenotyping. The genetic correlations are far lower for self-report vs all phenotypes than for all other pairwise comparisons.
c. Is the deeper phenotype dissection a better frame for the paper? For a broad readership beyond a journal focused on psychiatry, introducing the differences between BDI and BDII is important, since this is unlikely common knowledge.
2. Figure 2 presents perhaps the most interesting results of the paper, but the presentation is somewhat unintelligible. MiXeR estimates are unlikely to be totally obvious to the broad readership of Nature, and there is no description of what these estimates mean. In contrast to the unclear caption "The number and standard errors of influencing variants (???)…," the original MiXeR paper describes these estimates much more clearly: "The numbers indicate the estimated quantity of causal variants (in thousands) per component, explaining 90% of SNP heritability in each phenotype, followed by the standard error." The genetic correlations from LDSC shown below the Venn Diagrams are useful and interesting, but I have no idea what the red bar is supposed to mean. It corresponds to neither the genetic correlation, nor polygenicity as far as I can gather, and is not described anywhere.
3. At the bottom of pg 6, it is unclear why the authors chose to contrast directionality of BD with vs without self-report between MDD and SCZ. Table S11 indicates larger contrasts between ADHD and SCZ, which is a little visually easier to see when plotted (including error bars) and perhaps would make an interesting supplementary figure. The discussion mentions the absolute rankings rather than the relative differences, so MDD and SCZ may make sense to discuss there, but overall a more nuanced cross-phenotype discussion seems warranted.

4. The PRS accuracy results are improved as expected. However, some of the language around interpretation is strange: "therefore the BD liability explained remains insufficient for diagnostic prediction in the general population." I don't think there is any area of medicine where people use or think of PRS alone as diagnostic. Instead, they are predictive tools, used together with other clinical measures, and the prediction accuracy described here is fine.
a. However, in Figure 3, it is a little odd that the all-cohort analysis with more phenotypic variation is not predicted as accurately compared to all subtypes analyzed. Is the self-report data pulling the PRS accuracy way down? The "Ascertainment" column in Table S20 is I think incorrect – w23andMe should be a mixed ascertainment, not "Clinical", correct?
b. "When the multi-ancestry PRS was applied to a clinically ascertained AFR target cohort, assuming 2% disease prevalence, the inclusion of self-report data greatly increased the explained variance (no self-report R2 = 0.010, with self-report R2 = 0.023)." The majority of AFR participants have self-report data. Unless I missed something, I think a more accurate interpretation of this would be that when there are more ancestry-matched participants in the discovery GWAS, the PRS performs better (as opposed to emphasizing that AFR self-report vs other types of phenotype ascertainment are especially relevant given that the sample sizes are too small to draw conclusions about this). The discussion seems to convey this more accurately.
5. Like other figures, the caption of Figure 4 assumes a level of familiarity of someone having directly made or discussed with the plot creator the interpretation of a sunburst plot rather than the broad readership of Nature. It is unclear what the hierarchy in this plot is indicating or what it is based on, aside from vaguely relating to pre- vs post-synapse activity. It simply says child terms are subsets of the adjacent inner ring. This description needs to be much clearer. Most regions of the sunburst plot are unannotated – what are they? Is this plot showing distance from synapse in some way? Further description of what specifically was done here is needed.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)

The authors have satisfactorily addressed my comments and implemented my suggestions. I thank the authors for doing so. The manuscript now reads better than the earlier version. But still I feel it could be improved further to suit for Nature's broad readership. Some suggestions include:
- Avoid unnecessary acronyms such as GWS.
- Avoid repeating the method names, for example, GSA-MiXeR, every time when describing the results. After the first time, directly describe the findings and if curious, the reader could refer to methods to find what statistical method or tool was used.

- In the end of each section, add a concluding statement(s) summarizing what was the take home from that part of the analysis. Just by reading the first and last few lines of each section, the reader should be able to grab the main messages that the authors were trying to make.
- Please use color-blind friendly palettes for the figures.
- I'd advise against using the phrase "prediction" in the polygenic score analysis. Instead, just use the word "association".
Specific comments:
- The authors prioritize two lists of causal genes. Initially, they prioritize 71 genes using fine-mapping, which they use for the common-rare variants convergence analysis. Later, they describe further analyses in which they prioritize a shorter list of 36 genes, which they use for lifespan gene expression and drug target enrichment analysis. The abstract also reports the 36 genes. This is a bit confusing. I'd advise sticking to one list, perhaps just the 36 genes and do all downstream analysis based on that list. I'd guess the borderline significance of rare variant enrichment would disappear when using 36 genes. If so, I'd advise reporting as it is and not claim convergence of common and rare variants.
- In the fine mapping analysis, how many genes implicated by coding variants? If there are any, mentioning that will be informative.

Referee #2

(Remarks to the Author)
It was a pleasure to read the revised manuscript. The authors have done an excellent and thoughtful revision that has significantly enhanced the quality of the manuscript. The presentation is clear and engaging, making the paper an interesting read. I particularly appreciate the additional analyses, which provide valuable biological insights, such as the evaluation of identified BD risk genes across brain developmental stages and the SNP-heritability enrichment analyses of cell types. The reordering of the manuscript, where heterogeneity among BD subtypes or ascertainment is presented before the GWAS, is also a notable improvement.

I only have a few minor comments:

1. Figure 2:
- I appreciate the new Figure 2, which is very informative. For clarity, I will suggest to edit the text on page 6 to the following: “Most psychiatric disorders, including major depressive disorder (MDD), post-traumatic stress disorder (PTSD), attention deficit/hyperactivity disorder (ADHD), borderline personality disorder, and autism spectrum disorder (ASD), were more strongly correlated with the full meta-analysis, BDII, and BD in Community and Self-reported samples, than with BDI and BD in clinical cohorts (Figure 2)”.
- The figure itself needs some polishing. I suggest demonstrating the genetic correlations as dots (± standard error) and removing the gridlines. Additionally, information about the horizontal lines should be added to the figure legend. I assume they represent standard errors.
- The ADHD GWAS was published in 2023, so this should be corrected in the figure and in the supplementary table.

2. Figure 4:
- I think this figure adds great value to the manuscript. Could the authors please include information in the figure legend about the small bar in the middle — specifically, what do the numbers 1-6 represent?

3. Supplementary Information:
- Related to my previous comment regarding the explanation of headers in the Supplementary Information. I appreciate the authors clarification of the headers of the MiXeR results in their rebuttal letter. However, I could not find any details in the Methods section about the number of runs performed in the trivariate MiXeR, which makes it difficult for the reader to understand what “Run” means. Please add more information about this to the Methods section and the header of Supplementary Table 4.

4. Rare Variant Analyses:
- The authors have performed a Fisher's exact test, which does not correct for covariates, and I understand their rationale for using this method. However, I think it is important to evaluate whether the signal is not due to sequencing or batch errors. Could the authors include results from an analysis of rare synonymous variants in the BD risk gene set, as we would not expect differences between cases and controls for this category of variants?

5. Clustering of genes based on expression patterns across brain developmental stages:
- Page 22: The authors write: “Outliers in gene expression more than 4 standard deviations from the mean were removed.” What was the rationale for removing outlier genes if QC of the expression data is assumed to be properly done? The genes removed might not be outliers if the tested set had included other or more genes.
And related, why was 4 standard deviations chosen?

- Include a reference or URL to the TmixClust R-package.

- Page 22: the authors write: “We compare the average silhouette width across the K=2 to K=10 clusters and select that with the maximum value as the optimal number of clusters”. Since high silhouette values (i.e., close to one) indicate a good support for the clustering, it is important to know the estimated silhouette value for K=2 clusters – please add that in the manuscript.

- Supplementary Figure 10. There is no explanation of the grey shaded area around the lines, but I don't think this is confidence intervals related to expression because I would expect these to be larger. Please add confidence intervals for the average expression of the genes in the two clusters at each brain developmental stage to the figure.

- Additionally, the scale on the x-axis is in years but not linear in the figure, please add a brief explanation to the S10 figure legend explaining the scale.

Referee #3

(Remarks to the Author)
The authors have been responsive to most reviewer comments, including all of mine. Reviewer 1's major request for an honest discuss about new *biological* insights gleaned from increasing sample sizes here could be strengthened with a potential roadmap and discussion around how future bipolar genetic studies to deepen insights. This was partially addressed but could perhaps be strengthened in collaboration with the editorial team. Other efforts to strengthen and reframe the paper have been responsive.

Version 3

Reviewer #2

I would like to thank the authors for addressing many of my previous questions. However, I have a few additional items that I would like to see further addressed:

1. Although I am not an expert in clustering methods, I have gathered from the literature that a silhouette width of 0.5 is generally indicative of meaningful clustering, while a value below 0.25 is considered weak. If my understanding is accurate, it seems there may not be sufficient support for identifying two distinct gene clusters, as mentioned on page 12. Therefore, the conclusion regarding the identification of two gene clusters seems overly strong.

2. In Supplementary 8 figure legend, the authors write: "To illustrate the variability in gene expression within each cluster, we plot each donor's expression value in each sampled brain region for the 34 credible genes as individual points." However, I noticed that only 33 genes are listed as belonging to a cluster in the supplementary table from the previous revised version. Please clarify whether the correct number is 33 or 34.

Concerning the analyses of rare variants, I am puzzled as to why access to counts of synonymous variants for bipolar risk genes was not possible. Both the schema browser and the BipEX browser provide counts for synonymous variants. I believe it is crucial to include analyses of synonymous variants, and I recommend strongly this to be incorporated.

We thank all reviewers for their thorough review of the manuscript and insightful comments. The differences in genetic architecture of bipolar disorder (BD) based on ascertainment (i.e. clinical vs. self-reported) and subtype are important and represent key findings in this paper, and we now present the results in a less technical and method-focused way, as suggested by the reviewers. Prompted by the reviews, we have also included additional data and analyses to add new biological insights. In particular, we have added new data from single-nucleus RNA sequencing and temporal neurodevelopmental expression. We have also expanded our variant/locus-to-gene analyses and strategy, implementing additional lines of evidence including mQTL, sQTL and enhancer-promoter interactions. The revised results now include additional novel findings that provide new biological insights. Accordingly, the discussion has been reordered as well.

Below we provide responses (in blue) to the specific comments from each reviewer. Changes to the text are marked in yellow in the resubmitted main and supplementary documents.

Referees' comments:

Referee #1 (Remarks to the Author):

In the manuscript, researchers from the Psychiatric Genomics Consortium (PGC) report on a genome-wide association study (GWAS) of bipolar disorder (BD) in an impressively large sample size of ~150k cases and 2.8 million controls. This is a big jump from their previous sample size of ~42k cases and ~370k controls (Mullins et al. Nat Gen 2021). The analysis in the manuscript follows a standard workflow that the PGC has been perfecting for more than a decade now (so there should be little concern about the robustness of the results). In the new GWAS of bipolar disorder, the authors have linked close to 300 genome-wide loci influencing the risk of BD. The authors have thoughtfully investigated the heterogeneity in the results based on disease subtypes (BD1 and BD2) and mode of ascertainment (self-reported, clinical and community cases) and report significant differences between the groups in terms of genetic correlations. The authors find that the self-reported cases show low genetic correlation with clinical cases and also differ slightly in terms of the pattern of genetic correlations with the other psychiatric disorders. The authors argue that this difference is mainly due to differences in the distribution of BD subtypes. As expected, the increase in sample size has increased the explanatory power of the polygenic risk score. In the downstream analysis of the GWAS results, the authors prioritize a list of genes using different commonly used and well-validated statistical approaches and show that the prioritized genes are enriched for synaptic pathways, certain classes of drug targets like antiepileptics and rare deleterious variants observed in an independent cohort of BD and schizophrenia cases.

1) Overall, this is a well-executed, well-written classic PGC paper that anyone familiar with the psychiatric genetics literature will appreciate. However, looking at the bigger picture of the

overall progress in the genetics of BD, compared to previous GWASs of BD (Stahl et al. 2019, Mullins et al. 2021), I am finding it hard to appreciate significant major advancements in our understanding of the genetics of BD. Of course, the PGC has made a huge leap in the sample size, thanks to the impressive team efforts of hundreds of researchers from the consortium. And, yes, there is a slight increase in the representation of non-Europeans. In terms of effective sample size, 17.7% of non-Europeans, which when broken down into individual ancestries, the proportions get much smaller compared to Europeans. I appreciate the increase in non-European samples, but I don't feel that just adding some non-European samples into the meta-analysis justifies the title "diversity fuels gene discovery and biological insight in bipolar disorder". This will only attract unnecessary criticisms from the scientific community and will undermine the very purpose of the title which is to highlight the importance of studying underrepresented populations.

> We agree with the reviewer that the title of the manuscript may not be fully justified. We have now changed it to "Genomics yields biological and phenotypic insights into bipolar disorder", which is more in line with the new framing and structure of the manuscript to focus more on ascertainment and subtype within BD. We have addressed the reviewer's comment about lack of major advancements in our understanding of the genetics of BD with a series of new biological analyses, which are specified in the response to the relevant comments below.

2) The sample size is the main thing that stood out for me when I started reading the manuscript and it naturally raised the bars high in terms of what to expect from the new results. But to my disappointment, the increase in the sample size didn't appear to bring any major insights into the genetic architecture of BD, in comparison to what has been known already based on previous GWASs of BD (Stahl et al. 2019, Mullins et al. 2021). Compared to previous GWASs, the number of genetic loci for BD has tripled, but I am finding it difficult to appreciate what new major biological insights (emphasis on the word 'biological') these newly discovered loci bring except for the enrichments for brain tissues, cell types, neuronal pathways and certain classes of drug targets, all of which are known from previous studies. This makes me wonder if this is going to be the case when the sample size becomes even higher in future studies. If so, what is the roadmap the authors envision in terms of making meaningful progress in the genetics of BD? I'd have appreciated some honest discussions about this in the paper.

> We think that the large increase in sample size and the increased number of associated loci are a significant addition to our knowledge. They provide added depth to the followup analyses, strengthening enrichment analyses. For example, the brain specific cell-type enrichment analyses of the full meta-analysis highlights novel enriched cell-types beyond those from previous findings, such as GABAergic interneurons and pyramidal neurons of the prefrontal cortex.[1] We also identified novel enrichment in enteroendocrine cells of the large intestine and delta cells of the pancreas (see response to comment 10 below). Furthermore, the increase in associated loci led to a large enough set of fine-mapped genes from which we were able to identify convergence of common and rare variant signals in BD, which is also novel.

To specifically address this reviewers' comment and further expand the biological insight gained in the revised manuscript, we have carried out additional analyses of the multi-ancestry meta-analysis as well as analyses by subtype and ascertainment source. We have updated our enrichment analyses to include brain developmental stages, providing further context for the noted enrichment within brain tissue. These new results highlight the role of early- to mid-prenatal brain development in the aetiology of BD. Furthermore, cell-type enrichment by ascertainment and BD subtype showed similar patterns of enrichment. Together, these new results provide new insights in the molecular biology of BD.

Moreover, the inclusion of the new and larger data sources from biobanks and 23andMe allowed us, for the first time, to investigate differences in ascertainment, which is now a major focus of the paper.

Updated results text reads as follows, with new text highlighted in yellow (page 9, Pathway, tissue and cell type enrichment):

> *"The association signal was enriched among genes expressed in the brain (Supplementary Table 24), and specifically in the early- to mid-prenatal stages of development (Supplementary Table 25). Single-cell enrichment analyses of brain cell types indicate involvement of neuronal populations from different brain regions, including hippocampal pyramidal neurons and interneurons of the prefrontal cortex and hippocampus (Supplementary Figure 5)... Similar patterns of enrichment were observed based on ascertainment and subtype (Supplementary Figure 6)."*

We have also added enrichment analyses of gene expression profiles derived from single-nucleus RNA-seq (snRNAseq) data from adult postmortem brain tissue samples,[2] showing enrichment of specific brain tissue cell clusters, distinct from those identified in schizophrenia and depression.[3]

New results text reads as follows (page 9, Pathway, tissue and cell type enrichment):

> *"A recent study[2] analysed single-nucleus RNA sequencing (snRNAseq) data of 3.369 million nuclei from 106 anatomical dissections within 10 brain regions and divided cells into 31 superclusters and 481 clusters, respectively, based on principal component analysis of sequenced genes. These superclusters were then annotated based on their regional composition within the brain (Figure 4). We used stratified LD score regression (S-LDSC)[4] to estimate SNP-heritability enrichment for the top decile of expression proportion (TDEP) genes in each of the 31 superclusters and 481 clusters, as described previously.[3] Heritability was significantly enriched in 9 of the 31 superclusters (Figure 4), and 49 of the 481 clusters (Supplementary Figure 7). No enrichment was seen in non-neuronal clusters. Interestingly, two clusters within the medium spiny neurons, not*

observed at the supercluster level, are significantly enriched further supporting the involvement of striatal processes in BD."

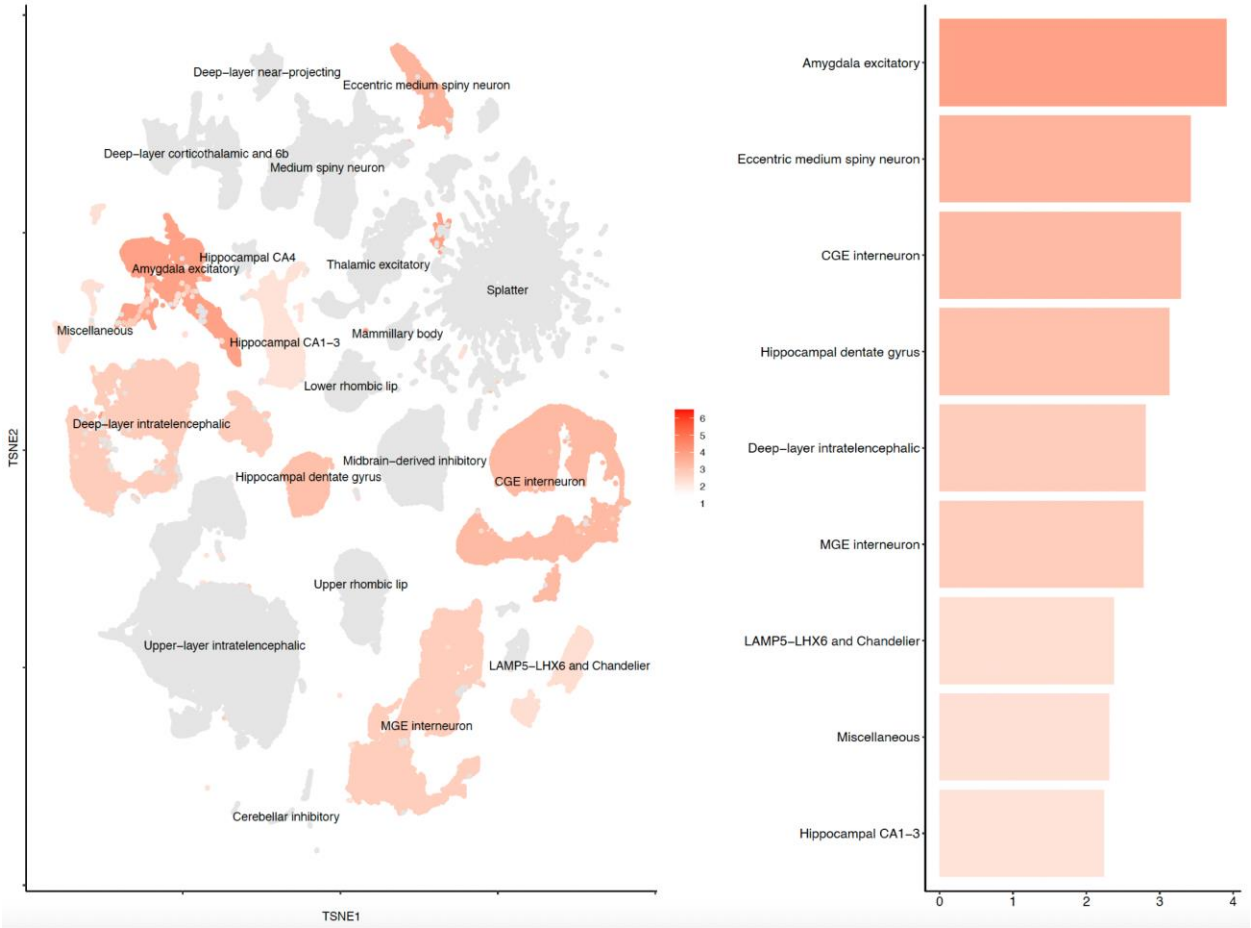


Figure 4. *Supercluster-level SNP-heritability enrichment for bipolar disorder. The t-distributed stochastic neighbour embedding (tSNE) plot (from Siletti et al.[2]) (left) is coloured by the enrichment z-score. Grey indicates non-significantly enriched superclusters (FDR > 0.05). The barplot (right) shows the nine significantly enriched superclusters.*

New discussion text reads as follows (page 13, Discussion):

"These findings were further corroborated by enrichment analyses in single-nucleus RNA-seq data from adult postmortem brain tissue, which highlighted specific clusters of interneurons derived from the caudal and medial ganglionic eminences and medium spiny neurons predominantly localised in the striatum. Medium spiny neurons are not enriched in depression using the same dataset.[3] Although interneurons derived from ganglionic eminences were also enriched in schizophrenia, stronger signals were observed for amygdala excitatory and hippocampal neurons.[3]"

In addition, we updated our analyses examining credible genes associated with BD. In our updated analyses we have now included data sources from brain mQTLs, sQTLs, and enhancer-promoter interactions in addition to the eQTL- based methods originally used. This updated approach identified a set of 36 credible genes. We explored the temporal expression patterns of this set of credible genes and identified two clusters of genes with opposite expression patterns, and major changes in gene expression before birth.

New results text reads as follows (page 11, Identification of credible BD-associated genes):

> *"In addition to the 71 genes annotated to the fine-mapped putatively causal SNPs as described above, we annotated a further 45 genes to the 80 fine-mapped SNPs by SMR using eQTL and sQTL data, as well as by proximity, i.e. the nearest gene to each SNP (Supplementary Figure 9, Supplementary Tables 30 and 31). No genes were annotated to the CpGs identified by the mQTL analysis (Supplementary Table 30). We then determined if any of these 116 genes were also identified through the genome-wide gene-based analysis using MAGMA,[5] eQTL analyses using TWAS as implemented in FUSION[6] and isoTWAS,[7] or through enhancer-promoter (E-P) interactions.[8,9] This resulted in seven possible approaches by which loci could be mapped to genes including, eQTL evidence (eQTL or TWAS or FOCUS or isoTWAS), mQTL, sQTL, VEP, proximity, MAGMA and E-P interactions.*
>
> *We integrated the results from the post-GWAS analyses described above and identified a credible set of 36 genes identified by at least three of the described approaches (Supplementary Table 31). The SP4 gene was identified by six of these approaches, and astrocyte and GABAergic neuron specific regulation of SP4, by the GWS variant rs2107448, were identified from cell-type specific enhancer-promoter interaction results (Supplementary Table 31). Moreover, the TTC12 and MED24 genes were identified by five of the approaches. Eight of the 36 credible genes have synaptic annotations in the SynGO database.[10] Three genes (HTT, ERBB4 and LR5NF) were mapped to both postsynaptic and presynaptic compartments. One gene (CACNA1B) was mapped to only the presynapse and four genes (SHANK2, OLFM1, SHISA9 and SORCS3) were mapped to only the postsynapse (Supplementary Table 32).*
>
> *Based on the lifespan gene expression data from the Human Brain Transcriptome project (www.hbatlas.org),[11] we identified two clusters of credible genes with distinct peaks in temporal expression (Supplementary Figure 10, Supplementary Table 31). The first cluster shows reduced prenatal gene expression, with gene expression peaking at birth and remaining stable over the life-course. Conversely, the second cluster shows a peak in gene expression during fetal development with a drop-off in expression before birth."*

We investigated and identified differences in BD genetic architecture related to ascertainment and subtype. These results suggest that BD diagnoses include a spectrum of mood phenotypes

across subtypes (BD I and II) and type of ascertainment. This heterogeneity needs to be accounted for when interpreting the current findings, and when designing future BD genetics research. As suggested by the reviewer we have now included this as a major point of consideration for the roadmap of BD genetics research.

Updated discussion text reads as follows, with new text highlighted in yellow (page 14, Discussion):

> *"We identified differences in the genetic architecture of BD subtypes related to ascertainment. BD within Clinical and Community samples was highly but imperfectly correlated, with varying correlations with Self-reported BD. The low genetic correlation and genetic overlap between cases ascertained through clinical studies and cases with self-reported BD is driven by a greater proportion of BDI within the Clinical samples. In line with these results, PRS derived from meta-analyses excluding the self-reported data performed better in Clinical and BDI target samples, while the inclusion of self-reported data improved the PRS in Community and BDII target samples. Moreover, the pattern of correlations between BD and other psychiatric disorders differed with the inclusion of self-reported data. Schizophrenia had the highest genetic correlation with BD without the inclusion of the self-reported data, while major depressive disorder was most strongly correlated with BD after the inclusion of the self-reported data. These results suggest that the Self-reported samples may include a high proportion of people with BDII. Moreover, this is in line with recent findings in individuals diagnosed with BDII, which showed increasing polygenic scores for depression and ADHD and decreasing polygenic scores for BD over time.[12] However, a diagnosis of BD in the outpatient setting may be overdiagnosed in people with conditions such as chronic depression or borderline personality disorder, highlighting a higher rate of comorbid disorders and potential for 'overdiagnosis' of BD within cohorts of this nature.[13,14] We showed that the differences in genetic architecture and phenotypic proportions of the Clinical, Community and Self-reported BD cohorts impacted the replication of prior BD-associated loci. Previously associated loci that fell short of meeting GWS in the current study were GWS in the Clinical samples and in the meta-analyses that excluded Self-reported data, and all top SNPs (12,151 SNPs with p < 1 × 10−5) from the previous GWAS were consistent in direction of association in this multi-ancestry meta-analysis of all samples (Supplementary Table 9)."*

Further updated discussion text, with new text highlighted in yellow (pages 14-15, Discussion):

> *"Another limitation is the inclusion of samples with minimal phenotyping. Although this allowed us to achieve large sample sizes, especially in under-represented non-European ancestry cohorts, and greatly increase the number of loci identified, it has some shortcomings. For example, minimal phenotyping may result in low specificity*

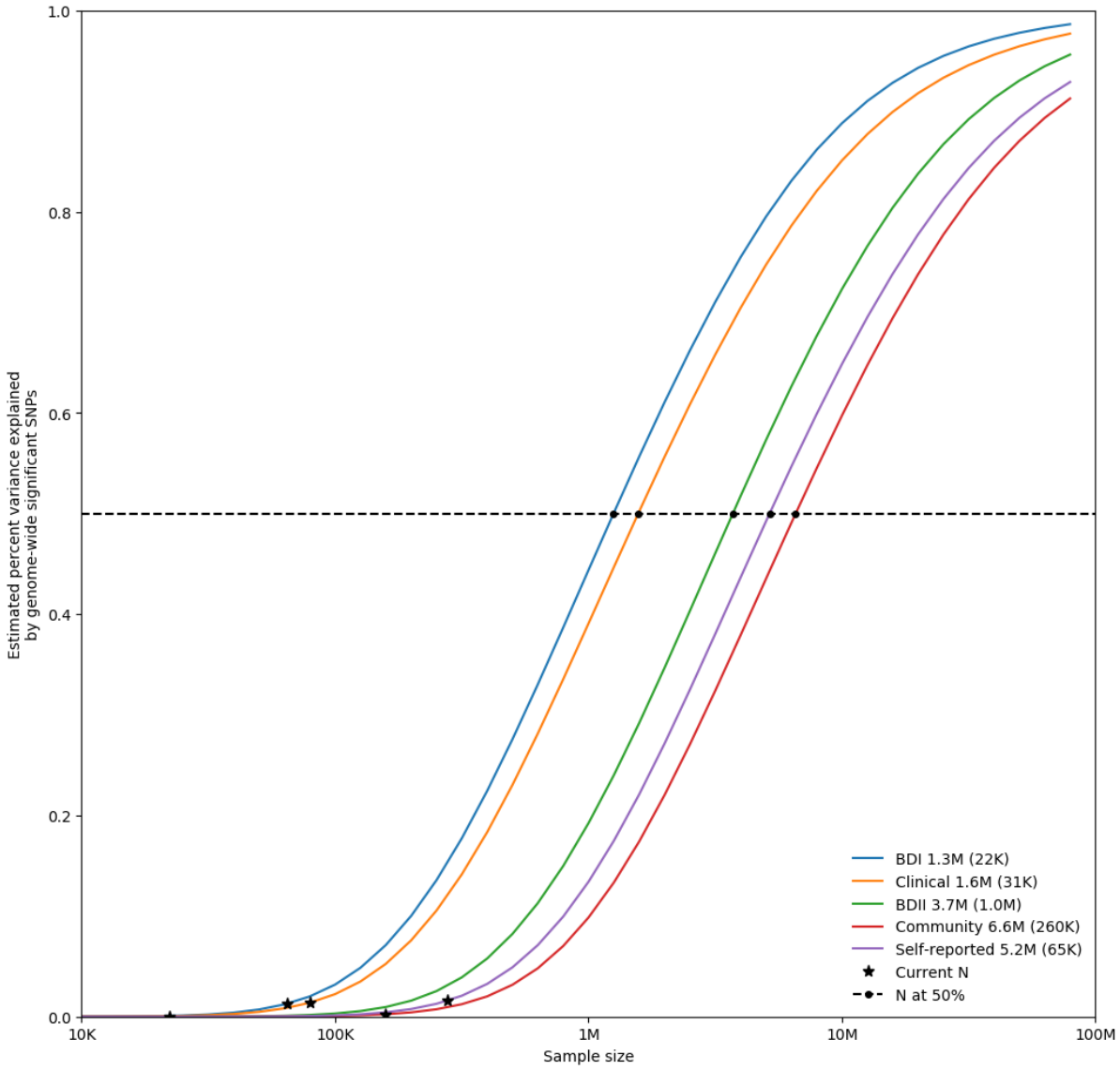
association signals, as shown in major depression,[15,16] and individuals in community-based biobanks may represent those less severely affected, as shown in schizophrenia.[17]"

"In conclusion, in this first large-scale multi-ancestry GWAS of BD, we identified 298 significant BD-associated loci, from which we demonstrate convergence of common variant associations with rare variant signals and highlight 36 genes credibly implicated in the pathobiology of the disorder. We identified differences in the genetic architecture of BD based on ascertainment and subtype, suggesting that stratification by subtype will be important in BD genetics moving forward. Several analyses implicate specific cell types in BD pathophysiology, including GABAergic interneurons and medium spiny neurons, and enrichment of dopamine- and calcium-related biological processes were also identified, further contributing to our understanding of the biological aetiology of BD."

In response to the comment about sample size and future progress in the genetics of BD, we also now provide statistical power calculations, estimated using univariate MiXeR,[18] which indicate the required effective sample size needed in order to capture specific proportions of SNP-based heritability with genome-wide significant variants for each BD ascertainment and subtype. An effective sample size of 1.3 million is required to capture 50% of genetic variance associated with BDI, while for BDII the effective sample size would be approximately 3.7 million.

These results are presented in Supplementary Figure 3 and below. We have also updated the results text as follows (page 3, Genetic architecture of BD differs by ascertainment and subtype):

"*MiXeR estimates show the greatest polygenicity for BD ascertained from Self-reported samples, followed by Clinical and then Community samples (Figure 1, Supplementary Table 4). BD in Clinical samples is estimated to be the most discoverable, while Self-reported BD had the lowest discoverability (Supplementary Figure 3, Supplementary Table 4).*"



Supplementary Figure 3: *Univariate MiXeR estimates of the required effective sample size needed to capture 50% of the genetic variance (horizontal dashed line) associated with each BD ascertainment and subtype. N and Sample size refer to the effective sample size. The estimated effective sample size (and standard errors) are given in the legend alongside each trait name.*

Besides the overall concern I described above, I have a list of comments and suggestions about the paper.

Robustness of the genetic loci

3) * The authors write that they have identified 298 genome-wide significant loci in the new meta-analysis. But there is no mention of how many novel and known loci. There is no discussion about how many of the previously identified 64 loci continued to stay genome-wide significant, and how many have lost their peaks in the Manhattan plot. It would be informative if

the authors could report the genetic correlation between meta-analysis based on new vs old cohorts.

> The genetic correlation between meta-analyses based on all new cohorts and the EUR cohorts from the previous PGC BD GWAS (old cohorts) is rg=0.64 (se=0.02). When excluding the self-reported BD cohorts, this correlation increases to rg=0.91 (se=0.04).
>
> The following text is added to the methods (page 5, Ancestry-specific GWAS meta-analyses):
>
> > *"The genetic correlation between meta-analyses based on all new cohorts (118,284 cases and 2,448,096 controls) and EUR cohorts from our previous PGC BD GWAS[1] was rg = 0.64 (se = 0.02), and rg=0.91 (se = 0.04) when excluding self-reported cohorts."*
>
> We have addressed the reviewer's comment regarding novel loci in our response to comment 4 below.

4) Also, a plot comparing the P values of known and novel loci between old and new GWASs.
* Comparing the Manhattan plots from 2019, 2021 and the current BD GWASs, I see some major differences. For example, the tallest tower in chromosome 3 seen in 2019 and 2021 (rs9834970; P=6.6e-19) GWAS results has now lost its first place in the new results. In the new GWAS, which has many fold larger sample size than previous ones, the P value of rs9834970 has become less significant (P=2.2e-14) instead of getting stronger. This is concerning as it suggests that the cohorts are probably not adding much to this signal but rather diluting the signal. If so, then that suggests that makes one question the quality of the phenotype. I wonder if this is the same for other loci as well. Hence, I think it's important the authors present a P value plot visualizing how the P values of 64 previously reported loci have changed in the current analysis and how the P values of newly identified loci look in the previous analysis.

> The reviewer is correct that there are differences in the observed association signals, which appear to be driven by phenotype (ascertainment and subtype). Of the 64 loci reported previously,[1] 31 met GWS in this new, multi-ancestry analysis containing all samples, and of the 33 that did not, 25 met GWS in either the Clinical samples or in the meta-analysis that excluded Self-reported data. These results are consistent with what has previously been reported for ADHD.[19]
>
> We have added the following text to the results (pages 5-6, Multi-ancestry meta-analysis):
>
> > *"Of the 298 loci identified in this multi-ancestry meta-analysis, 267 are novel for BD. Of the 64 previously reported BD-associated loci,[1] 31 met GWS in this new, multi-ancestry analysis containing all samples, and of the 33 that did not, 25 met GWS in either the Clinical samples or in the meta-analysis that excluded Self-reported data (Supplementary Table 9). Moreover, the direction of association for all top SNPs (12,151 SNPs with p < 1*

× 10^{-5}) from the previous GWAS was consistent with the direction of association in this multi-ancestry meta-analysis of all samples (Supplementary Table 9).”

And the following to the discussion section (page 14, Discussion):

“We showed that the differences in genetic architecture and phenotypic proportions of the Clinical, Community and Self-reported BD cohorts impacted the replication of prior BD-associated loci. Previously associated loci that fell short of meeting GWS in the current study were GWS in the Clinical samples and in the meta-analyses that excluded Self-reported data, and all top SNPs (12,151 SNPs with $p < 1 \times 10^{-5}$) from the previous GWAS were consistent in direction of association in this multi-ancestry meta-analysis of all samples (Supplementary Table 9).”

And the following text has been added to the methods section (page 17, Genome-wide association study (GWAS)):

“Concordance of the direction of associations in the present GWAS with associations in the previously published BD data were evaluated as described previously.[19]

In addition to the observed differences based on ascertainment and subtype above, to further examine the source of GWS variability across the present and previous studies, we also examined if any of the association signals amongst the 298 identified loci are impacted by ancestry. For almost all loci, the association signal was strongest in the EUR-ancestry meta-analysis, which is unsurprising given its sample size and corresponding power. Interestingly, for one locus (lead SNP rs7248481, chr19:13079957-13122567) the association p-value was lowest in the EAS ancestry meta-analysis. The majority of loci were nominally significant ($p < 0.05$) within the AFR (290/298 loci), EAS (257/298 loci) and LAT (293/298 loci) ancestry meta-analyses, highlighting consistency of signal across the ancestry groups.

We have therefore added the following text to the results (page 6, Multi-ancestry meta-analysis):

“When considering the impact of ancestry on the discovery of these 298 loci, one locus (lead SNP rs7248481, chr19:13079957-13122567) was most strongly associated in the EAS ancestry meta-analysis. For all other loci, the association was strongest in the EUR ancestry meta-analysis. The majority of loci were nominally significant ($p < 0.05$) within the AFR (290/298 loci), EAS (257/298 loci) and LAT (293/298 loci) ancestry meta-analyses, highlighting consistency of signal across the ancestry groups (Supplementary Table 8). Moreover, for 38 locus lead SNPs, the effect allele frequency is greater in the EAS, AFR and LAT ancestry samples than in the EUR ancestry samples, suggesting that these loci might not be discovered in a similar sample sized EUR-only meta-analysis (Supplementary Table 8).”

5) * What about the MHC locus? It is one of the strongest associations for schizophrenia where the causal gene is believed to be C4 (encoding complement factor 4). Despite the compelling demonstrations of the association of C4 copy number with schizophrenia, a recent study by the iPSYCH consortium failed to reproduce the association using genetic predictors of serum C4 levels, challenging the C4 hypothesis (https://www.medrxiv.org/content/10.1101/2022.11.09.22281216v1). Given this background, it would be great if the authors discussed the MHC locus in BD GWAS results. In the 2021 BD GWAS paper, Mullins et al. write that they observed a genome-wide locus in the MHC region, but they did not find a significant association with imputed C4 copy number. Is that the same even in the new data with a substantially increased sample size? Is there a colocalization in the MHC locus between schizophrenia and bipolar disorder? Perhaps, also with SLE? If so, testing the association with the C4 copy number might bring new evidence that might either strengthen or weaken the schizophrenia-C4 hypothesis.

> We agree that this would be an interesting analysis. However, additional imputation and analyses of the MHC/C4 locus were not possible since the number of samples for which individual-level genotype data were accessible did not increase much since the Mullins et al. paper[1]. The analysis in the previous paper was performed in 32,749 BD cases and 53,370 controls, and unfortunately the number with individual-level data has only increased by 3,850 BD cases and by 3,935 controls.
>
> We have added the following text to the limitations section of the discussion. New text is highlighted in yellow. (page 14, Discussion).
>
> > *"One limitation is the lack of in-sample LD estimates for all cohorts, due to a lack of in-house raw genotype data for some cohorts. For instance, analysis of the MHC/C4 locus was not considered since the number of samples for which individual-level genotype data were accessible did not increase much since the previous analysis[1]. "*

Heritability analysis

6) * The authors report SNP heritability (SNP-h2) based on genome-wide SNPs, which is around 22% for clinical samples, close to what has been reported for schizophrenia (24%; Trubetskoy et al. Nature 2022). It will be informative to also report on how much of this SNP-h2 is explained by the genome-wide significant variants and if the 298 loci explain more of the SNP-h2 compared to the previously known 64 loci.

> Using GSA-MiXeR, we estimated the proportion of SNP-heritability (SNP-h2) explained by the 64 loci from the previous BD GWAS[1] and the 298 loci identified in the multi-ancestry analysis presented here. SNPs within the 64 loci from the previous GWAS accounted for 8.3% of the

estimated SNP-h2 in the previous GWAS, whereas SNPs within the 298 loci account for 18.5% of the estimated SNP-h2.

We further examined these loci when targeting the clinical, community and self-reported meta-analyses, as well as BDI and BDII stratified samples. SNPs within the 298 loci accounted for higher proportions of SNP-h2 in the clinical (64 loci: 8.5%; 298 loci: 17.8%), BDI (64 loci: 8.3%; 298 loci: 17.5%), community (64 loci: 4.8%; 298 loci: 22.6%), and self-reported (64 loci: 2.0%; 298 loci: 21.1%) samples. The GSA-MiXeR model fits for BDII indicate that these estimated should be interpreted with caution (negative AIC values), but also show a greater proportion of SNP-h2 accounted for by SNPs in the 298 loci (42.5%) when compared to the previous 64 loci (13.1%).

We have added the following text to the results section, and also included the results in Supplementary Table 10.

New text (page 6, Multi-ancestry meta-analysis):

> *"We used GSA-MiXeR[20] to estimate the proportion of SNP-heritability (SNP-h^2) accounted for by SNPs within GWS loci. Compared to only 8.3% accounted for by SNPs within the 64 previously identified loci,[1] SNPs within the 298 loci account for 18.5% of the SNP-h^2 of BD (Supplementary Table 10). Moreover, SNPs within the 298 loci also accounted for higher proportions of SNP-h2 in the clinical (64 loci: 8.5%; 298 loci: 17.8%), BDI (64 loci: 8.3%; 298 loci: 17.5%), Community (64 loci: 4.8%; 298 loci: 22.6%), and Self-reported (64 loci: 2.0%; 298 loci: 21.1%) samples."*

Drug repurposing analysis

7) * I am not a big fan of the "drug repurposing" analysis being presented in the recent GWASs where researchers look at the correlation of drug-induced gene expression patterns from in vitro experiments with gene expression patterns inferred from GWAS results. I am yet to be convinced that such analyses produce any meaningful results. For example, lithium is one of the most effective medicines used in the treatment of bipolar disorder. Did the authors find that lithium gene targets are enriched for bipolar disorder?

We agree that the correlation analyses of drug-induced gene expression from in vitro experiments with gene expression patterns inferred from GWAS results cannot confirm whether the identified drugs would be effective. We have now removed this correlation-based analysis from the manuscript.

The therapeutic mechanism of lithium has been difficult to determine, [21,22] and this is an open question, as lithium has non-specific effects on brain tissue.[23] Still, we tested if our list of credible bipolar disorder genes are enriched for drug target genes (GSEA) and did not find that they were significantly enriched for lithium target genes. As the Drug Gene Interaction Database

(DGIdb)[24] has been updated since original submission, we re-ran these analyses. In this updated analysis we considered the 139 genes in the DGIdb listed as targets of lithium or interaction partners of the listed targets genes.
Among the 36 credible genes, two (*ALDH2* and *ESR1*) were within this list of 139 lithium-related genes. Since the credible gene list is primarily derived from our fine-mapping analysis, it is possible that lithium target genes (and interaction partners) may be within loci for which significant fine-mapped putatively causal SNPs were not identified. It is also possible that a lack of specificity exists in the dataset for lithium target genes and that the geneset needs to be improved.

Rare variant enrichment analysis

8) * The statistical evidence for the convergence between common and rare variants is borderline. The authors report that the 80 prioritized genes for BD were enriched for ultra-rare damaging missense and loss of function variants observed in BD and schizophrenia cases. The effect size for enrichment is 1.16 (P=0.002) and 1.21 (P=0.02). How do these numbers compare to the enrichment of rare variants in genes linked to other psychiatric disorders such as major depression, autism, ADHD etc. or just a general list of brain-expressed genes and mutation-intolerant genes? Is the enrichment for rare variants in the bipolar GWAS genes observed in the background of brain-expressed genes or mutation-intolerant genes?

We agree that it is of interest to evaluate how these numbers compare to what has been identified for other psychiatric disorders. We found similar approaches had been undertaken for schizophrenia and ADHD, and include these as comparators.

In the 2022 SCHEMA paper, similar enrichment was observed for schizophrenia.[25]
"Statistical fine-mapping prioritized the likely underlying protein-coding gene at 64 of these associations (Supplementary Table 11, Supplementary Fig. 18), and we found case–control enrichment of URVs in these genes (Pmeta = 3.1 × 10−3; class I: OR = 1.37, 95% CI = 1.13–1.67)".

A preprint analysis of ADHD cases and controls shows similar enrichment as well.[26]
"Results from this burden analysis demonstrate a greater rate of both rare and ultra-rare de novo damaging variants (PTVs + Mis-D) in ADHD cases versus unaffected controls. For rare de novo damaging variants (non-neuro gnomAD allele frequency < 0.001), the rate ratio was 1.55 (95% CI 1.02-2.30). We found a greater difference between cases and controls when narrowing our analysis to ultra-rare de novo damaging variants (non-neuro gnomAD allele frequency < 0.00005), with a rate ratio of 1.86 (95% CI 1.21- 2.83, p=0.009)."

We have now added these references to the results text. New text is highlighted in yellow (page 11, Converging evidence of common and rare variation).

> "Within loci associated with BD in the multi-ancestry meta-analysis, the 71 genes annotated to putatively causal fine-mapped SNPs (Supplementary Table 29) were

enriched for ultra rare (<=5 minor allele count) damaging missense and protein-truncating variants in BD cases in the Bipolar Exome (BipEx) consortium dataset[27] (Odds ratio (OR) = 1.16, 95% confidence interval (CI) = 1.05 - 1.28, P = 0.002), and in schizophrenia cases in the Schizophrenia Exome Meta-analysis (SCHEMA) dataset[25] (OR = 1.21, 95% CI = 1.02 - 1.43, P = 0.024). This enrichment is similar to that observed for schizophrenia[25] and ADHD.[26] "

Using the same approach as taken in the SCHEMA[25] and BipEx[27] papers, background genes included all genes surveyed in each sequencing study, respectively.

We have now updated the text to clarify the approach.

New text added to the methods section (page 20, Convergence of common and rare variant signal):

"Using the same approach as taken in the SCHEMA[25] and BipEx[27] papers, background genes included all genes surveyed in each sequencing study, respectively."

9) * A previous ExWAS of BD, in combination with schizophrenia, has reported an association with AKAP11, which created excitement in the field as AKAP11 is known to interact with the GSK3B, a known target lithium. Speaking of convergence between common and rare variants, did the authors find any evidence of common variant associations at the AKAP11 locus? Or are there common variant associations at the other lithium target genes? Such lines of investigation will be insightful (irrespective of the positive or negative outcome) rather than the drug repurposing analysis.

We investigated if any lithium target genes, as well as their interaction partners, are among the 36 credible BD genes. We first used the DGIdb to extract all lithium-gene interactions. The extracted genes were then investigated in the latest version of the human protein interactome from the Barabasi lab focusing on high-quality protein-protein interactions[28]. Among the 36 credible genes, two (*ALDH2* and *ESR1*) were within this list of 139 lithium-related genes. The gene *AKAP11* (13:42846288:42897397, build GRCh37) is not within any of our GWS loci, and was not implicated in any of our locus to gene mapping strategies. The nearest locus to this gene identified in the multi-ancestry meta-analysis of all cohorts is locus 213 (index SNP rs9591009 13:47684394:47818121).

To examine if the lithium target genes and the credible BD genes show a significant overlap in the human protein interactome, we calculated the network-based separation (S_{AB}) between BD genes and lithium target genes, where an overlapping network neighborhood would be indicative of functional similarity[29]. The results from this analysis do not indicate any overlap between BD genes and lithium target genes in the human protein interactome (S_{AB}=0.124, z-score=1.710, p=0.044). The positive SAB value indicates that the lithium target genes and the

credible BD genes are separated from each other in the network of protein-protein interactions (significantly separated, as $p<0.05$), so there is no support for a biological relationship between them.

The following text was added to the methods section (page 23, Drug enrichment analyses)

> *"Moreover, we investigated if any lithium target genes, as well as their interaction partners, were among the 36 credible genes using the latest version of the human protein interactome[28]. We calculated the network-based separation (S_{AB}) between credible genes and lithium target genes, where a significant overlapping network neighborhood would be indicative of functional similarity[29]."*

The following text was added to the results (page 12, Drug target analyses):

> *"Among the 36 credible genes, two (ALDH2 and ESR1) were within the list of 139 lithium target and interaction partner genes. The results of the network-based separation (S_{AB}) analysis do not indicate a general overlap between the credible genes and lithium target genes in the human protein interactome (S_{AB}=0.124, z-score=1.710, p-value=0.044). The positive S_{AB} value indicates that the lithium target genes and the 36 credible genes are separated from each other in the network of protein-protein interactions."*

Tissue and cell type enrichment analysis

10) * I expected that the increase in sample size would have powered the tissue and cell-type specific enrichment analysis revealing new insights. In the single-cell enrichment analysis, the authors write that, among the non-brain cell types, they find enrichment for enteroendocrine cells of the large intestine and delta cells of the pancreas. What is the relevance of this finding to bipolar disorder?

We have now revised the cell type enrichment analysis. In addition to providing relevance of these findings to bipolar disorder, we also performed a sensitivity analysis to show that the results are independent of overlapping genes between the identified cell-types and those expressed in neurons.

The novel finding of enrichment in enteroendocrine cells may be of interest in bipolar disorder. When stimulated with short-chain fatty acids (SCFAs) these cells secrete serotonin, which may circulate to the brain[30,31]. In addition, lithium is shown to upregulate SCFA-producing bacteria[32], highlighting a potential mechanism of action via the gut-brain axis.

The following text has been added to the discussion to address this (page 13, Discussion):

> *"A novel finding is that single-cell enrichment analysis of non-brain murine tissues identified significant enrichment in the enteroendocrine cells of the large intestine and*

delta cells of the pancreas. Conditional analyses suggest that this enrichment is independent of overlapping genes between these cell-types and those expressed in neurons. Stimulation of enteroendocrine cells by short-chain fatty acids (SCFAs) promotes serotonin production in the colon which leads to enhanced levels of serotonin in systemic circulation and in the brain, and is a proposed mechanism by which microbiota influence the gut-brain axis.[30,31] Notably, lithium treatment is shown to upregulate SCFA-producing bacteria highlighting a potential mechanism of action.[32]"

11) The authors write that they found enrichment of neuronal populations. Are there differences in the tissue and cell type enrichments between the BD subtypes or between BD and schizophrenia or BD and depression etc.?

Based on the reviewer's suggestion, we have updated the framing of the paper, and investigated cell type enrichment in BD based on ascertainment and subtype. We observed similar patterns of enrichment across ascertainment and subtypes. We have amended the results and discussion sections to describe these findings.

Updated results text. New text highlighted in yellow (page 9, Pathway, tissue and cell type enrichment):

"Single-cell enrichment analyses of brain cell types indicate involvement of neuronal populations from different brain regions, including hippocampal pyramidal neurons and interneurons of the prefrontal cortex and hippocampus (Supplementary Figure 5), and were largely consistent with findings from the previous PGC BD GWAS[1]. Similar patterns of enrichment were observed based on ascertainment and subtype (Supplementary Figure 6)."

Moreover, we now provide a comparison of our results from the single-nucleus RNA-seq analysis with those obtained for depression and schizophrenia using the same data.[3]

Updated discussion text. New text highlighted in yellow (page 13, Discussion):

"Enrichment of the common variant associations from this multi-ancestry meta-analysis highlights the synapse, interneurons of the prefrontal cortex and hippocampus, and hippocampal pyramidal neurons as particularly relevant. Exploratory analyses using GSA-MiXeR[20] suggest enrichment of dopamine- and calcium-related biological processes and development of GABAergic interneurons. These findings were further corroborated by enrichment analyses in single-nucleus RNA-seq data from adult postmortem brain tissue, which highlighted specific clusters of interneurons derived from the caudal and medial ganglionic eminences and medium spiny neurons predominantly localised in the striatum. Medium spiny neurons are not enriched in depression using the same dataset.[3] Although interneurons derived from ganglionic eminences were also enriched in

schizophrenia, stronger signals were observed for amygdala excitatory and hippocampal neurons.[3]"

Non-European ancestries

12) * The addition of non-European cohorts to the new meta-analysis is great. However, I am not convinced that the added samples had any major impact on the results except for a slight increase in improvement in fine mapping. The new genome-wide significant locus identified in the East Asian cohort is interesting, but it's not clear if it will replicate. Also, no clear biological link between genes at this locus (TET2, CXXC4) and bipolar disorder.

The lack of representation of non-European ancestry cohorts is an issue within the GWAS field,[33] and so the inclusion of such samples is of high importance, even if the results do not substantially differ. As indicated by the reviewer, this led to improved fine-mapping, suggesting that the inclusion of additional non-European ancestry cohorts will continue to improve our understanding of BD genetics.

To provide more context for the novel East Asian associated locus we searched the GWAS catalog utilizing the locus boundaries and identified additional complex traits with overlapping implicated loci, including verbal-numerical reasoning measurement, sleep duration, and educational attainment. We have now removed the text related to TET2 and CXXC4 since these genes were approximately 300 kb up- and downstream of lead SNP, respectively, with little evidence to support a role in BD.

The following text has been added to the discussion to address this (page 14, Discussion):

"Investigation of the novel ancestral-specific association in the EAS meta-analysis in the GWAS catalog[34] highlights overlaps with genome-wide significant loci for verbal-numerical reasoning measurement[35], sleep duration[36], and educational attainment[37], as well as a suggestive locus ($p < 2 \times 10^{-6}$) for the interaction between cognitive function and MDD[38]. These findings suggest a role for this genomic region in complex brain-related phenotypes."

13) * In the discussion the author writes that "The multi-ancestry PRS provided the greatest improvement over the EUR-PRS in an AFR target cohort, achieving an over two fold greater increase in variance explained". But the description in the results section does not agree with this statement: "When the multi-ancestry PRS was applied to a clinically ascertained AFR target cohort, … the inclusion of self-report data greatly increased the explained variance (no self-report R2 = 0.010, with self-report R2 = 0.023). A similar pattern was observed for the EUR ancestry PRS (no self-report R2 = 0.010, with self-report R2 = 0.022)" According to the above statement, both EUR PRS and multi ancestry PRS explain similar R2. Only the inclusion of self-report cases increases the R2 value from 0.01 to 0.02. Am I interpreting that right?

The reviewer is correct, the observed increase in R2 in the AFR target cohort seems due to the greater power resulting from the increased sample size.
Given that the AFR target cohort is likely the most genetically heterogeneous of the target cohorts[39], it is possible that the similarity (in terms of genetic architecture) between the self-report AFR samples and the target AFR cohort is not as high as, for example, between the self-report EAS and target EAS cohorts. Thus, the increase in sample size and power increases the signal irrespective of ancestry representation in the training data. However, without access to the individual-level genetic data for the self-report AFR cohort this is difficult to investigate. Thus, we choose not to speculate on this in the discussion.
We have now updated the text to more accurately reflect what is reported in the results.

Updated discussion text. New text is highlighted in yellow (page 14, Discussion):

> *"The multi-ancestry PRS provided the greatest improvement over the EUR-PRS in two of the three EAS target cohorts (Korean and Taiwanese). More subtle improvements were seen when the EUR target cohorts were analysed. Multi-ancestry training data provided little improvement in the AFR target cohort, which may be due to the genetic heterogeneity of this target cohort[39]. These results highlight the benefits of multi-ancestry representation in the PRS training data, in line with findings from other diseases[40]."*

Less technical and more biological discussions needed

14) * The manuscript overall reads more like a technical paper, spending more time on describing the differences between self-report vs clinical case cohort. I'd rather be interested more in discussions on interesting loci and the nearby genes and their associated pathways and their biological links to bipolar disorder.

We agree with the reviewer that the manuscript can be improved with more discussions on interesting loci and the nearby genes and their associated pathways, and the biological links to bipolar disorder. Thus, we have now revised and restructured the results and discussion sections of the manuscript, with more emphasis on biological interpretations and implications. This includes a revised section focused on 'Pathway, tissue and cell enrichment" which includes results from new analyses of additional QTL and single-nucleus RNA sequencing data.
We have also now included additional data types for gene-mapping in our credible gene identification. These include, mQTL, sQTL and enhancer-promoter interactions. Based on these updated analyses we identify the *SP4* gene as our top credible gene (identified by 6/7 approaches used) and we show that this gene is specifically regulated in astrocyte and GABAergic neurons by the genome-wide significant (GWS) variant rs2107448.

We have added the following text to the results (page 11, Identification of credible BD-associated genes)

"The SP4 gene was identified by six of these approaches, and astrocyte and GABAergic neuron specific regulation of SP4, by the GWS variant rs2107448, were identified from cell-type specific enhancer-promoter interaction results (Supplementary Table 32). Moreover, the TTC12 and MED24 genes were identified by five of the approaches. Eight of the 36 credible genes have synaptic annotations in the SynGO database.[10] Three genes (HTT, ERBB4 and LR5NF) were mapped to both postsynaptic and presynaptic compartments. One gene (CACNA1B) was mapped to only the presynapse and four genes (SHANK2, OLFM1, SHISA9 and SORCS3) were mapped to only the postsynapse (Supplementary Table 32)."

We have added the following text to the discussion (page 13, Discussion):

"The top credible gene, identified by six gene-mapping approaches, was SP4, which has also been implicated in schizophrenia through both rare[25] and common variation.[41]"

Moreover, we have also examined the temporal expression of identified credible BD-associated genes highlighting two distinct clusters with changes in expression before birth.

Still, we believe that the differences in genetic architecture between ascertainment and subtypes represent some of the key findings in this study and are important to discuss.

Referee #2 (Remarks to the Author):

Review of the manuscript: “Diversity fuels gene discovery and biological insight in bipolar disorder”.

The manuscript presents results from a large multi-ancestry GWAS of bipolar disorder (BD) including 158,036 BD cases and 2.8 million controls, identifying 298 genome-wide significant loci. This is a considerable increase in number of samples analyzed and identified loci compared to the latest published GWAS which included 41,917 BD cases, 371,549 controls and identified 64 loci. The major increase in sample size is primarily due to a large contribution from 23andMe of data on self-reported BD (90,088 cases and 1,928,789 controls).
The study includes a careful evaluation of how ascertainment (“clinical”, “community” and “self-report”) affects the results. Besides the multi-ancestry GWAS, the authors also report results from ancestry specific GWAS of individuals with EUR, EAS, AFR and LAT ancestries, the latter three groups compose 17,7% (proportion of Neff) of the samples.
Enrichment analyses highlight potential affected tissues and cell types, and 47 BD risk genes are highlighted based on convergent evidence from seven different methods that link risk variants to genes by integration with functional genomics data.
The study contributes to our understanding of the genetics underlying BD and represents the next step towards identification of the biological underpinnings of BD, not only due to the current results but also from the wave of future studies that will be based on the new GWAS results.

I have some comments and questions:

1) The analyzed samples are collected using three overall approaches which the authors refer to as “clinical”, “community” and “self-report”. For most of the analyses the authors present results from analyses both with and without the self-reported samples. It would have made the manuscript and messages clearer if there had been one main phenotype/GWAS and the other analyses e.g. “excluding the self-reported data” were presented as sensitivity analyses in the supplementary. However, I understand the authors choice of reporting both versions of the results, due to the low genetic correlation between self-reported and clinical samples (rg=0.47), and the observation that only around 50% of self-reported influencing common variants are shared and concordant with variants influencing clinically ascertained BD. Self-reported data demonstrates considerable high genetic correlation with the community samples which justifies the inclusion in the meta-analysis.
However, I think it would be relevant to investigate the heterogeneity between the clinical and self-reported data bit further (as sensitivity analyses in the supplementary). What are the genetic correlations of the self-reported data alone with other psychiatric disorders compared to the genetic correlations of the clinical samples alone with other psychiatric disorders?

> We agree with the reviewer about clarifying the overall design and approach of the paper. We have now revised the presentation of the results and discussion to clarify (see above). Further, we have investigated the BD spectrum heterogeneity (subtypes and ascertainment) to provide a better understanding of the differences in underlying genetic architecture.

We have generated a new plot to show the differences in genetic correlation, as suggested by the reviewer, and reviewer 3 comment 3 below. The figure is below and Figure 2 in the revised main text. We have also expanded on these in the results and discussion.

Updated results text. New text highlighted in yellow (page 6, Genetic correlations with other traits):

> *"Genome-wide genetic correlations (r_g) were estimated between EUR ancestry BD GWASs (with and without self-reported data, and when stratified by ascertainment and subtypes) and human diseases and traits via the Complex Traits Genetics Virtual Lab (CTG-VL; https://vl.genoma.io) web platform[42] (Figure 2, Supplementary Tables 13-15). Most psychiatric disorders, including major depressive disorder (MDD), post-traumatic stress disorder (PTSD), attention deficit/hyperactivity disorder (ADHD), borderline personality disorder, and autism spectrum disorder (ASD), were more strongly correlated with the full meta-analysis, as well as with bipolar disorder type II, and BD in community and self-reported samples (Figure 2). In contrast, schizophrenia was more strongly correlated with the full BD meta-analysis excluding self-reported data and with bipolar disorder type I and BD in clinical samples (Figure 2)."*

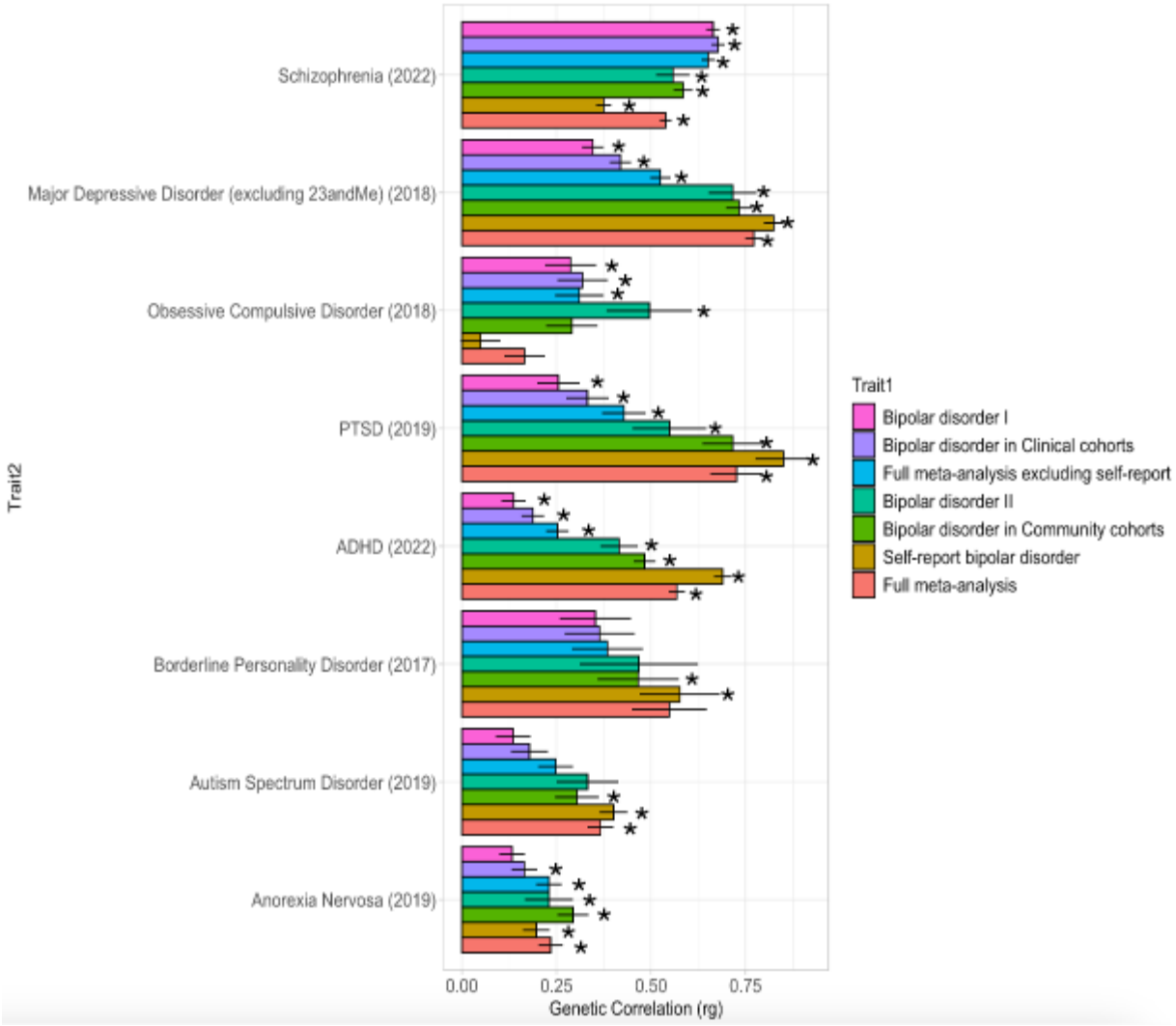


Figure 2. *Genetic correlations between bipolar disorder and other psychiatric disorders. The y-axis (Trait2) is ordered based on the significance and magnitude of genetic correlation of each trait with bipolar disorder type I. Stars indicate results passing theBonferroni corrected significance threshold of $P < 3.6 \times 10^{-5}$. ADHD, attention deficit/hyperactivity disorder. PTSD, post-traumatic stress disorder. Details are provided in Supplementary Table 13.*

Discussion text (page 14, Discussion):

> *"Moreover, the pattern of correlations between BD and other psychiatric disorders differed with the inclusion of self-reported data. Schizophrenia had the highest correlation with BD without the inclusion of the self-reported data, while major depressive disorder was most strongly correlated with BD after the inclusion of the self-reported data."*

2) Related to my point above – when reading the "Multi-ancestry meta-analysis" section it would be helpful to know the reason why results from both with- and without self-reported BD are

reported. I would therefore suggest that the section that evaluates the genetic heterogeneities between ascertainment approaches comes before the GWAS results. I know that this does not follow the traditional structure in GWAS papers, but it will give a better understanding of the reporting of GWAS results with and without the self-reported data.

We agree with the reviewer that reporting the genetic heterogeneities between ascertainment approaches and subtypes before the GWAS results will clarify why results are reported with and without the self-reported samples. In the revised manuscript we have made this change, in addition to other restructuring of the text to better frame the results of the study.

3) What is the proportion of BDI and BDII among the clinical and community ascertained individuals? I could not locate that information.

There are 23,158 cases within the Clinical ascertainment group, and 8,878 cases within the Community ascertainment group for whom subtype data are available (extracted from Supp Table 1). 82.5% of cases in the Clinical ascertainment group were BDI, while only 68.7% of cases in the Community ascertainment group were BDI. These data are shown in the table below, and this information is now included in Supplementary Table 2.

Ascertainment	Case N	BDI N (%)	BDII N (%)
Clinical	23,158	19,115 (82.5)	4,043 (17.5%)
Community	8,878	6,101 (68.7)	2,777 (31.3)

X^2=730.51, p < 2.2e-16

We have also added the following text to the results (page 3, Study populations)

"BD subtype data were available for a subset of individuals within the Clinical and Community groups. 82.5% of cases in the Clinical ascertainment group were BDI, while only 68.7% of cases in the Community ascertainment group were BDI (X^2=730.51, p < 2.2e-16; Supplementary Table 2)."

Moreover, we examined how BD subtypes within Clinical and Community cohorts affected their genetic correlation with Self-reported BD. To do so, we conditioned the Clinical and Community meta-analyses on the genetic risk for BDI and BDII.

We have added the following text to the results (page 5, Genetic architecture of BD differs by ascertainment and subtype)

"Given the difference in proportion of BDI and BDII cases within the Clinical and Community cohorts, we evaluated the genetic correlation between BD within Clinical and Community cohorts, and Self-reported BD, after conditioning on the genetic risk for BDI and BDII. After conditioning, the genetic correlation between Self-reported BD and BD within Community cohorts (rg = 0.92; s.e. = 0.09) = was not significantly different (p = 0.10) than with BD in Clinical cohorts (rg = 0.71; s.e. = 0.13). "

We have also added the following text to the discussion (page 13, Discussion)

"The low genetic correlation and minimal genetic overlap between cases ascertained through clinical studies and cases with self-reported BD is driven by a greater proportion of BDI within the Clinical and Community samples."

4) On page 3 the authors write: "There was minimal evidence of residual population stratification (LDSC intercept = 1.052 (se = 0.016), attenuation ratio=0.071 (s.e. = 0.013))." As far as I have understood the result is based on the multi-ancestry GWAS, so I find the wording a bit strange. Since the analysis includes samples with different ancestries, we would expect population stratification, I assume.

We agree with the reviewer that this statement requires clarification. Our choice of wording was poor. We agree that the sample contains population stratification. What we intended to convey is that this analysis shows that we have effectively controlled for population stratification such that any residual contributions to the genome wide test statistics are small, and that the overall genome wide test statistic inflation above the null represents true polygenic association signal.

Updated results text. New text highlighted in yellow (page 5, Multi-ancestry meta-analysis):

"There was minimal test statistic inflation due to uncontrolled population stratification after correction for principal components in each dataset (LDSC intercept = 1.052 (se = 0.016), attenuation ratio=0.071 (s.e. = 0.013))."

Updated results text. New text highlighted in yellow (page 6, Multi-ancestry meta-analysis):

"There was minimal test statistic inflation due to uncontrolled population stratification after correction for principal components in each dataset (LDSC intercept = 1.050 (se = 0.012), attenuation ratio=0.086 (s.e. = 0.018))."

5) Figure 1, I think it should be added to the figure legend that the locus marked with a star is the genome-wide significant locus in EAS.

We agree with the reviewer. The following text has been added to the figure legend for Supplementary Figure 1.

"The star indicates the position of the East Asian GWS locus (rs117130410, 4:105734758, build GRCh37)."

6) Most supplementary tables miss the explanation of the column headers. All headers should be explained not only those that the authors think are essential.

We agree with the reviewer that the column headers should be explained. We have now included explanations for all headers in all tables.

7) What is the genetic correlation of BD between the ancestries? – this could be evaluated with e.g., POPCORN. I think this would be informative also for future GWAS. It would probably make most sense to evaluate this using the self-reported individuals, due to lack of clinical samples with non-European ancestry.

We have now computed cross-ancestry genetic correlations as suggested by the reviewer (And reviewer 3, comment 1b, below). These are included below and in Supplementary Table 3. Given the large standard errors on many of these correlation results they are not currently informative.

Genetic correlations between EUR, EAS, AFR and LAT meta-analyses using popcorn		
Trait1	Trait2	Genetic impact correlation (se)
EUR meta-analysis including self-report data	EAS meta-analysis including self-report data	0.71 (0.17)
EUR meta-analysis excluding self-report data	EAS meta-analysis excluding self-report data	0.83 (0.17)
EUR meta-analysis including self-report data	AFR meta-analysis including self-report data	1.19 (0.53)
EUR meta-analysis excluding self-report data	AFR meta-analysis excluding self-report data	0.67 (0.82)
EAS meta-analysis including self-report data	AFR meta-analysis including self-report data	0.06 (0.62)
EAS meta-analysis excluding self-report data	AFR meta-analysis excluding self-report data	-
EUR meta-analysis including self-report data	LAT meta-analysis including self-report data	0.77 (0.05)
EAS meta-analysis including self-report data	LAT meta-analysis including self-report data	0.59 (0.20)
AFR meta-analysis including self-report data	AFR meta-analysis excluding self-report data	-

'-' = the correlation could not be computed.

We have also added the following text to the results section (page 5, Ancestry-specific GWAS meta-analyses):

"Ancestry-specific estimates of SNP-heritability and cross-ancestry genetic correlations are provided in Supplementary Table 3."

8) In relation to my point above, what is the SNP-h2 of the other ancestries? I assume there should be enough power to calculate this using LDSC, at least for individuals with LAT ancestry.

As with the above suggestion, we have now also computed the SNP-h2 for the other ancestral meta-analyses. These are included below and in Supplementary Table 3.

Liability scale SNP-heritability estimates in EUR, EAS, AFR and LAT meta-analyses using popcorn	
SNP-heritablity (se) using 2% population prevalence of bipolar disorder for conversion to liability scale	
EUR meta-analysis excluding self-report data	0.095 (0.012)
EUR meta-analysis including self-report data	0.065 (0.006)
EAS meta-analysis excluding self-report data	0.099 (0.035)
EAS meta-analysis including self-report data	0.060 (0.021)
AFR meta-analysis excluding self-report data	0.025 (0.075)
AFR meta-analysis including self-report data	0.021 (0.027)
LAT meta-analysis including self-report data	0.063 (0.012)

9) Page 10, the authors write "Within loci associated with BD in the multi-ancestry meta-analysis, genes (N = 80) implicated by putatively causal fine-mapped SNPs (Supplementary Table 31)". In supplementary Table 31 there are much more than 80 genes, so please highly the 80 genes included in the rare-variant analysis.

We agree with the reviewer that this information was difficult to retrieve from the initial Supplementary Table 31. We have now revised all supplementary tables and removed additional unnecessary information. The list of genes included in the rare variant analysis are now listed in Supplementary Table 29 "Fine-mapping results for multi-ancestry meta-analysis including self-reported data".

Moreover, the original submission indicated 80 genes identified via fine-mapped SNPs; however, this was a typographical error. The correct number, and that which was used for all analyses, is 71 genes. The manuscript has been updated accordingly.

10) Add sample sizes to Supplementary Table 9 (the SNP-heritability table).

Sample sizes have now been added for all meta-analyses now presented in Supplementary Table 3.

11) Page 5, add sample sizes to the sentence: "We further stratified Clinical BD samples into BDI and BDII subtypes and performed similar analyses".

We have now updated the sentence to reflect the number of BDI and BDII cases.

New results text (page 4, Genetic architecture of BD differs by ascertainment and subtype):

"To analyse BD subtypes, we used available GWAS summary statistics for BDI (25,060 cases) and BDII (6,781 cases)[1], which come from a subset of the Clinical and Community samples."

12) Page 5 and Supplementary Table 10, MiXeR: I do not see strong support for the "best" model vs the "max"/"min" models in many of the analyses (i.e., negative AIC/BIC values). I think this needs to be addressed in the manuscript, and can the authors comment on why they think the analyses are valid.

The reviewer is correct that some of the models, mostly those incorporating the BD I and II data, do not have strong support for 'best' model vs the 'max'/'min' models. We have now amended the figure legend to reflect this.

Updated figure legend. New text highlighted in yellow (page 4, Genetic architecture of BD differs by ascertainment and subtype):

*"**Figure 1.** Genetic correlation and bivariate MiXeR estimates for the genetic overlap of BD ascertainment and subtypes. Trait-influencing genetic variants shared between each pair (grey) and unique to each trait (colours) are shown. The numbers within the Venn diagrams indicate the estimated number of trait-influencing variants (and standard errors) (in thousands) that explain 90% of SNP heritability in each phenotype. The size of the circles reflects the polygenicity of each trait, with larger circles corresponding to greater polygenicity. The estimated genetic correlation (rg) and standard error between BD and each trait of interest from LDSC is shown below the corresponding Venn diagram. Clinical and Community samples were stratified into bipolar I disorder (BDI) and bipolar II disorder (BDII) subtypes if subtype data were available. Model fit statistics*

indicated that MiXeR-modeled overlap for bivariate comparisons including the bipolar subtypes (BDI and BDII) were not distinguishable from minimal or maximal possible overlap, and therefore to be interpreted with caution (see Supplementary Table 4).”

13) Supplementary Table 10, tri-variate MiXeR, it is not clear to me what column one represents, e.g. what does “bdCommunity_bdclinical_bdselfreport_12.json” mean? please use another term or explain in the header.

We agree with the reviewer that this was not clear. In this case “bdCommunity_bdclinical_bdselfreport_12.json” refers to the 12th run (out of 16) used to compute the trivariate estimates. We have renamed this as Run_12, and applied the same to the other runs. We have also provided additional explanation for all headers in this Supplementary Table and the others.

14) Page 5: the authors could consider to include block-jack-knife estimated p-values for differences in rg to quantify their statements in the sentences “In contrast, the genetic correlation between BDI and Self-report BD (rg = 0.42 (s.e. = 0.02)) was considerably lower than between BDII and Self-report BD (rg = 0.76 (s.e. = 0.05))”, and “… the genetic correlation for Self-report BD was considerably greater with Community samples (rg = 0.79 (s.e. = 0.02)) compared to Clinical samples (rg = 0.47 (s.e. = 0.02))..”.

The following has now been added to the methods and results sections of the manuscript to address this.

New methods text (page 17, Heritability and Genetic Correlation):

“Differences in rg between phenotype pairs were tested as a deviation from 0 using the block jackknife approach implemented in LDSC.[43]”

Updated results text. New text highlighted in yellow (page 3, Genetic architecture of BD differs by ascertainment and subtype):

“While there was a strong genetic correlation between Clinical and Community samples (r_g = 0.95; s.e. = 0.03), the genetic correlation for Self-reported BD was significantly greater ($p = 7.4 \times 10^{-28}$) with Community samples (r_g = 0.79; s.e. = 0.02) than with Clinical samples (r_g = 0.47; s.e. = 0.02) (Supplementary Figure 2).”

Updated results text. New text highlighted in yellow (page 4, Genetic architecture of BD differs by ascertainment and subtype):

"In contrast, the genetic correlation between BDI and Self-reported BD (r_g = 0.42; s.e. = 0.02) was significantly lower ($p = 7.1 \times 10^{-13}$) than between BDII and Self-reported BD (r_g = 0.76; s.e. = 0.05) (Supplementary Figure 2)."

In line with the above, we also provide a similar block jackknife p-value for the comparison of genetic correlation after conditioning on BD subtypes (see reviewer 2 comment 3 above) (page 5, Genetic architecture of BD differs by ascertainment and subtype).

"Given the difference in proportion of BDI and BDII cases within the Clinical and Community cohorts, we evaluated the genetic correlation between BD within Clinical and Community cohorts, and Self-reported BD, after conditioning on the genetic risk for BDI and BDII. After conditioning, the genetic correlation between Self-reported BD and BD within Community cohorts (rg = 0.92; s.e. = 0.09) = was not significantly different (p = 0.10) than with BD in Clinical cohorts (rg = 0.71; s.e. = 0.13)."

15) On page 7 the authors write: "The variance explained by the multi-ancestry GWAS without the self-report data was significantly greater than that explained by the multi-ancestry GWAS including self-report data (R2 = 0.058, s.e. = 0.017, P = 2.72 × 10−4), and the EUR ancestry GWAS excluding the self-report data (R2 = 0.084, s.e. = 0.018, P = 5.62 × 10−3) (Supplementary Table 19)." Could the authors add information about the variance explained (i.e. the R2 estimate) by the multi-ancestry GWAS without the self-report data, to the text.

We have now added this information to the results text.

Updated results text. New text highlighted in yellow (page 7, Polygenic prediction of bipolar disorder):

"The variance explained by the multi-ancestry GWAS without the self-report data (R2 = 0.090, s.e. = 0.019) was significantly greater than that explained by the multi-ancestry GWAS including self-report data (R2 = 0.058, s.e. = 0.017, P = 2.72 × 10−4), and the EUR ancestry GWAS excluding the self-report data (R2 = 0.084, s.e. = 0.018, P = 5.62 × 10−3) (Supplementary Table 19)."

16) On page 9 the authors write: "…indicating involvement of neuronal populations from different brain regions, including hippocampal pyramidal neurons and interneurons of the prefrontal cortex and hippocampus". It would be informative to know more about the brain data in the result section (e.g. pre- or post-natal tissues). In addition to this could the authors discuss the potential brain developmental stages affected (if they have the results to do so), it would be very interesting to know if genes affected by common risk variants play a role in early brain development or later in life.

We agree with the reviewer that the addition of details regarding developmental stages would be of interest. In addition to the enrichment analyses in the GTeX data, we have now also

included enrichment analyses within the Brainspan dataset, which includes brain-tissue from 524 samples over 11 developmental stages (from early prenatal to middle adulthood). This provides new insight into the early- to mid-prenatal stages and identifies two clusters of genes with changes in expression before birth, implicating early neurodevelopmental aspects in BD.

The following text has been added to the results (page 9, Pathway, tissue and cell type enrichment):

> *"The association signal was enriched among genes expressed in the brain (Supplementary Tables 24), and specifically in the early- to mid-prenatal stages of development (Supplementary Table 25). "*

In addition to the enrichment of the common variation signal in brain tissue from different developmental stages, we also examined patterns of temporal variation in expression of the identified credible genes using the lifespan gene expression data from the Human Brain Transcriptome project (www.hbatlas.org).[11] We identified two clusters of genes with changes in expression before birth.

The following text has been added to the results (page 12, identification of credible BD-associated genes):

> *"Based on the lifespan gene expression data from the Human Brain Transcriptome project (www.hbatlas.org),[11] we identified two clusters of credible genes with distinct peaks in temporal expression (Supplementary Figure 10, Supplementary Table 31). The first cluster shows reduced prenatal gene expression, with gene expression peaking at birth and remaining stable over the life-course. Conversely, the second cluster shows a peak in gene expression during fetal development with a drop-off in expression before birth."*

The following text has been added to the discussion (page 13, Discussion):

> *"Moreover, we clustered the credible genes based on similar patterns of temporal variation in expression over the lifespan and identified two clusters. The first cluster shows reduced gene expression pre-birth, with gene expression peaking at birth and remaining stable over the life-course, while the second cluster shows a peak in gene expression during fetal development aligning with the neurodevelopmental hypothesis of mental disorders.[44]"*

The following text has been added to the methods (page 22, Clustering of credible genes by temporal variation)

"Lifespan gene expression from the Human Brain Transcriptome project (www.hbatlas.org)[11] was used to cluster the list of credible genes based on their temporal variation. The gene expression and associated metadata were acquired from the gene expression omnibus (GEO; accession: GSE25219). The data consists of 57 donors aged 5.7 weeks post conception to 82 years old with samples extracted across regions of the brain. Prior to filtering gene expression for the list of credible genes, gene symbols of both credible genes and the gene expression dataset were harmonized using the "limma" package in R which updates any synonymous gene symbols to the latest Entrez symbol. Gene expression was available for 33 of the 36 credible genes. Within a given brain region, each gene's expression was then mean-centered and scaled. Outliers in gene expression more than 4 standard deviations from the mean were removed. To generate a single gene expression profile for each gene across the lifespan, at a given age, the mean gene expression for a given gene was taken across brain regions, and in some cases across donors. This resulted in a matrix where each gene had a single expression value for each age across the lifespan. This gene-expression by age matrix was then used to cluster the credible genes by the lifespan expression profiles using the R package "TMixClust". This method uses mixed-effects models with nonparametric smoothing splines to capture and cluster non-linear variation in temporal gene expression. We tested K=2 to K=10 clusters performing 50 clustering runs to analyse stability. The clustering solution with the highest likelihood (i.e., the global optimum using an expectation maximization technique) is selected as the most stable solution across the 50 runs for each of the trials testing 2-10 clusters. We compare the average silhouette width across the K=2 to K=10 clusters and select that with the maximum value as the optimal number of clusters."

17) Supplementary Figure 4, I think the figure needs a better explanation. It is not clear to me how the CS-values/the colors relate to the PIP threshold (or if there is a relation).

We have updated the figure legend to clarify that all SNPs regardless of PIP threshold were used to assess the size of each credible set. We have also added an explanation for the colours used.

New figure legend:

"Supplementary Figure 8: Number of SNPs within the smallest 95% credible sets (CS) from meta-analysis of European and multi-ancestry meta-analyses when excluding and including self-report data. Colours represent CS of varying size, with blue CS containing 0 SNPs and red CS containing 15+ SNPs. All fine-mapped SNPs regardless of their PIP were used to assess the size of the 95% credible sets."

18) Page 10, was the VEP annotation of causal variants based on position alone?

For VEP annotation, genes are assigned to variants based on their position relative to annotated Ensembl transcripts and known regulatory features (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences). The text has been updated to reflect this.

Updated results text. New text highlighted in yellow (page 11, Fine-mapping):

> *"Putatively causal SNPs with a PIP > 0.5 were mapped to genes by performing variant annotation with Variant Effect Predictor (VEP) (GRCh37) Ensembl release 109, based on their position relative to annotated Ensembl transcripts and known regulatory features."*

19) On page 14 the authors write "Thus, the previous BDI and BDII GWAS summary statistics data were used for BDI and BDII analyses in this study." Add information about sample sizes.

We have now updated the sentence to reflect the number of BDI and BDII cases.

Updated methods text. New text highlighted in yellow (page 16, Sample description):

> *"Thus, the previous BDI (25,060 cases and 449,978 controls) and BDII (6,781 cases and 364,075 controls) GWAS summary statistics data were used for BDI and BDII analyses in this study."*

20) Have the authors evaluated the genetic correlation between males and females based on sex-stratified BD GWAS?

Other ongoing work by PGC BD indicated the genetic correlation between males and females is 0.92 (se = 0.03), consistent with what has been reported previously[45].
However, including sex-stratified analyses were considered beyond the scope for this current manuscript.

21) Page 18, the authors write "TWAS p-values were Bonferroni-corrected for the number of genes included in the respective analysis." Please state the p-value threshold for declaring significance and the number of genes corrected for.

This has now been added to the methods text for TWAS which was moved to the supplementary note.

New methods text (Supplementary note page 2, Detailed description of TWAS, FOCUS and isoTWAS eQTL analyses):

> *"... TWAS p-values were Bonferroni-corrected at a p-value threshold of $3 \times 10{-}6$, derived by taking the nominal alpha level of 0.05 and correcting for the 14,773 genes tested."*

22) Page 18, isoTWAS results: the authors write “We then performed probabilistic fine-mapping of the significant associations, …” how did the authors define significant associations and did they apply Bonferroni correction?

We have now revised the paragraph describing the isoTWAS procedure for clarification. These details were moved to the supplementary note.

Updated text. New text highlighted in yellow (Supplementary note page 2, Detailed description of TWAS, FOCUS and isoTWAS eQTL analyses):

“We also applied the recently developed isoform-level TWAS method (isoTWAS)[7], using precomputed weights per isoform derived from adult PsychENCODE data, as provided by the isoTWAS developers (https://zenodo.org/record/6795947#.Y8mi2-zMLBI). In total we tested the 7,530 genes, each with varying numbers of isoforms, that had positive heritability with a p-value < 0.05 within the PsychENCODE data. We subset the provided PsychENOCDE SNP-isoform weights and GWAS summary statistics SNPs to those that intersect with the HRC reference panel SNPs, then perform the isoTWAS burden test to generate z-scores for each SNP-isoform pair. We then performed probabilistic fine-mapping, filtering the resulting SNP-isoform statistics to those isoforms with a burden z-test p-value < 0.05 and a permutation p-value < 0.05. The permutation p-value was generated by performing 100 random shuffles of SNP-to-isoform weights to generate a null distribution. Credible sets were defined as SNPs passing these two p-value thresholds and having a posterior inclusion probability (PIP) >= 0.8.”

23) Very little information is provided about the exome-sequencing data. Even though the authors refer to papers explaining these data it would help the reader to have more information. E.g. I could not locate the sample size of the individuals included in the sequencing studies in the manuscript and how damaging missense and protein-truncating variants are defined.

We have now updated the methods section to include these omitted details.

Updated methods text. New text is highlighted in yellow (page 20, Convergence of common and rare variant signal):

“Data from the Bipolar Exome (BipEx) consortium[27] (13,933 bipolar cases and 14,422 controls) were used to assess the convergence of common and rare variant signals, using a similar approach as previously used for schizophrenia[41]. This dataset includes approximately 8,2 k individuals with BDI and 3,4 k individuals with BDII, while the remainder of the sample lack BD sub-type information. Ultra rare variants (<=5 minor allele count) for damaging missense (missense badness, PolyPhen-2 and regional constraint (MPC) score >3) and protein-truncating variants (including: transcript ablation, splice acceptor variants, splice donor variants, stop gained and frameshift

variants) were considered. An enrichment of rare variants in genes prioritised through fine-mapping in cases relative to controls were assessed using a Fisher's Exact Test. Given the genetic overlap between bipolar disorder and schizophrenia, we repeated the analysis in data from the Schizophrenia Exome Meta-analysis (SCHEMA) cohort (24,248 schizophrenia cases and 97,322 controls)[25]."

24) In the enrichment analysis of exome-sequencing data the authors use a Fisher´s exact test. I understand the wish to apply Fisher´s exact test to increase power but I think it would be more appropriate to apply logistic regression which allows for the inclusion of covariates to correct for technical issues related to the sequencing data and potential remaining population stratification. Could the authors comment on their choice of test?

The use of Fisher's exact test here follows the same approach as undertaken in the SCHEMA[25] and BipEx[27] papers.
For the BipEx, SCHEMA and GWAS analyses, necessary covariates were already included. Moreover, since the counts are so rare we don't expect inclusion of covariates to change the result much, and logistic regression may end up miscalibrated with such small counts.

Referee #3 (Remarks to the Author):

O'Connell et al produce the largest GWAS to date of BD and identify many new loci. This is an impressive sample size, but I have several comments particularly regarding the frame and readability for the authors.

1. The frame of the manuscript as written is oriented around the multi-ancestry analysis of BD. However, from Table S2, 69/79 cohorts are of European descent (83% of total participants) with 131,969 cases and 2,322,416 controls, versus a combined total of 26,067 cases and 474,083 controls from all other ancestries. Further, there is only one genome-wide significant association in any non-European ancestry group. The vast majority of the non-European ancestry participants are self-report cases from 23andMe. A few questions/comments arise from this:
a. Is the multi-ancestry frame for this paper the strongest way to introduce this work, given that it consists of participants who are by a large majority of European descent? As framed, I'm surprised there is not more emphasis on the multi-ancestry fine-mapping results, although Figure S4 suggests CS size identified in the meta-analysis closely reflects the European only results, which is a bit underwhelming. Using the test statistics and LD information within vs across populations, some simulation or downsampling/projection work showing how many loci in the multi-ancestry meta-analysis were discovered that wouldn't have been discovered from a similar sample size of European ancestry group participants alone could be interesting.

> We agree that the multi-ancestry frame for this paper is not the strongest way to introduce this work, and have now revised the title and the approach, with less emphasis on the multi-ancestry results (see also response to reviewer 1, comment 1 and 2). This also involved changing the order of presentation.
>
> The reviewer is correct that the CS sizes identified in the multi-ancestry meta-analyses, although smaller, closely match those of the EUR ancestry meta-analyses. This may in part be due to the use of an EUR ancestry reference panel used for both multi-ancestry and EUR fine-mapping analyses due to a lack of in-house raw genetic data or in-sample LD estimates for all cohorts.
>
> With regard to the reviewers' suggestion of simulations or downsampling work, we do not believe that this could be performed accurately. As mentioned, the lack of access to raw genetic data prevents us from downsampling the meta-analyses while maintaining the same proportion across ancestral groups. Moreover, lack of raw data makes simulation of the multi-ancestry LD to match the current meta-analysis impossible.
>
> We have included a description of this limitation in the discussion (page 14):
>
> > *"One limitation is the lack of in-sample LD estimates for all cohorts, due to a lack of in-house raw genotype data for some cohorts. We used a EUR LD reference panel to*

analyse the multi-ancestry meta-analyses[46] where LD patterns and interindividual heterogeneity within the ancestry groups are not fully captured."

b. There seem to be significant limitations in polygenic cross-ancestry comparisons of genetic architecture in this study due to differences in phenotyping. The genetic correlations are far lower for self-report vs all phenotypes than for all other pairwise comparisons.

As also requested in reviewer 2 comment 7 above, we have now computed cross-ancestry genetic correlations. These are included below and in Supplementary Table 3, and in the response to reviewer 2 comment 7 above.

Further, we agree that the inclusion of self-report data does alter the pattern of genetic correlations. We have included this data and a new figure addressing this point in our response to comment 3 below.

c. Is the deeper phenotype dissection a better frame for the paper? For a broad readership beyond a journal focused on psychiatry, introducing the differences between BDI and BDII is important, since this is unlikely common knowledge.

We agree with the reviewer that the phenotypic dissection is an important aspect to the manuscript. As part of the restructuring and reframing of the manuscript we have now provided additional details related to the heterogeneity of bipolar disorder and the subtypes.

The following text was added to the second paragraph of the introduction:

Updated text (page 2, Main):

"The heterogeneous nature of the disorder is noted in the Diagnostic and Statistical Manual of Mental Disorders, fifth edition (DSM-5), which includes the category "bipolar and related disorders," encompassing bipolar disorder type I (BDI), bipolar disorder type II (BDII) and cyclothymic disorders[47]. Similarly, the International Classification of Diseases, 11th Revision (ICD-11), also includes a section on bipolar disorders which recognises BDI and BDII as distinct subtypes[48]. BDI is characterised by episodes of mania, while BDII includes episodes of hypomania and depression."

2. Figure 2 presents perhaps the most interesting results of the paper, but the presentation is somewhat unintelligible. MiXeR estimates are unlikely to be totally obvious to the broad readership of Nature, and there is no description of what these estimates mean. In contrast to the unclear caption "The number and standard errors of influencing variants (???)…," the original MiXeR paper describes these estimates much more clearly: "The numbers indicate the estimated quantity of causal variants (in thousands) per component, explaining 90% of SNP heritability in each phenotype, followed by the standard error." The genetic correlations from LDSC shown below the Venn Diagrams are useful and interesting, but I have no idea what the

red bar is supposed to mean. It corresponds to neither the genetic correlation, nor polygenicity as far as I can gather, and is not described anywhere.

We agree with the reviewer's concern that the MiXeR estimates and accompanying figure legend were unclear. We have updated the figure legend as suggested by the reviewer. In addition we have removed the red scale bars that were reflective of the genetic correlations since this information is provided by text anyway.
Moreover, based on comment 12 by reviewer 2 we have also updated the figure legend to indicate that MiXeR model fit may not be reliable for all analyses.

Updated figure legend. New text is highlighted in yellow (page 4, Genetic architecture of BD differs by ascertainment and subtype):

> *"**Figure 1.** Genetic correlation and bivariate MiXeR estimates for the genetic overlap of BD ascertainment and subtypes. Trait-influencing genetic variants shared between each pair (grey) and unique to each trait (colours) are shown. The numbers within the Venn diagrams indicate the estimated number of trait-influencing variants (in thousands) per component, explaining 90% of SNP heritability in each phenotype, followed by the standard error. The size of the circles reflects the polygenicity of each trait, with larger circles corresponding to greater polygenicity. The estimated genetic correlation (rg) and standard error between BD and each trait of interest from LDSC is shown below the corresponding Venn diagram. Clinical and Community samples were stratified into bipolar I disorder (BDI) and bipolar II disorder (BDII) subtypes. Model fit statistics indicated that MiXeR-modeled overlap for bivariate comparisons including the bipolar subtypes (BDI and BDII) were not distinguishable from minimal or maximal possible overlap, and therefore to be interpreted with caution (see Supplementary Table 4)."*

3. At the bottom of pg 6, it is unclear why the authors chose to contrast directionality of BD with vs without self-report between MDD and SCZ. Table S11 indicates larger contrasts between ADHD and SCZ, which is a little visually easier to see when plotted (including error bars) and perhaps would make an interesting supplementary figure. The discussion mentions the absolute rankings rather than the relative differences, so MDD and SCZ may make sense to discuss there, but overall a more nuanced cross-phenotype discussion seems warranted.

We initially focused on SCZ and MDD as comparators since these disorders are often referred to when considering the range of the bipolar disorder spectrum, i.e. that bipolar disorder type I (BDI) is more similar to SCZ while bipolar disorder type II (BDII) is more similar to MDD. The reviewer is correct that ADHD does also result in large contrasts when considering the results with and without the self-report data. We have generated a new plot to show these differences in genetic correlation as suggested by the reviewer, and reviewer 2 comment 1 above. Further, we have investigated the BD spectrum heterogeneity (subtypes and ascertainment) to provide a better understanding of the differences in underlying genetic architecture. The figure is below

and Figure 2 in the revised main text. We have also expanded on these in the results and discussion.

Updated results text. New text highlighted in yellow (page 6, Genetic correlations with other traits):

> *"Genome-wide genetic correlations (r_g) were estimated between EUR ancestry BD GWASs (with and without self-reported data, and when stratified by ascertainment and subtypes) and human diseases and traits via the Complex Traits Genetics Virtual Lab (CTG-VL; https://vl.genoma.io) web platform[42] (Figure 2, Supplementary Tables 13-15). Most psychiatric disorders, including major depressive disorder (MDD), post-traumatic stress disorder (PTSD), attention deficit/hyperactivity disorder (ADHD), borderline personality disorder, and autism spectrum disorder (ASD), were more strongly correlated with the full meta-analysis, as well as with BDII, and BD in Community and Self-reported samples (Figure 2). In contrast, schizophrenia was more strongly genetically correlated with the full BD meta-analysis excluding self-reported data and with BDI and BD in clinical samples (Figure 2)."*

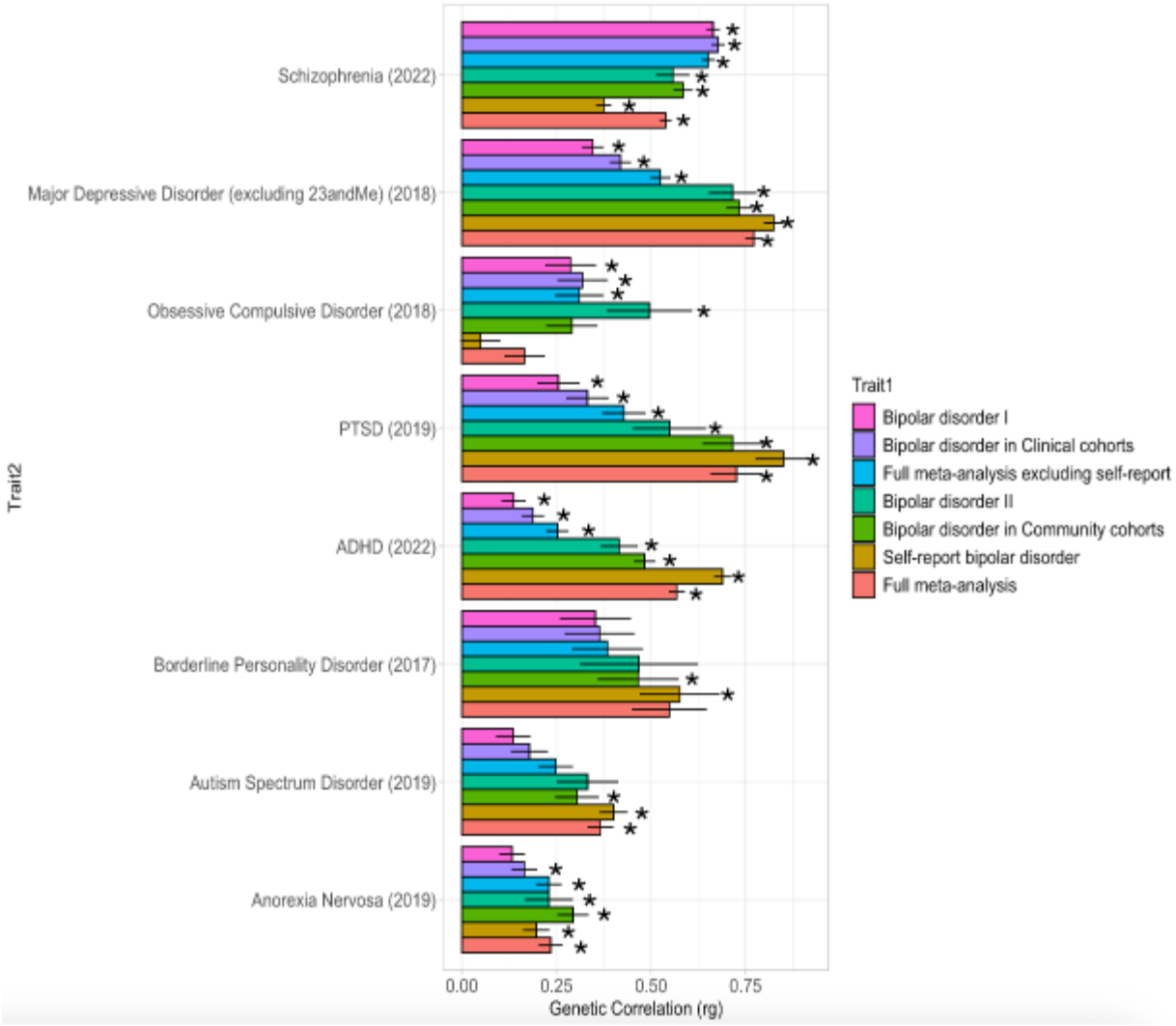


Figure 2. *Genetic correlations between bipolar disorder and other psychiatric disorders. The y-axis (Trait2) is ordered based on the significance and magnitude of genetic correlation of each trait with bipolar disorder type I. Stars indicate results passing theBonferroni corrected significance threshold of $P < 3.6 \times 10^{-5}$. ADHD, attention deficit/hyperactivity disorder. PTSD, post-traumatic stress disorder. Details are provided in Supplementary Table 13.*

Discussion text (page 14):

> *"Moreover, the pattern of correlations between BD and other psychiatric disorders differed with the inclusion of self-reported data. Schizophrenia had the highest correlation with BD without the inclusion of the self-reported data, while major depressive disorder was most strongly correlated with BD after the inclusion of the self-reported data."*

4. The PRS accuracy results are improved as expected. However, some of the language around interpretation is strange: "therefore the BD liability explained remains insufficient for diagnostic prediction in the general population." I don't think there is any area of medicine where people

use or think of PRS alone as diagnostic. Instead, they are predictive tools, used together with other clinical measures, and the prediction accuracy described here is fine.

> While we agree with the reviewer that medical experts are unlikely to think of PRS alone as diagnostic, we believe that a caution is still appropriate for the wider audience of readers since evidence suggests that understanding of PRS reports can be limited,[49] and polygenic embryo screening services are becoming more prevalent.[50,51] Therefore, we have opted to keep this text in the manuscript.

a. However, in Figure 3, it is a little odd that the all-cohort analysis with more phenotypic variation is not predicted as accurately compared to all subtypes analyzed. Is the self-report data pulling the PRS accuracy way down?

> The reviewer is correct that the PRS accuracy is diminished in the 'all-cohort' analysis when the discovery dataset includes self-reported samples. This pattern is similar to what is observed for the BDI and Clinical samples analyses, while inclusion of the self-reported samples improves the PRS in BDII and Community samples. These results are in line with our observations of lower genetic correlations between Self-reported BD and BDI (r_g = 0.42) or BD in Clinical samples (r_g = 0.47), since the majority of participants included in the 'all-cohort' analysis were from Clinical samples (27,833/40,992 BD cases).

The "Ascertainment" column in Table S20 is I think incorrect – w23andMe should be a mixed ascertainment, not "Clinical", correct?

> We acknowledge that there were inconsistent and sometimes confusing headers used in the supplementary tables. All supplementary tables have now been reviewed and the column headers clarified with explanations where necessary, to avoid further confusion or misunderstanding.

b. "When the multi-ancestry PRS was applied to a clinically ascertained AFR target cohort, assuming 2% disease prevalence, the inclusion of self-report data greatly increased the explained variance (no self-report R2 = 0.010, with self-report R2 = 0.023)." The majority of AFR participants have self-report data. Unless I missed something, I think a more accurate interpretation of this would be that when there are more ancestry-matched participants in the discovery GWAS, the PRS performs better (as opposed to emphasizing that AFR self-report vs other types of phenotype ascertainment are especially relevant given that the sample sizes are too small to draw conclusions about this). The discussion seems to convey this more accurately.

> We agree with the reviewer that when there are more ancestry-matched participants in the discovery GWAS, the PRS performs better.
> Interestingly, the observed increase in R2 in the AFR target cohort was independent of whether the PRS training data included only EUR or multi-ancestry data. Rather, the increase in sample size, gained by the inclusion of the self-report data, improves the PRS R2.

Given that the AFR target cohort is likely the most genetically heterogeneous of the target cohorts[39], it is possible that the similarity (in terms of genetic architecture) between the self-report AFR samples and the target AFR cohort is not as high as, for example, between the self-report EAS and target EAS cohorts. However, without access to the individual-level genetic data for the self-report AFR cohort this is difficult to investigate. Thus, we choose not to speculate too much on this in the discussion.

Updated discussion text. New text highlighted in yellow (page 14, Discussion):

> *"The multi-ancestry PRS provided the greatest improvement over the EUR-PRS in two of the three EAS target cohorts (Korean and Taiwanese). More subtle improvements were seen when the EUR target cohorts were analysed. Multi-ancestry training data provided little improvement in the AFR target cohort, which may be due to the genetic heterogeneity of this target cohort.[39] These results highlight the benefits of multi-ancestry representation in the PRS training data, in line with findings from other diseases[40]."*

5. Like other figures, the caption of Figure 4 assumes a level of familiarity of someone having directly made or discussed with the plot creator the interpretation of a sunburst plot rather than the broad readership of Nature. It is unclear what the hierarchy in this plot is indicating or what it is based on, aside from vaguely relating to pre- vs post-synapse activity. It simply says child terms are subsets of the adjacent inner ring. This description needs to be much clearer. Most regions of the sunburst plot are unannotated – what are they? Is this plot showing distance from synapse in some way? Further description of what specifically was done here is needed.

We agree with the reviewer that additional clarification and description was needed for this plot. However, with the restructure and revision of the results we have chosen to remove this figure from the manuscript. The text description and results in Supplementary Table 32 remain.

REFERENCES

1. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021).
2. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).
3. Yao, S. *et al.* Connecting genomic results for psychiatric disorders to human brain cell types and regions reveals convergence with functional connectivity. *medRxiv*

2024.01.18.24301478 (2024) doi:10.1101/2024.01.18.24301478.

4. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
5. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
6. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
7. Bhattacharya, A. *et al.* Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain. *Nat. Genet.* **55**, 2117–2128 (2023).
8. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
9. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
10. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103**, 217–234.e4 (2019).
11. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
12. Jonsson, L., Song, J., Joas, E., Pålsson, E. & Landén, M. Polygenic scores for psychiatric disorders associate with year of first bipolar disorder diagnosis: A register-based study between 1972 and 2016. *Psychiatry Res.* **339**, 116081 (2024).
13. Zimmerman, M., Ruggero, C. J., Chelminski, I. & Young, D. Is bipolar disorder overdiagnosed? *J. Clin. Psychiatry* **69**, 935–940 (2008).
14. Zimmerman, M., Ruggero, C. J., Chelminski, I. & Young, D. Psychiatric diagnoses in patients previously overdiagnosed with bipolar disorder. *J. Clin. Psychiatry* **71**, 26–31 (2010).

15. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
16. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.* **52**, 437–447 (2020).
17. Legge, S. E. *et al.* Genetic and Phenotypic Features of Schizophrenia in the UK Biobank. *JAMA Psychiatry* (2024) doi:10.1001/jamapsychiatry.2024.0200.
18. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet.* **16**, e1008612 (2020).
19. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat. Genet.* **55**, 198–208 (2023).
20. Frei, O. *et al.* Improved functional mapping of complex trait heritability with GSA-MiXeR implicates biologically specific gene sets. *Nat. Genet.* 1–9 (2024).
21. Alda, M. Lithium in the treatment of bipolar disorder: pharmacology and pharmacogenetics. *Mol. Psychiatry* **20**, 661–670 (2015).
22. Toker, L., Belmaker, R. H. & Agam, G. Gene-expression studies in understanding the mechanism of action of lithium. *Expert Rev. Neurother.* **12**, 93–97 (2012).
23. Ching, C. R. K. *et al.* What we learn about bipolar disorder from large-scale neuroimaging: Findings and future directions from the ENIGMA Bipolar Disorder Working Group. *Hum. Brain Mapp.* **43**, 56–82 (2022).
24. Cannon, M. *et al.* DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms. *Nucleic Acids Res.* **52**, D1227–D1235 (2024).
25. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509–516 (2022).
26. Olfson, E. *et al.* Ultra-rare de novo damaging coding variants are enriched in attention-deficit/hyperactivity disorder and identify risk genes. *medRxiv* 2023.05.19.23290241 (2023) doi:10.1101/2023.05.19.23290241.

27. Palmer, D. S. *et al.* Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. *Nat. Genet.* **54**, 541–547 (2022).
28. Morselli Gysi, D. & Barabási, A.-L. Noncoding RNAs improve the predictive power of network medicine. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2301342120 (2023).
29. Menche, J. *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
30. Reigstad, C. S. *et al.* Gut microbes promote colonic serotonin production through an effect of short-chain fatty acids on enterochromaffin cells. *The FASEB Journal* **29**, 1395–1403 (2015).
31. Ortega, M. A. *et al.* Microbiota–gut–brain axis mechanisms in the complex network of bipolar disorders: potential clinical implications and translational opportunities. *Mol. Psychiatry* **28**, 2645–2673 (2023).
32. Lithium carbonate alleviates colon inflammation through modulating gut microbiota and Treg cells in a GPR43-dependent manner. *Pharmacol. Res.* **175**, 105992 (2022).
33. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).
34. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
35. de la Fuente, J., Davies, G., Grotzinger, A. D., Tucker-Drob, E. M. & Deary, I. J. A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. *Nat Hum Behav* **5**, 49–58 (2021).
36. Li, X. & Zhao, H. Automated feature extraction from population wearable device data identified novel loci associated with sleep and circadian rhythms. *PLoS Genet.* **16**, e1009089 (2020).
37. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).

38. Thalamuthu, A. *et al.* Genome-wide interaction study with major depression identifies novel variants associated with cognitive function. *Mol. Psychiatry* **27**, 1111–1119 (2022).
39. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).
40. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
41. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
42. Cuéllar-Partida, G. *et al.* Complex-Traits Genetics Virtual Lab: A community-driven web platform for post-GWAS analyses. *bioRxiv* 518027 (2019) doi:10.1101/518027.
43. Hübel, C. *et al.* Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **180**, 428–438 (2019).
44. Neurodevelopmental pathways in bipolar disorder. *Neurosci. Biobehav. Rev.* **112**, 213–226 (2020).
45. Martin, J. *et al.* Examining Sex-Differentiated Genetic Effects Across Neuropsychiatric and Behavioral Traits. *Biol. Psychiatry* **89**, 1127–1137 (2021).
46. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
47. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Pub, 2013).
48. Reed, G. M. *et al.* Innovations and changes in the ICD-11 classification of mental, behavioural and neurodevelopmental disorders. *World Psychiatry* **18**, 3–19 (2019).
49. Peck, L., Borle, K., Folkersen, L. & Austin, J. Why do people seek out polygenic risk scores for complex disorders, and how do they understand and react to results? *Eur. J. Hum. Genet.* **30**, 81–87 (2022).

50. Lencz, T. *et al.* Utility of polygenic embryo screening for disease depends on the selection strategy. *Elife* **10**, (2021).
51. Lencz, T. *et al.* Concerns about the use of polygenic embryo screening for psychiatric and cognitive traits. *Lancet Psychiatry* **9**, 838–844 (2022).

We again thank all reviewers for their thorough review of the manuscript. Below we provide responses (in blue) to the remaining specific comments from each reviewer.

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have satisfactorily addressed my comments and implemented my suggestions. I thank the authors for doing so. The manuscript now reads better than the earlier version. But still I feel it could be improved further to suit for Nature's broad readership. Some suggestions include:
- Avoid unnecessary acronyms such as GWS.

We have replaced all unnecessary acronyms as the reviewer suggests.

- Avoid repeating the method names, for example, GSA-MiXeR, every time when describing the results. After the first time, directly describe the findings and if curious, the reader could refer to methods to find what statistical method or tool was used.

We have now revised the manuscript and removed repetition of method names when describing the results in both the 'results' and 'discussion' sections.

- In the end of each section, add a concluding statement(s) summarizing what was the take home from that part of the analysis. Just by reading the first and last few lines of each section, the reader should be able to grab the main messages that the authors were trying to make.

We have now updated the manuscript to include summarizing statements at the end of each section, when relevant. Listed below are the sections and new summarizing statement texts:

Genetic architecture of BD (page 13):

"We show that genetic architecture is different across ascertainment and subtypes, and that these differences appear to be driven by the proportion of BD subtype within the sample. Despite these observed differences, the high correlations of variant effects in the shared components across ascertainment groups supports our decision to meta-analyse all BD cases."

Multi-ancestry meta-analysis (page 14)

"Our multi-ancestry meta-analysis identified 298 loci implicating 337 LD independent genome-wide significant variants."

Genetic correlations with other traits (page 15)"

“This pattern of correlations, together with the observed patterns of genetic architecture, suggest that the Self-reported samples include a high proportion of people with BDII.”

Polygenic association with BD (page 15). New text is highlighted in yellow:

“Similarly, PRS derived from GWAS excluding self-reported data explained significantly more variance in cases of BDI (Figure 3B, Supplementary Tables 17) and in Clinical cohorts (Figure 3D, Supplementary Tables 19) than when self-reported data were included. Conversely, inclusion of the self-reported data yielded greater median R2 estimates for the PRS in cases of BDII (Figure 3C, Supplementary Tables 18) and in Community cohorts (Figure 3E, Supplementary Tables 20), although these differences were not significant. These results are likely due to increased phenotypic heterogeneity when the self-reported data are included in the PRS discovery sample (see Figure 2).”

Pathway, tissue and cell type enrichment (page 16):

“Together these results implicate the synapse, interneurons of the prefrontal cortex and hippocampus, and hippocampal pyramidal neurons as particularly relevant in molecular biology of BD.”

Credible BD-associated genes (page 18)

“Together, these results implicate 36 credible genes in BD.”

Drug target analyses (page 19):

“Since the credible gene list is primarily derived from our fine-mapping analysis, it is possible that lithium target genes (and interaction partners) are within loci for which significant fine-mapped putatively causal SNPs were not identified. The identification of evidence of tractability with small molecules for some of the credible genes indicates opportunities for novel drug development.”

- Please use color-blind friendly palettes for the figures.

We have updated the figures to use color-blind friendly palettes.

- I'd advise against using the phrase "prediction" in the polygenic score analysis. Instead, just use the word "association".

We have now replaced the term ‘prediction’ with ‘association’ for all polygenic score analyses and results discussion.

Specific comments:
- The authors prioritize two lists of causal genes. Initially, they prioritize 71 genes using fine-mapping, which they use for the common-rare variants convergence analysis. Later, they describe further analyses in which they prioritize a shorter list of

36 genes, which they use for lifespan gene expression and drug target enrichment analysis. The abstract also reports the 36 genes. This is a bit confusing. I'd advise sticking to one list, perhaps just the 36 genes and do all downstream analysis based on that list. I'd guess the borderline significance of rare variant enrichment would disappear when using 36 genes. If so, I'd advise reporting as it is and not claim convergence of common and rare variants.

> The reviewer is correct that the final list of 36 credible genes is under-powered to idenitify a signficant enrichment of rare variants, and this is why we used the initial list of 71 genes annotated to fine-mapped SNPs. Our procedure to identify credible genes for BD is initially based on the significant fine-mapped SNPs, and as such we believe that performing this analysis with genes annotated to these SNPs has some reliability. Moreover, the majority (n=29) of the credible genes are included in the list of annotated genes from fine-mapped SNPs. Thus, although the genes in this annotated list (n=71) are not confirmed by numerous in silico evidences, they are based on the approach which we deemed to have the highest level of evidence. We therefore believe that these results are still valid and important to convey.

- In the fine mapping analysis, how many genes implicated by coding variants? If there are any, mentioning that will be informative.

> None of the signficant fine-mapped variants are coding variants. We have updated Supplementary Tables 28 and 29 to include the 'functional' consequences of all fine-mapped variants.

Referee #2 (Remarks to the Author):

It was a pleasure to read the revised manuscript. The authors have done an excellent and thoughtful revision that has significantly enhanced the quality of the manuscript. The presentation is clear and engaging, making the paper an interesting read. I particularly appreciate the additional analyses, which provide valuable biological insights, such as the evaluation of identified BD risk genes across brain developmental stages and the SNP-heritability enrichment analyses of cell types. The reordering of the manuscript, where heterogeneity among BD subtypes or ascertainment is presented before the GWAS, is also a notable improvement.

I only have a few minor comments:

1. Figure 2:
- I appreciate the new Figure 2, which is very informative. For clarity, I will suggest to edit the text on page 6 to the following: "Most psychiatric disorders, including major depressive disorder (MDD), post-traumatic stress disorder (PTSD), attention deficit/hyperactivity disorder (ADHD), borderline personality disorder, and autism spectrum disorder (ASD), were more strongly correlated with the full meta-analysis, BDII, and BD in Community and Self-reported samples, than with BDI and BD in clinical cohorts (Figure 2)".

> We have updated the text with the reviewers recommendation.

The text on page 6 is now as follows. The reviewers suggestion is highlighted in yellow:

"Genome-wide genetic correlations (rg) were estimated between EUR ancestry BD GWASs (with and without self-reported data, and when stratified by ascertainment and subtypes) and human diseases and traits via the Complex Traits Genetics Virtual Lab (CTG-VL; https://vl.genoma.io) web platform20 (Figure 2, Supplementary Tables 13-15). Most psychiatric disorders, including major depressive disorder (MDD), post-traumatic stress disorder (PTSD), attention deficit/hyperactivity disorder (ADHD), borderline personality disorder, and autism spectrum disorder (ASD), were more strongly correlated with the full meta-analysis, BDII, and BD in Community and Self-reported samples, than with BDI and BD in clinical cohorts (Figure 2). In contrast, schizophrenia was more strongly genetically correlated with the full BD meta-analysis excluding self-reported data and with BDI and BD in clinical samples (Figure 2)."

- The figure itself needs some polishing. I suggest demonstrating the genetic correlations as dots (± standard error) and removing the gridlines. Additionally, information about the horizontal lines should be added to the figure legend. I assume they represent standard errors.

We have now ammended Figure 2 based on the reviewer's suggestion and the requirements for figures as outlined by the journal. We have updated the figure legend accordingly. The new figure legend reads as follows, with new text highlighted in yellow:

"Figure 2. Genetic correlations (with standard errors) between bipolar disorder and other psychiatric disorders. The y-axis (Trait2) is ordered based on the significance and magnitude of genetic correlation of each trait with bipolar disorder type I. P-values were calculated from the two-sided z statistics computed by dividing the estimated genetic correlation by the estimated standard error, without adjustment. The standard error for a genetic correlation was estimated using a ratio block jackknife over 200 blocks. Triangles indicate significant results passing the Bonferroni corrected significance threshold of two-sided $P < 3.6 \times 10{-}5$. Each error bar represents the standard error of the estimate. ADHD, attention deficit/hyperactivity disorder. PTSD, post-traumatic stress disorder. The year indicated in parentheses after each trait refers to the year in which the GWAS was published. Details are provided in Supplementary Table 13."

- The ADHD GWAS was published in 2023, so this should be corrected in the figure and in the supplementary table.

This has now been corrected in the new figure as well as in the supplementary table.

2. Figure 4:

- I think this figure adds great value to the manuscript. Could the authors please include information in the figure legend about the small bar in the middle — specifically, what do the numbers 1-6 represent?

We have updated the figure to include the label 'Enrichment Z-score' for the legend (numbers 1-6). This was already mentioned in the figure legend so it should now be more clear.
Figure legend:

"Figure 4. Supercluster-level SNP-heritability enrichment for bipolar disorder. The t-distributed stochastic neighbour embedding (tSNE) plot (from Siletti et al.23) (left) is coloured by the enrichment z-score. Grey indicates non-significantly enriched superclusters (FDR > 0.05). The barplot (right) shows the nine significantly enriched superclusters."

3. Supplementary Information:
- Related to my previous comment regarding the explanation of headers in the Supplementary Information. I appreciate the authors clarification of the headers of the MiXeR results in their rebuttal letter. However, I could not find any details in the Methods section about the number of runs performed in the trivariate MiXeR, which makes it difficult for the reader to understand what "Run" means. Please add more information about this to the Methods section and the header of Supplementary Table 4.

We have now updated the methods text for MiXeR to better explain what 'run' means. The text now reads as follows, with new text highlighted in yellow:

"We first computed univariate analyses to estimate the polygenicity, discoverability and heritability of each trait. These were followed by bivariate analyses to compute the number of shared trait-influencing variants between pairs of traits, and finally trivariate analyses to compute the proportion of shared variants between all three traits analysed. We also determined the correlation of effect sizes of SNPs within the bivariate shared components. For trivariate MiXeR analyses, model optimization procedures are repeated 20 times (20 runs) to obtain the means and standard errors of model parameters. Estimated parameters from the 'run' with the smallest deviation from the median overlap pattern are then selected and reported."

Moreover, we also include an additional explanation in Supplementary Table 4 as follows:

"run_id = iteration (1-20) of model optimization procedures to obtain means and standard errors of model parameters."

4. Rare Variant Analyses:
- The authors have performed a Fisher's exact test, which does not correct for covariates, and I understand their rationale for using this method. However, I think it is important to evaluate whether the signal is not due to sequencing or batch errors. Could the authors include results from an analysis of rare synonymous variants in

the BD risk gene set, as we would not expect differences between cases and controls for this category of variants?

> While we agree that this could be an interesting analysis, the synonymous counts are not readily available to perform it. However, the initial BipEx analysis included extensive quality control of the data, including elimination of batch effects. Approximately 1.9% of the initial sample was excluded during quality control based on identified batch effects.[1] Similar extensive batch quality control was performed in the SCHEMA study.[2]

5. Clustering of genes based on expression patterns across brain developmental stages:
- Page 22: The authors write: “Outliers in gene expression more than 4 standard deviations from the mean were removed.” What was the rationale for removing outlier genes if QC of the expression data is assumed to be properly done? The genes removed might not be outliers if the tested set had included other or more genes.
And related, why was 4 standard deviations chosen?

> For clarification, we removed an individual’s gene expression value if it was an outlier (>4 standard deviations) for a given gene in a given brain region. This did not remove genes themselves from the analyses. Instead, this prevented any extreme expression values from skewing the analyses. Less thant 0.0008% of the expression values were removed as outliers. Including these outliers in expression may have affected modeling of age-related variation and added additional noise which would ultimately effect clustering. While the threshold for defining outliers can vary across studies, we used the same cut-off as in our previous publication[3] to be in-line with the standards in the field.

- Include a reference or URL to the TmixClust R-package.

> We have now udpated the manuscript text to include URLs for all R-packages used for analyses, including the fmsb (https://cran.r-project.org/web/packages/fmsb/index.html), pROC (https://cran.r-project.org/web/packages/pROC/index.html), limma (https://www.bioconductor.org/packages/release/bioc/html/limma.html) and TMixClust packages (https://www.bioconductor.org/packages/release/bioc/html/TMixClust.html), respectively.

- Page 22: the authors write: “We compare the average silhouette width across the K=2 to K=10 clusters and select that with the maximum value as the optimal number of clusters”. Since high silhouette values (i.e., close to one) indicate a good support for the clustering, it is important to know the estimated silhouette value for K=2 clusters – please add that in the manuscript.

> We have updated the text in the methods to now include this information. New text is highlighted in yellow (Temporal clustering of credible genes, page 38):

"This method uses mixed-effects models with nonparametric smoothing splines to capture and cluster non-linear variation in temporal gene expression. We tested K=2 to K=10 clusters performing 50 clustering runs to analyse stability. The clustering solution with the highest likelihood (i.e., the global optimum using an expectation maximization technique) is selected as the most stable solution across the 50 runs for each of the trials testing 2-10 clusters. We compare the average silhouette width across the K=2 to K=10 clusters and select that with the maximum value as the optimal number of clusters. The highest average silhouette width was 0.24 for two clusters, while the lowest was 0.17 for four clusters."

- Supplementary Figure 10. There is no explanation of the grey shaded area around the lines, but I don't think this is confidence intervals related to expression because I would expect these to be larger. Please add confidence intervals for the average expression of the genes in the two clusters at each brain developmental stage to the figure.

To clarify, in this extended figure, we are plotting the smoothed age trajectory for expression using a generalized additive model (GAM) for each cluster. The 95% confidence interval for each cluster's trajectory is based on predictions from this GAM. We agree that information on the confidence intervals for this plot are necessary and that additional clarity on the variation in expression across individual donors and brain regions is important. For this reason, we have now added an additional statement to the figure legend regarding the confidence intervals. In addition, we have added individual points for the donor expression values across brain regions for all 34 credible genes to the figure. This will allow readers to understand the variability in expression within the clusters.

The figure legend has been updated with new text highlighted in yellow:

"Extended Figure 8. Clustering of patterns of temporal variation in expression of 34 credible genes.
Cluster 1 (n=21 genes) shows reduced prenatal gene expression, with gene expression peaking at birth and remaining stable over the life-course. Cluster 2 (n=13 genes) includes genes with a peak gene expression during fetal development with a drop-off in expression before birth. Genes within each cluster are described in Supplementary Table 31. To illustrate the variability in gene expression within each cluster we plot each donor expression value in each sampled brain region for the 34 credible genes as individual points. Smoothing splines used to illustrate the age trajectory for each cluster is based on generalized additive models with the predicted 95% confidence interval in grey. We use age in days to plot the variation in gene expression with the x-axis on a log2 scale and labels for birth, 10, 18, and 65 years of age as reference points."

- Additionally, the scale on the x-axis is in years but not linear in the figure, please add a brief explanation to the S10 figure legend explaining the scale.

The x-axis of figure S10 is on a log2 scale. We have now added this informatin to the figure legend.
We have added the following text to the figure legend to clarify this:

"We use age in days to plot the variation in gene expression with the x-axis on a log2 scale and labels for birth, 10, 18, and 65 years of age as reference points."

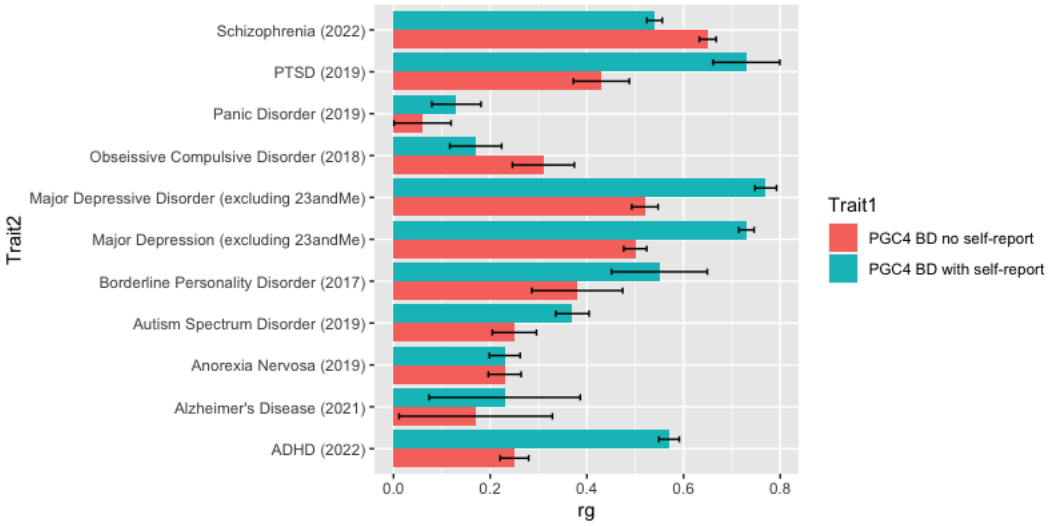
Referee #3 (Remarks to the Author):

The authors have been responsive to most reviewer comments, including all of mine. Reviewer 1's major request for an honest discuss about new *biological* insights gleaned from increasing sample sizes here could be strengthened with a potential roadmap and discussion around how future bipolar genetic studies to deepen insights. This was partially addressed but could perhaps be strengthened in collaboration with the editorial team. Other efforts to strengthen and reframe the paper have been responsive.

We have followed the input from the editor and the comments from the Reviewers to further, strengthen and reframe the paper.

References

1 Palmer, D. S. et al. Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. Nat Genet 54, 541-547 (2022). https://doi.org:10.1038/s41588-022-01034-x
2 Singh, T. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature 604, 509-516 (2022). https://doi.org:10.1038/s41586-022-04556-w
3 Parker, N. et al. Psychiatric disorders and brain white matter exhibit genetic overlap implicating developmental and neural cell biology. Mol Psychiatry 28, 4924-4932 (2023). https://doi.org:10.1038/s41380-023-02264-z


Schizophrenia (2022)
PTSD (2019)
Panic Disorder (2019)
Obseissive Compulsive Disorder (2018)
Major Depressive Disorder (excluding 23andMe)
Major Depression (excluding 23andMe)
Borderline Personality Disorder (2017)
Autism Spectrum Disorder (2019)
Anorexia Nervosa (2019)
Alzheimer's Disease (2021)
ADHD (2022)
Trait2
0.0
0.2
0.4
0.6
0.8
rg
Trait1
PGC4 BD no self-report
PGC4 BD with self-report

Reviewer #2:

I would like to thank the authors for addressing many of my previous questions. However, I have a few additional items that I would like to see further addressed:

1. Although I am not an expert in clustering methods, I have gathered from the literature that a silhouette width of 0.5 is generally indicative of meaningful clustering, while a value below 0.25 is considered weak. If my understanding is accurate, it seems there may not be sufficient support for identifying two distinct gene clusters, as mentioned on page 12. Therefore, the conclusion regarding the identification of two gene clusters seems overly strong.

   It is true that the average silhouette width of 0.24 is only suggestive of a two-cluster solution for credible genes. Therefore, we have now adjusted the text in several locations to address this point.

   Main Text (page 18):
   "Based on the lifespan gene expression data from the Human Brain Transcriptome project (www.hbatlas.org),[35] ~~we identified~~ *suggestive evidence for* two clusters of credible ~~genes with distinct peaks in~~ *was observed based on* temporal expression… *However, both clusters show high variability in gene expression across the lifespan.*"

   Discussion (page20):
   "Moreover, we clustered the credible genes based on similar patterns of temporal variation in expression over the lifespan and ~~identified~~ *found suggestive evidence for* two clusters. ~~The first cluster shows reduced gene expression pre-birth, with gene expression peaking at birth and remaining stable over the life-course, while~~ *Although within cluster gene expression was highly variable across the lifespan*, the second cluster had a peak in expression during fetal development aligning with the neurodevelopmental hypothesis of mental disorders."

   Methods (page 35)
   "*Overall, there was suggestive evidence for a two-cluster solution for the temporal expression of credible genes*."

2. In Supplementary 8 figure legend, the authors write: "To illustrate the variability in gene expression within each cluster, we plot each donor's expression value in each sampled brain region for the 34 credible genes as individual points." However, I noticed that only 33 genes are listed as belonging to a cluster in the supplementary table from the previous revised version. Please clarify whether the correct number is 33 or 34.

   We thank the reviewer for catching this typo in the Methods section. There was in fact 34 credible genes with temporal gene expression data. We have now made the correction to the text.

3. Concerning the analyses of rare variants, I am puzzled as to why access to counts of synonymous variants for bipolar risk genes was not possible. Both the schema browser and the BipEX browser provide counts for synonymous variants. I believe it is crucial to include analyses of synonymous variants, and I recommend strongly this to be incorporated.

We have now performed analysis of synonymous variants, as suggested by the reviewer, in the BipEx cohort. The 71 genes annotated to putatively causal fine-mapped SNPs were not enriched for synonymous variants in BD cases (odds ratio = 0.96, 95% confidence interval 0.935-0.985).

Methods (page 33)
"As a sensitivity analysis we further evaluated the enrichment of synonymous variants in the credible genes in BD cases of the BipEx cohort and found no enrichment (odds ratio = 0.96, 95% confidence interval 0.935-0.985)."