nature portfolio

Peer Review File

Robust inference and correlates from genetic associations with personality

Corresponding Author: Dr Ted Schwaba

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
I must preface this review with confessions of my lack of expertise. I'm not a twin or molecular genetics expert. I am, however, an avid consumer of this field's work with more than mild interest in the meaning and import of the findings. With my limitations in mind, I have not examined the methods or statistics closely and will focus mostly on the significance of the findings, and how the paper is framed and discussed.

Globally, this is a very important paper for the field of personality genetics and the field of personality more broadly. It is as definitive as we can get with the technologies at hand to ask basic and important questions about genetic signatures found in personality phenotypes.

I found the write up to be enthusiastic, maybe a bit too enthusiastic. Statements like "personality genetics have widespread, potentially causal associations with a wide range of consequential behaviors and life outcomes" while true in some senses is probably too enthusiastic given the caveats that need to accompany those statements which neither the abstract, nor the brief nature of a Nature article format afford. Given the history of how genetics research has been misused by problematic parties (mass murderers, for example), I think it would be prudent if the authors contemplated the fact that their work will reach less sophisticated audiences than the typical Nature reader. Temperance and humility might be a better approach than enthusiasm.

Similarly, after a compelling section outlining the importance of personality to human functioning the authors argue that "Given the fundamental relevance of personality to core phenomena studied across scientific disciplines, understanding the molecular genetics of personality is crucially important for developing biopsychosocial models of both human behavior and life outcomes." Is it? How? We know things are genetic. We know the GWAS data are challenging. We know there are no simple "gene for" associations. As an avid consumer of these types of studies, I already knew going in that the big issue is the rather large discrepancy between twin estimates of heritability and GWAS estimates with the latter being disturbingly small, comparitavely speaking. This context does not appear to inform the enthusiasm for the effort. Nor is the discrepancy discussed enough for my taste. Couple this with the way genetics studies and heritability findings have been used by authors in the past, and one would be concerned with the lack of perspective taking. Some authors took the most extreme twin estimates to argue that there is no nurture in personality's nature. Others gleefully cite the low GWAS heritabilities of personality as evidence for the Blank Slate. While it might be a mistake to infer the opposite with the small heritabilities characteristic of GWAS efforts, it is hard not to take the findings as an indication that genetics are far less important than originally thought. There are many complicated reasons not to draw this conclusion and most of those complexities are not registered here. The framing comes across as "look, its genetic!", which, after all is not a very interesting research question at this point in history. Somehow, some of the complexity of the findings needs to be infused into the write up despite page limitations.

To that end, I believe the authors have a responsibility to more consistently and directly handle the effect size issue. The fact that the variance accounted for is quite small and much smaller than twin studie should be a theme throughout.

Also, I understand the authors impulse to correct their results for unreliability, especially in the context of a meta-analysis. There are always studies that have glaring issues and it seems wrong to simply merge those estimates together with better conducted studies. That said, believing in something fantastical like perfect measurement reliability is equally problematic. To date, we have never achieved it and we are unlikely to to so anytime soon. I propose a compromise. Adjust the results for the highest achievable average reliability estimates we find in the field–.80 or so. That way, you can report more realistic estimates (e.g., when the study is done well) rather than in some hypothetical universe with unicorns and perfect reliability.

I like very much the material on what the PGS predict, though I would have prefered some discussion of mechanism here. Are we to infer that if I get some good genes, these genes will act directly on whether another human sees me as well-groomed (as we currently define it)? Or, should we assume there are intervening factors that act on these genetic propensities to manifest the phenotype of "well-groomed"? If it is the latter, a more detailed discussion of this process as well as some descriptive hedging of the PGS predictions wouild be good. At this juncture, given the way it is written, some of your readers can conclude that being born with a certain suite of genes causes, directly, you being a criminal. Is that really what the authors are proposing?

Again, it would be preferable to condition the predictive relationships on the magnitude of the effects rather than statistical signficance. The current write up uses the latter and leaves the mistaken impression that everything is strongly related to everything else. Relatedly, describing the findings on Figure 4b as "robust" is more marketing than science. Robust is meaningless here unless the authors have given the reader a metric of effect sizes to understand the magnitude of the effects.

The most profound findings, in my opinion, are the null findings. There was no strong signal in regions related to dopaminergic and serotonergic systems which have been the pillars of personality neuroscience if not psychological neuroscience. There was no strong difference between self-reports and observer ratings of personality. The PGS worked equally well across ancestries. There is little or no assortative mating on personality. There is little or no environmental confounding. These are quite possibly some of the most significant null effects I've read, as the alternative findings or threatened findings have been so significant to the field. The dopamine and serotonic hypotheses have ruled physiological personaltiy psychology for decades. Self-reports are commonly dismissed as invalid because of reporting biases. Lack of generalizability of the genetic findings is commonly used to argue against genetic findings. Assortative mating is often invoked as another problematic confound for genetic as well as phenotypic research on personality. This paper will reset the literatures of multiple fields.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
The study reports findings from large-scale GWAS on the Big-5 personality traits. This was a big gap in the field – there were no published well-powered personality GWAS until very recently(Gupta et al., 2024). The current study, by also incorporating data from the latest GWAS (Gupta et al., 2024), conducts the largest personality GWAS to date and obtains important insights about the genetics of personality, as well as facilitating future studies on a wide range of research questions involving personality that could not be conducted before due to the low statistical power of the GWAS. As such, this is a very important contribution. There is a lot to like a about this paper. It is very comprehensive, carefully conducted and was a pleasure to read. However, I believe there are some points that could be improved which I list below. I realize this is a very lengthy list of comments but a lot of them are just clarification questions that I hope will be quick to address.

QC and meta-analysis
1. It was unclear to me what QC filters have been applied at the cohort-level prior to meta-analysis other than dropping multiallelic SNPs or those that are not in 1000 Genomes. There is no mention of a MAF filter, and only an INFO filter of 0.4 appears to have been applied prior to GWAS by cohorts. Is that all?

2. It would be good to see the INFO scores for the lead SNPs in Supplementary Tables 5-9.

3. I was surprised to see that the authors used a 250kb window when clumping the genome-wide significant SNPs. That is a window much smaller than is used in the literature and is likely to identify SNPs in long-range LD as independent hits. As the authors also report, we know that at least for neuroticism, some of the associated SNPs lie on inversion regions. 250kb distance will not be enough to cover these regions, leading to overreporting of approximately independent SNPs. Do the authors have a justification for using such a small window?

4. Related to my point above, in order to identify novel SNPs, the authors check whether any of the genome-wide significant SNPs found in previous research are within the 250kb loci that they've identified. This, also, is a very anti-conservative definition of a novel hit. It would be good to see how many SNPs identified in the current study have a squared correlation less than 0.1 with all of the previously identified ones on the same chromosome.

5. It would be interesting to look at what the dosage compensation parameter actually is, whether the heritability due to the X

chromosome differs between males and females, and what the X chromosome male-female genetic correlation is. Some of these analyses might require individual-level genetic data but it would be possible to estimate them in UKB at least for neuroticism.

6. The authors apply inverse-variance-weighted meta-analysis to combine the EUR and AFR GWAS. Why not use something more sophisticated designed for this purpose, e.g. MAMA(Turley et al., 2021)?

7. In Section 2.7 of the SI, the authors say "we aligned SNPs to the 1000 genomes 3v5 EUR reference panel". I am not sure what "aligning" means here. Do they mean the summary stats were clumped using EUR data? If so, why not use the combined EUR and AFR reference data that they've generated to identify the novel SNPs?

8. It would be nice to see the ancestry-specific MAF and INFO scores for the SNPs in SI Table 10.

Environmental confounding
9. The LDSC intercept for the neuroticism meta-analysis is reported to be less than 1. Similarly, many of the cohort-level LDSC intercepts in Supplementary Tables 1 and 2 appear to be below 1. This often happens because the effective sample size is lower than the reported sample size (e.g. when the GWAS is run with mixed linear models). The authors say that they have re-calculated effective sample size of the meta-analyses because of this. A few questions and suggestions about this:
a. Was this not done at the cohort level before running LDSC? Why are so many intercepts still below 1?
b. Why is the neuroticism intercept still below 1 even though it was estimated with effective N?
c. It would be more informative to see the LDSC ratios rather than the intercepts in all cases to understand how the confounding compares to signal.

10. Why have the authors not used the Tan et al.(Tan et al., 2024) within-family neuroticism GWAS instead of Howe et al. given the worries about the low INFO score filter applied? Isn't the Tan et al. GWAS better powered as well?

11. SI Section 6.4 states the QC procedure mirrored the full-population analysis except for the stated modifications. Does that mean no MAF filter was applied here, either?

12. The authors find a within-family GWAS LDSC intercept significantly higher than 1, and also not lower than the population GWAS LDSC intercept. This is surprising and requires some discussion. It would also be useful to see the ratio statistics rather than intercepts here. Do the LD scores not match the GWAS, or is it the genetic architecture of these traits that don't match LDSC assumptions?

13. Were the within-family - population-level genetic correlations estimated by LDSC so that population stratification is accounted for? Since the authors claim there is negligible confounding in the GWAS, how do they interpret the correlations for agreeableness and neuroticism being significantly less than 1?

14. It would be interesting to see the genetic correlations of the within-family GWAS with the traits in Figure 3, especially for neuroticism and agreeableness.

15. I think some discussion on the difference between the within-twin polygenic prediction and direct/indirect genetic effects analyses is missing. As it stands, it's unclear why these two different analyses are conducted because they both aim to get at the same result. The way they are presented as if they are complimentary is misleading. The authors do not mention that the PGI effects obtained using the within-twin analysis may be biased by indirect genetic effects from siblings. Imputing the genotypes of the parents and controlling for parental PGIs is a way to deal with this. I am not sure why the within-twin prediction has been done at all. It would be better to run the analysis controlling for parental PGIs in two cohorts and meta-analyze the results. If the authors do have a reason why they think these two analyses are complimentary and we get something out of the within-twin prediction that we don't get out of controlling for parental PGIs, they should make it clearer in the text.

16. The terminology in this section seems somewhat imprecise. The authors refer to the coefficient of the non-transmitted parental PGIs sometimes as "indirect effects" and sometimes "non-transmitted genetic effects". Given that this coefficient captures biases due to population stratification and assortative mating in addition to genetic effects from parents, calling it an "effect" is a misnomer. I realize this terminology has been used before in other papers but it becomes especially misleading when it gets used interchangeably with "genetic nurture" later on in this section (also in the Figure 4B title) when the finding that the non-transmitted coefficient is non-significant is interpreted as there is negligible genetic nurture. It is also not clear to me that this interpretation follows from the finding. Although not too likely, it is possible that genetic nurture operates in the reverse direction (e.g. growing up with neurotic parents makes the child less neurotic as a reaction?) and it cancels out the inflation in predictive power due to population stratification such that there appears to be no confounding. The fact that heritability estimated using the within-family GWAS appears to be close to the one estimated using the population GWAS is evidence against this to the extent that LDSC accounts for population stratification because then the difference is attributable to genetic nurture, but the authors need to synthesize these findings better in their discussion.

Replication
17. I'm having a hard time understanding why the replication was done with overlapping samples. It seems straightforward to exclude the Gupta et al.(Gupta et al., 2024) samples from the meta-analysis and try to replicate their findings in independent data. If that is not possible for whatever reason, it's possible to analytically subtract the Gupta et al. GWAS from the meta-analysis. Am I missing something? As it stands, this cannot be called a replication analysis and is not meaningful at all.

Heritability differences across groups/ancestries

18. The analyses looking at heterogeneity in heritability estimates across different groups of samples is interesting, but I feel like the results are somewhat glossed over in the main text. The authors discuss only mean genetic correlations across the big-5 between groups, which do not reflect the interesting differences at the trait level. For example, the genetic correlation between the old and young for neuroticism is 0.33, which is quite striking. Moreover, even if we just look at mean rg's, the conclusion that I would draw from a mean rg of 0.8 across age groups would not be that genetic effects are highly similar. An rg of 0.8 for the same trait indicates quite some heterogenetiy. I think these findings need more emphasis for this analysis to be informative.

19. As an explanation for the heritability difference between EUR and AFR GWAS, have the authors considered the set of SNPs used to estimate heritability? Are these analyses all based on HapMap3 SNPs? Might using a different/larger set of SNPs lead to higher heritability in AFR?

Prediction

20. I'm surprised by the predictive power of the PGI for neuroticism. The PGI from the second release of Polygenic Index Repository(Alemu et al., 2025) has a predictive power of 3.31% in HRS, and it is based only on UKB, 23andMe(Lo et al., 2016) and GPC(de Moor et al., 2015) with a sample size less than half of the current one. I would expect a much higher predictive power from this PGI. One difference between the two is the set of SNPs (HapMap3 vs ~2.9million SNPs(Lloyd-Jones et al., 2019)) but I don't think that's enough of an explanation. Have the authors looked into why this is the case?

21. The authors have used the EUR GWAS weights to make PGIs for the AFR samples in HRS and Add Health. I am curious why they haven't used the trans-ancestry meta-analysis. Is it because it would be complicated to obtain LD estimates reflecting the ancestry composition of the combined sample to use with SBayesR? It would still be informative to see if there is an improvement in predictive power for AFR samples using a method like PRS-CSx(Zheng et al., 2024) to incorporate both EUR and AFR meta-analyses.

22. Are the PGI associations in Panel B of Figure 2, or the genetic correlations in Figure 3 are corrected for multiple testing? Apologies if this information is reported, I couldn't find it anywhere.

23. In the residential analyses, the authors control for 5 PCs. I was surprised by this choice; it seems very few PCs especially for an analysis like this. Does controlling for more PCs change the results?

24. In SI section 6.2 about the parent-offspring PGI analysis in deCODE, the authors state there were N=34506 probands with at least one genotyped parent. Were the missing parents imputed here?

Assortative mating

25. I couldn't see any information about how the spouse pairs in UKB were identified.

Editorial

26. I find it a bit confusing that the biological follow-up paragraph on page 10 cuts in between the discussion on heritability. I think it would be better if this was a section on its own that comes after "Characterizing Common-Variant Heritability".

References

Alemu, R., Terskaya, A., Howell, M., Guan, J., Sands, H., Kleinman, A., Team, 23andMe Research, Bann, D., Morris, T., Ploubidis, G. B., Fitzsimons, E., Harris, K. M., Caspi, A., Corcoran, D. L., Moffitt, T. E., Poulton, R., Sugden, K., Williams, B. S., Steptoe, A., … Okbay, A. (2025). An Updated Polygenic Index Repository: Expanded Phenotypes, New Cohorts, and Improved Causal Inference. BioRxiv, 9, 2025.05.14.653986. https://doi.org/10.1101/2025.05.14.653986

de Moor, M. H. M., van den Berg, S. M., Verweij, K. J. H., Krueger, R. F., Luciano, M., Arias Vasquez, A., Matteson, L. K., Derringer, J., Esko, T., Amin, N., Gordon, S. D., Hansell, N. K., Hart, A. B., Seppälä, I., Huffman, J. E., Konte, B., Lahti, J., Lee, M., Miller, M., … Boomsma, D. I. (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with Major Depressive Disorder. JAMA Psychiatry, 72(7), 642–650. https://doi.org/10.1001/jamapsychiatry.2015.0554

Gupta, P., Galimberti, M., Liu, Y., Beck, S., Wingo, A., Wingo, T., Adhikari, K., Kranzler, H. R., Stein, M. B., Gelernter, J., & Levey, D. F. (2024). A genome-wide investigation into the underlying genetic architecture of personality traits and overlap with psychopathology. Nature Human Behaviour, 8(11), 2235–2249. https://doi.org/10.1038/S41562-024-01951-3;TECHMETA=38,43,45;SUBJMETA=208,2811,4014,457,477,631,649;KWRD=GENETIC+VARIATION,HUMAN+BEHAVIOUR

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N. R., Goddard, M. E., Yang, J., & Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nature Communications, 10(1), 5086. https://doi.org/10.1038/s41467-019-12653-0

Lo, M.-T., Hinds, D. A., Tung, J. Y., Franz, C., Fan, C.-C., Wang, Y., Smeland, O. B., Schork, A., Holland, D., Kauppi, K., Sanyal, N., Escott-Price, V., Smith, D. J., O'Donovan, M., Stefansson, H., Bjornsdottir, G., Thorgeirsson, T. E., Stefansson, K., McEvoy, L. K., … Chen, C.-H. (2016). Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. Nature Genetics, 49(1), 152–156. https://doi.org/10.1038/ng.3736

Tan, T., Jayashankar, H., Guan, J., Nehzati, S. M., Mir, M., Bennett, M., Agerbo, E., Ahlskog, R., Pinto de Andrade Anapaz, V., Asvold, B. O., Benonisdottir, S., Bhatta, L., Boomsma, D. I., Brumpton, B., Campbell, A., Chabris, C. F., Cheesman, R., Chen, Z., de Geus, E., … Young, A. S. (2024). Family-GWAS reveals effects of environment and mating on genetic associations. MedRxiv. https://doi.org/10.1101/2024.10.01.24314703

Turley, P., Martin, A. R., Goldman, G., Li, H., Kanai, M., Walters, R. K., Jala, J. B., Lin, K., Millwood, I. Y., & Carey, C. E. (2021). Multi-Ancestry Meta-Analysis yields novel genetic discoveries and ancestry-specific associations. BioRxiv.

Zheng, Z., Liu, S., Sidorenko, J., Wang, Y., Lin, T., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I. M., Snieder, H., Yang, J., Wray, N. R., Goddard, M. E., Visscher, P. M., & Zeng, J. (2024). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nature Genetics, 56(5), 767–777. https://doi.org/10.1038/S41588-024-01704-Y;SUBJMETA=205,208,212,2138,631;KWRD=GENOME-WIDE+ASSOCIATION+STUDIES,GENOMICS

(Remarks on code availability)
I cannot seem to access the code using the OSF link above. I get an error message that says "Failed to fetch file metadata from WaterButler. Received response"

Referee #3

(Remarks to the Author)
Review of “Robust inference and widespread genetic correlates from a large-scale genetic association study of human personality”
This manuscript represents an extraordinary scientific contribution, offering the most comprehensive analysis to date of the genetic architecture of the Big Five personality traits. The authors leverage over one million participants across 46 cohorts to generate a robust set of GWAS findings, enriched by within-family analyses, polygenic score prediction, and Mendelian Randomization (MR). Their work is remarkable not only for its scale but also for the care taken to replicate findings across subpopulations and to explore the genetic correlates of personality across a wide range of outcomes and to test causal hypotheses using MR. This is a landmark study and is, in my view, appropriate for publication in Nature pending revisions addressing the points below.
Major Comments:
1. A central claim of the manuscript is that genetic associations with personality are largely free from shared environmental confounding. While the within-family and parent-offspring results are compelling, sample sizes are modest (N ≈ 30k–50k; see p. 17–18), and the ability to detect indirect genetic effects (e.g., from parental genotype or shared environment) is limited. Null findings here should be interpreted in light of modest statistical power. The authors should focus on the confidence intervals of the indirect effect estimates (or the ratio of direct/(indirect+direct)) and revise their interpretations accordingly. This concern arises in multiple places, including the Abstract, pp. 6 and 18–19.
2. Equation 14 in the Supplement uses a family fixed effects model but does not include both individual deviations from the family PGI mean and the family mean PGI itself as separate predictors. The approach used by the authors estimates only the within-family effect and precludes estimation of the between-family component. The comparison to "matched population-level GWAS in cohorts that contributed to within-family GWAS" (p. 19) would be more informative if the same exact individuals were used to estimate both within- and between-family effects (i.e., including both within- and between-family PGI predictors). If this was not done, please justify this decision and consider re-estimating the model accordingly.
3. The inferred genetic correlations between spouses (Figure 2C; lines 310–326) are modest, but readers will want to know how these compare to phenotypic spousal correlations and may not understand that spousal rg’s are expected to be lower than spousal phenotypic correlations. Given the non-unity heritability of personality, the rg between spouses should be expected to be lower than the rp under phenotypic assortment. The authors could provide inferred rp estimates from rg values (by correcting for the heritability assuming phenotypic homogamy) and compare them to known phenotypic spousal correlations (e.g., Horwitz et al., 2022). This would clarify how the observed genetic rg's relate to reported rp’s.
4. The MR analysis is underreported in the main text. Two short paragraphs on pp. 16–17 do not do justice to the depth of the results. I recommend including a dedicated figure and/or expanded main-text table listing the significant MR results and the method used (e.g., CAUSE, weighted median). In addition, please report whether corrections for multiple testing were applied.
5. The decision to exclude SNPs only when INFO < 0.40 is unusually permissive. The authors note (p. 6 and supplement) that more stringent thresholds were applied elsewhere but do not clarify when and where. Please provide more detailed justification and clearly indicate the imputation reference panels and procedures used across cohorts. This should be included in the main text or, at minimum, expanded in the supplement.
6. Some rg results in Figure 3 are counterintuitive. For example, openness shows positive genetic correlations with ASD and OCD, and negative correlations with COVID-19 infection and sports participation. Conscientiousness negatively impacts educational attainment in MR (p. 16). The authors should help the reader interpret these surprising findings, perhaps in a new paragraph in the Discussion.
Specific Comments:
• Abstract: Final sentence is awkward—reword “The genetic architecture of personality… is fundamental to being a human.”
• Page 4: Define what is meant by “assemble” in “assemble data from 46 cohorts.” Clarify cohort selection, harmonization, and QC steps.
• Page 4 (WF GWAS): When stating “little to no shared environmental confounding,” provide confidence intervals for

estimates of indirect effects or IGE magnitudes.
• Page 5: Supplement the Roberts & Yoon (2022) citation with a more foundational reference (e.g., McCrae & John, 1992).
• Page 5 (“predict”): Clarify that predictive relationships of Big Five traits with outcomes are phenotypic, not genetic.
• Page 6 (“k = 46”): Define “k” explicitly as the number of cohorts per meta-analysis.
• Page 6: Fix typo: “across social relevant” -> “across socially relevant”
• Table 1: Excellent and informative table. Consider adding average cross-cohort LDSC rg values (possibly weighted by the geometric mean of the pairs of cohort sizes). This can help in interpreting RE-meta vs. fixed effect meta-analysis results.
• Page 7: Explain why conscientiousness shows lower disattenuated h^2_SNP relative to other traits.
• Page 8: The claim that Big Five rg’s roughly equal phenotypic correlations (Soto & John, 2017) could be tested formally via matrix comparison of rg vs. rp using maximum likelihood.
• Figure 1: Consider adding SNP h^2 estimates to diagonal of the rg matrix.
• Page 10 (MAGMA): Consider rerunning the gene-set analysis excluding neuroticism to examine its influence on shared signal.
• Page 10, Negative Selection: I think that the evidence purported to be in support of negative selection (enrichment in protein truncating variant intolerant gene sets) is pretty tenuous. Perhaps this can be omitted, or perhaps other types of evidence can be brought to bear on this issue.
• Page 11: Address power limitations in the 20K other-rated Estonian Biobank subsample when interpreting rg = .84 with self-report.
• Figure 2A: AFR PGIs: Report predictive performance when using AFR-derived PGIs (if available). Otherwise, acknowledge this omission.
• Figure 2E: Convert effect size into a correlation or similar metric for easier interpretation.
• Lines 203–207: Clarify the “M_Beta” notation (clearer would be a bar of beta) and confirm that coefficients are standardized.
• Fig 3: Age-at-first-sex: Analyze potential sex interactions. This outcome may be differentially associated with personality by gender.
• Page 16 and Fig. 2: I understand how relatives are used in addition to spouses to estimate the signatures of assortative mating, but this will be confusing to most readers. Consider adding a few sentences here explaining this in better detail.
• Page 16 (MR): Report correction for multiple testing and list the full set of exposure-outcome pairs tested.
• Page 19 (Figure legend): Clarify what “104%” and similar values mean. Specify units and estimation approach.
• Supp. p. 57: Add missing reference for Tucker-Drob (2017).
• Supp. Tables PDFs: Some PDFs failed to render; rely on Excel files for clarity.
• Data Availability: Clarify whether neuroticism sumstats will be available both with and without UKB. Currently ambiguous. Certainly hope it is both (and if not, justify).

(Remarks on code availability)

Referee #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
I have no changes to recommend. The authors have done a spectacular job addressing my prior concerns.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
I sincerely thank the authors for their thorough and thoughtful revision. They have addressed all of my comments and added extensive additional analyses, greatly strengthening the manuscript. I have no further suggestions.

(Remarks on code availability)
I'm still having trouble accessing the code. The initial OSF page opens, but if I click anything on there, I get a "You Don't Have Permission To Perform This Action" error.

Referee #3

(Remarks to the Author)
This manuscript represents an extraordinary scientific contribution, offering the most comprehensive analysis to date of the genetic architecture of the Big Five personality traits. The authors leverage over one million participants across dozens of cohorts to generate a robust and highly replicable set of GWAS findings, complemented by within-family analyses, polygenic score prediction, and Mendelian Randomization (MR). The scale and scope of the work are exceptional, and the manuscript has been substantially improved in response to prior reviews. I am satisfied with the majority of the authors' revisions, although some interpretive issues would benefit from further tempering or clarification. Overall, this remains a landmark study that is, in my view, appropriate for publication in Nature pending minor revisions.
Major Comments
1. The authors have improved the presentation and discussion of the within-family analyses, particularly by emphasizing confidence intervals and the detectable magnitude of indirect effects. These revisions are appreciated and address my prior concerns to a large extent. That said, I would still encourage slightly more cautious framing of conclusions regarding the limited role of confounding. The current results more strongly rule out large sources of confounding than they do smaller (but potentially meaningful) effects, and a brief acknowledgment of this would strengthen the interpretation.
2. The revisions have improved the section on genetic correlations among spouses. However, I wasn't aware of how spouses were detecdted until reading the authors' response to reviewer 2. Previous approaches to detecting spouses in the UKB have major problems - not just that over half of likely spouses were not included, but also that the remaining spouses were ascertained in odd and unintended ways that could bias findings (Kelly, Horwitz, & Keller, 2026). The Kelly 2026 paper explains this issue and provides a simple algorithm to detect spouses in the UKB that is much more accurate and leads to 89K – 93K spouses. We suggest authors reanalyze using this approach.
3. The MR analyses are extensive but remain somewhat underrepresented in the main text relative to their scope. While I appreciate the authors' need to streamline the MS, I recommend modestly expanding the presentation of these results, either via a summary figure or a more explicit table of key findings. In addition, it would be helpful to clearly state whether multiple testing correction was applied and, if so, how.
Comments on Prior Review Points
I appreciate the authors' careful and thorough responses to prior concerns. Most issues have been addressed satisfactorily.
• The revisions to the within-family analyses and their interpretation are appreciated, although, as noted above, I encourage continued caution in drawing strong conclusions about the absence of confounding.
• Re 3.6: It is satisfying to read the corresponding supplementary text explaining these interesting correlations. I do not believe ascertainment bias fully explains the negative correlation between conscientiousness and educational attainment. Exploring the genetic correlation between conscientiousness and cognitive ability might clarify the picture a little better. However, I recognize this is not the primary focus of the current paper, and I am perfectly satisfied with the authors' current approach.
• Re 3.16: Thank you for providing the comparison table between genetic correlations and phenotypic correlations. I agree with the authors that most are consistent. However, it is intriguing that the genetic correlation between openness and conscientiousness operates in the opposite direction of the phenotypic correlation. Do the authors have a brief interpretation for this divergence?
Minor Comments
● Abstract: The final sentence remains somewhat awkward and could be rephrased for clarity.
● p. 6: "Western geography" — This phrase is ambiguous. I recommend using a more precise geographic or demographic descriptor.
● p. 21: "β .158, compared to β .165" — Missing equality signs (e.g., β = .158, compared to β = .165).
● p. 24: "The ratio of direct to confounding estimates is presented at the bottom of the corresponding bar (Supplementary Table S44)" — To clarify, should this ratio be the direct effect over the sum of the direct and confounding effects (direct / [direct + confounds])?
● p. 25: "mean r between rg estimates = .83" — I suggest revising this to "mean correlation among r_g estimates" for greater clarity.
● p. 25: "within-family estimates replicated population-level estimates (from independent cohorts that did not contribute within-family data) at 95% of the magnitude that the population-level estimates from the same within-family cohorts replicated the population-level estimates (from the independent cohorts that did not contribute within-family data; Supplementary Table S47)." — The phrasing here is dense and confusing. I recommend breaking this into two distinct sentences to clarify the comparison.
● p. 27: "For agreeableness and neuroticism, genetic correlations between within-family and population GWAS were imperfect." — I suggest replacing "imperfect" with a more neutral and precise term, such as "not identical" or "attenuated."

(Remarks on code availability)

Referee #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Version 2:

Reviewer comments:

Referee #2

(Remarks to the Author)
I have no further comments.

(Remarks on code availability)
I can finally access the code and it looks fine to me. It has a README and is clearly annotated. It would be easy to reproduce the results of the paper with access to the same data.

Referee #3

(Remarks to the Author)
I have read the new manuscript and have no further comments. Nice work and we hope to see it published soon.

(Remarks on code availability)

Referee #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)
The code looks good.

Comments from AE:

You will see that the reviewers very much appreciate the study and it appears that all comments should be addressable.

Thank you! We have addressed each of these comments in full, with changes to the manuscript highlighted. Additionally, in this revision we have reformatted the document to align with Nature's formatting requirements. Specifically, we have added a supplemental Online Methods section and reformatted the supplementary text figures into Extended Data Figures. We welcome any comments on these additions to the manuscript.

Furthermore, after adding the requested revisions by the referees, and trimming points of redundancy across the manuscript, the main text is now at approximately 5,000 words, of which 500 are the introduction, 3,800 are the main results, and 700 are the discussion. If you or the reviewers have any comments on how to further trim the manuscript to comply with Nature's 4,300-word guideline, including content that can be moved to the Supplementary Text, we are happy to comply.

Besides the technical comments, we want to emphasize our strong agreement with R1 regarding the importance of cautious and nuanced presentation and interpretation of results. Thus, in addition to addressing the technical comments, we have paid particular attention to R1's comments about presentation and used them to make changes throughout the manuscript, especially in the introduction and discussion where effects are contextualized. We respond to these specifically in R1.1-5.

It would also seem a good idea to signpost the FAQ more prominently in the main text of the paper.

We now prominently signpost the FAQ in the introduction, on p. 6. We have added additional questions and answers to the FAQ, in response to Reviewer 1's suggestions. The FAQ is located at https://osf.io/hgnsm/files/crh9d?view_only=faa526bbeabc4615b5d6c0fbc52e10ba.

Referee #1 (Remarks to the Author):

Globally, this is a very important paper for the field of personality genetics and the field of personality more broadly. It is as definitive as we can get with the technologies at hand to ask basic and important questions about genetic signatures found in personality phenotypes.

We appreciate this reviewer's very positive evaluation of the manuscript.

R1.1 I found the write up to be enthusiastic, maybe a bit too enthusiastic. Statements like "personality genetics have widespread, potentially causal associations with a wide range

of consequential behaviors and life outcomes" while true in some senses is probably too enthusiastic given the caveats that need to accompany those statements which neither the abstract, nor the brief nature of a Nature article format afford. Given the history of how genetics research has been misused by problematic parties (mass murderers, for example), I think it would be prudent if the authors contemplated the fact that their work will reach less sophisticated audiences than the typical Nature reader. Temperance and humility might be a better approach than enthusiasm.

We agree about the value of careful messaging for this paper's audience. In revision, we have tempered our language considerably throughout the entire paper.

In the abstract, we now write that "*Common genetic variants explain a moderate 4.8%-9.3% of the variance in each trait"* (emphasis added) to convey that the h^2_{SNP} of personality is moderate in comparison to the h^2_{SNP} of other phenotypes. We choose the qualifier *moderate* both to signify that this magnitude is notable (for comparison, the mean h^2_{SNP} of continuous phenotypes in the UK Biobank is 2%; https://nealelab.github.io/UKBB_ldsc/h2_browser.html) while at the same time remain realistic that this is still a minority of the variance. Additionally, we now write that *"personality traits* (rather than personality genetics) *have widespread, potentially causal associations with consequential behaviors and life outcomes"* to convey a more precise and less problematically interpretable inference, which is that personality traits are the causes of these associations.

We also now more prominently signpost the FAQ in the introduction. The FAQ explains the rationale and results of this study in everyday language for non-academic audiences, immediately after introducing the GWAS (p. 6) and directly addresses challenging and important questions such as *"if I inherit a certain personality gene, does this mean I will have a different personality?" "Does this research promote eugenics?,*" and the newly added *"Are genetic associations causal?"* in plainer language, and in further detail, than is contained in the main text. Furthermore, in the introduction (p. 6) and discussion (pp. 26-27), we now describe estimates of h^2_{SNP} and contrast them with the higher h^2 estimates from twin and family studies, and have now added that *"personality is substantially influenced by a person's environment*." These changes explicitly communicate the substantial environmentality to personality traits, setting the paper up for a more balanced presentation that does not overstate the magnitude of genetic effects.

In the results section, we explicitly describe how PGI analyses *"predict outcomes later in life from genetic variants inherited at conception, via nondeterministic transactions with the environment*." (p. 18) We also caveat that this prediction is small in magnitude and ineffective for applied behavioral prediction for a given individual, writing as plainly as possible, *"These PGIs cannot predict a specific individual's future behavior*." (p. 19).

Finally, in the discussion section, we have added substantial new language that contextualizes the genetic effects on personality more carefully. We restate that personality's origins are partially genetic and partially environmental (p. 26), we have excised the word "causal" from the

description of genetic effects (which we now describe as “minimally confounded”), we have added language describing the evidence for this minimal (not absent) confounding (p. 27), and we state that even direct genetic effects *“do not determine an individual’s personality.”*

If the reviewer can identify other areas where re-wording or softening can facilitate appropriate interpretation of these findings by those less familiar with genetics research, we are happy to make additional changes.

R1.2 Similarly, after a compelling section outlining the importance of personality to human functioning the authors argue that “Given the fundamental relevance of personality to core phenomena studied across scientific disciplines, understanding the molecular genetics of personality is crucially important for developing biopsychosocial models of both human behavior and life outcomes.” Is it? How? We know things are genetic. We know the GWAS data are challenging. We know there are no simple “gene for” associations. As an avid consumer of these types of studies, I already knew going in that the big issue is the rather large discrepancy between twin estimates of heritability and GWAS estimates with the latter being disturbingly small, comparitavely speaking. This context does not appear to inform the enthusiasm for the effort. Nor is the discrepancy discussed enough for my taste. Couple this with the way genetics studies and heritability findings have been used by authors in the past, and one would be concerned with the lack of perspective taking. Some authors took the most extreme twin estimates to argue that there is no nurture in personality’s nature. Others gleefully cite the low GWAS heritabilities of personality as evidence for the Blank Slate. While it might be a mistake to infer the opposite with the small heritabilities characteristic of GWAS efforts, it is hard not to take the findings as an indication that genetics are far less important than originally thought. There are many complicated reasons not to draw this conclusion and most of those complexities are not registered here. The framing comes across as “look, its genetic!”, which, after all is not a very interesting research question at this point in history. Somehow, some of the complexity of the findings needs to be infused into the write up despite page limitations. To that end, I believe the authors have a responsibility to more consistently and directly handle the effect size issue. The fact that the variance accounted for is quite small and much smaller than twin studie should be a theme throughout.

In revision, we have made it clearer in the introduction how and why understanding the molecular genetics of personality is important for, as we write, “developing biopsychosocial models of human behavior and life outcomes.” In the remainder of the introductory paragraph (pp. 6-7), we preview key results from the study that inform a biopsychosocial model of human behavior, including findings for high consistency of genetic effects on personality across environments, neurobiological characterization of genetic effects on personality, and evaluation of confounding in genetic effects on personality. Each of these tests leverage molecular genetics to understand and link human behavior through biology and environments, in ways that cannot straightforwardly be studied using other research methods.

A second major point that the reviewer touches on in this comment is that it is important to discuss the discrepancy between molecular estimates of common variant SNP heritability and estimates of heritability from classical twin and adoption studies that compare similarity of individuals without measuring DNA, as a function of the extent of relatedness and shared rearing. We now do this in three places. In the introduction, we now specifically define h^2_{SNP} as the additive effects tagged by the common genetic variants we measured, and mention that our estimates are consistent with previous GWAS of personality, but lower than estimates of h^2 from classical twin and adoption studies. In the discussion section, we have added a new paragraph that compares and contrasts common variant SNP heritability with both rare-variant and pedigree-based estimates of heritability. We write that our h^2_{SNP} estimate are similar to past personality GWAS, but lower than h^2 estimates both from classical twin and family studies and molecular methods that index effects of rare variance and genetic interactions. In other words, these changes convey that h^2_{SNP} is an important quantity that nonetheless does not index all genetic effects.

We would like to emphasize here that we find it important to contrast h^2_{SNP} not just with classical twin and adoption estimates of h^2, but with other molecular genetic estimates of h^2. For example, an estimator for h^2 that relies on measured variation in genetic relatedness between sibling pairs, who share between 40-60% of their genetic variants on average, produces h^2 estimates that are expected to include effects of rare variants, structural variants, gene-by-shared environment interactions, and (to some extent) gene-by-gene interactions not well captured by estimates from standard GWAS. Using this method, Markel and colleagues (2025; https://dx.doi.org/10.2139/ssrn.5225447) estimate the h^2 of well-being at 34% and h^2 of risk-taking at 44%, and Yengo et al. (2025; https://doi.org/10.1101/2025.09.17.25336022) estimate the h^2 of “talk vs. listen” (a single item measure of a component of extraversion) at 48%. These estimates from molecular genetic data are far larger than estimates of common variant h^2 and much closer to the 40-60% h^2 estimates from classical twin and adoption studies. We would like to emphasize that we have not hand-picked the highest estimates; rather, these are currently the only estimates available for any personality traits (both papers also report estimates for other traits that are not as closely related to personality). We enthusiastically anticipate results of future research applying this and related approaches to even larger samples and a broader array of personality traits.

1.3 Also, I understand the authors impulse to correct their results for unreliability, especially in the context of a meta-analysis. There are always studies that have glaring issues and it seems wrong to simply merge those estimates together with better conducted studies. That said, believing in something fantastical like perfect measurement reliability is equally problematic. To date, we have never achieved it and we are unlikely to to so anytime soon. I propose a compromise. Adjust the results for the highest achievable average reliability estimates we find in the field–.80 or so. That way, you can report more realistic estimates (e.g., when the study is done well) rather than in some hypothetical universe with unicorns and perfect reliability.

We appreciate this excellent suggestion. The median alpha reliability across personality measurement instruments in this meta-analysis is .81, and we have now re-estimated reliability-disattenuated heritability estimates at this level of alpha (rather than only at alpha = 1.0). In the abstract, we now write,
"Common genetic variants explain a moderate 4.8%-9.3% of the variance in each trait, and 9.3%-13.3% among instruments with typical measurement reliability."

In the main text, and in Table 1, we now report and describe these estimates at this typical level of reliability alongside the estimates that fully correct for measurement unreliability.

Finally, in the supplement (pp. 16-17), we now describe these additional analyses and the logic behind them. Importantly, we contrast the reasons why one would want to disattenuate for *typical* measurement unreliability (when one's goal is to estimate the observed association with a manifest trait under conditions of imperfectly reliable measurement) or disattenuate for *all* measurement unreliability (when one's goal is to estimate the true association with a latent trait for which observed scores serve as an imperfect proxy).

1.4 I like very much the material on what the PGS predict, though I would have prefered some discussion of mechanism here. Are we to infer that if I get some good genes, these genes will act directly on whether another human sees me as well-groomed (as we currently define it)? Or, should we assume there are intervening factors that act on these genetic propensities to manifest the phenotype of "well-groomed"? If it is the latter, a more detailed discussion of this process as well as some descriptive hedging of the PGS predictions wouild be good. At this juncture, given the way it is written, some of your readers can conclude that being born with a certain suite of genes causes, directly, you being a criminal. Is that really what the authors are proposing?

Although many of these PGI predictions are silent with respect to mechanism, we have added language (p. 18) to better describe the theorized mechanism connecting genes with outcomes like interviewer perceptions, of evocative gene-environment transactions, whereby personality-relevant genetic variants partially predict traits and behaviors that affect which environments and experiences individuals seek out as well as how those individuals are perceived and responded to by others (Scarr & McCartney, 1983).

Additionally, in the introduction and conclusion of this PGI prediction section, we now emphasize that these effects are small in magnitude, operate through transactions with the environment, reveal average associations based on large samples of individuals, are only probabilistic, and cannot accurately predict a specific individual's future behavior (see R1.1). For instance, we use the qualifier "more often" to emphasize that individuals with higher PGIs didn't all evince the expected outcome, but simply more often experienced it.

Finally, as described earlier (R1.1.), we have taken care to create and signpost an FAQ to directly address questions such as *"if I inherit a certain personality gene, does this mean I will have a different personality?*

1.5 Again, it would be preferable to condition the predictive relationships on the magnitude of the effects rather than statistical signficance. The current write up uses the latter and leaves the mistaken impression that everything is strongly related to everything else. Relatedly, describing the findings on Figure 4b as "robust" is more marketing than science. Robust is meaningless here unless the authors have given the reader a metric of effect sizes to understand the magnitude of the effects.

We agree that it is important to be clear about the magnitude of the effects obtained, and have therefore added qualifiers such as partially," "slightly" and "small-to-moderate associations" that also communicate the probabilistic nature of these findings. At the same time, it is important to recognize that, while tests of PGI-external outcome associations are useful to identify associations and their signs, they do not provide direct information about the true magnitudes of genetic associations. This is because PGI associations are attenuated by the substantial unreliability in the genetic instrument, which is a direct function of the power of the discovery GWAS used to estimate PGI weights.

On the other hand, tests of genetic *correlation* between two GWASed traits are not attenuated by unreliability (of their genetic measure or the GWAS phenotype) and thus do contribute useful information about effect sizes, on a standardized correlational scale. For this reason, we test genetic associations using genetic correlations whenever possible and we have added effect sizes (r_gs) to our description of genetic correlation results across pp. 15-18. In the PGI section, we now explicitly state that the purpose of these PGI predictions is to substantiate connections between personality traits and life outcomes that have yet to be submitted to GWAS (p. 19). As we do not recommend that the PGIs be used for the prediction of individual outcomes in an applied clinical or educational setting, we refrain from emphasizing the magnitude of prediction, but we provide labelled x axis ticks in Figure 2B to denote effect sizes, and we report exact estimates in Supplementary Table S38.

Finally, we would like to clarify that we use the term robust when specifically referring to the finding that estimates of genetic effects on personality using population-level methods (population-GWAS and population-level PGI associations) are generalizable across groups and highly similar to estimates using within-family methods (within-family GWAS, within-DZ-twin-pair PGI associations, and parent-offspring PGI associations) that rigorously control for a wide range of potential confounds, including uncontrolled population stratification, dynastic effects, and assortative mating. As we discuss, this finding is especially notable given that that several other social science traits (e.g. educational attainment, cognitive performance, and externalizing behavior) exhibit marked reductions in prediction when comparing within-family analyses to population-level analyses. In revision, we now use the term "*robustly generalizable [and] minimally confounded*" to clarify its meaning.

1.6 The most profound findings, in my opinion, are the null findings. There was no strong signal in regions related to dopaminergic and serotonergic systems which have been the pillars of personality neuroscience if not psychological neuroscience. There was no

strong difference between self-reports and observer ratings of personality. The PGS worked equally well across ancestries. There is little or no assortative mating on personality. There is little or no environmental confounding. These are quite possibly some of the most significant null effects I've read, as the alternative findings or threatened findings have been so significant to the field. The dopamine and serotonic hypotheses have ruled physiological personaltiy psychology for decades. Self-reports are commonly dismissed as invalid because of reporting biases. Lack of generalizability of the genetic findings is commonly used to argue against genetic findings. Assortative mating is often invoked as another problematic confound for genetic as well as phenotypic research on personality. This paper will reset the literatures of multiple fields.

We appreciate this, and we now signpost these findings more visibly in the introduction for readers to take note (pp. 6-7). Specifically, we mention the null findings regarding personality neuroscience, environmental confounding, and assortative mating in the introduction, while also highlighting the generalizability of findings. We have also streamlined the discussion paragraph on confounding to better highlight the null and near-null findings of this study.

Referee #2 (Remarks to the Author):

The study reports findings from large-scale GWAS on the Big-5 personality traits. This was a big gap in the field – there were no published well-powered personality GWAS until very recently(Gupta et al., 2024). The current study, by also incorporating data from the latest GWAS (Gupta et al., 2024), conducts the largest personality GWAS to date and obtains important insights about the genetics of personality, as well as facilitating future studies on a wide range of research questions involving personality that could not be conducted before due to the low statistical power of the GWAS. As such, this is a very important contribution. There is a lot to like a about this paper. It is very comprehensive, carefully conducted and was a pleasure to read. However, I believe there are some points that could be improved which I list below. I realize this is a very lengthy list of comments but a lot of them are just clarification questions that I hope will be quick to address.

We appreciate the reviewer's careful read of the paper, and we have responded to each of their comments below.

QC and meta-analysis

2.1 It was unclear to me what QC filters have been applied at the cohort-level prior to meta-analysis other than dropping multiallelic SNPs or those that are not in 1000 Genomes. There is no mention of a MAF filter, and only an INFO filter of 0.4 appears to have been applied prior to GWAS by cohorts. Is that all?

We have now added an online-only methods section in line with *Nature*'s formatting requirements. In this new section, we now more clearly describe cohort-level QC in terms of SNPs and participants (pp. 33-35). For the primary population-level GWAS of each trait, we write that we excluded SNPs with MAF < 1%, Call rate <95%, deviation from Hardy-Weinberg

Equilibrium ($p < 1x10^{-5}$), INFO < .40, and those with poor clustering on visual inspection of intensity plots. We also write that we excluded participants if they had low overall call rates (< 95%), excess autosomal heterozygosity or homozygosity, were duplicated samples, had gender inconsistent with sex or had chromosomal abnormalities. Furthermore, in the supplement, we have added additional language clarifying particulars of QC that do not fit in the online methods section, including for newly added analyses in revision (pp. 3-12, highlighted text).

We also now provide full details (including QC filters) of other analyses as relevant. For instance, we provide full details of within-family GWAS in section 6 of the supplement, and we provide details of QC for within family GWAS in 6.4 ("Within-Family GWAS Quality Control", where we implemented a stringent filter of INFO ≥ .99).

2.2 It would be good to see the INFO scores for the lead SNPs in Supplementary Tables 5-9.

We have added these INFO scores as requested (Supplementary Tables 6-10). As expected, nearly all lead SNPs have very high INFO scores (e.g., the mean INFO for extraversion lead SNPs is .97).

2.3 I was surprised to see that the authors used a 250kb window when clumping the genome-wide significant SNPs. That is a window much smaller than is used in the literature and is likely to identify SNPs in long-range LD as independent hits. As the authors also report, we know that at least for neuroticism, some of the associated SNPs lie on inversion regions. 250kb distance will not be enough to cover these regions, leading to overreporting of approximately independent SNPs. Do the authors have a justification for using such a small window?

We used the 250kb window for clumping lead SNPs as it has often been used in other genome-wide meta-analyses of personality (Nagel et al., 2018) and psychopathology (Grotzinger et al., 2022) and serves as the default window in FUMA (Watanabe et al., 2017). We agree with the reviewer that this may have identified some SNPs that are in LD despite being more distal from one another on the genome. To address this, as the reviewer suggested we estimated transancestry-weighted LD across all lead SNPs on each chromosome to identify any lead SNPs with $r^2 > .10$ irrespective of distance.

The results of this re-analysis indicate that 1,239 of 1,260 (98%) of lead SNPs have $r^2 < .10$ regardless of genomic distance (see below table). In other words, this analytic decision has little effect on the identification of lead SNPs for this study. We now mention this in the main text when reporting lead SNPs (p. 7), flag in Supplementary Tables S6-S10 all SNPs that are in long-range LD with others on the chromosome (and which SNPs they are in LD with) so that researchers can identify these lead SNPs, if they are interested, and we discuss this additional analysis in the Supplementary Text. We have also added the below table to the Supplement (Table S13):

Trait	Total reported lead SNPs (*r*2 < .1 AND <250kb)	Lead SNPs in long distance LD (r2 < .1 AND >250kb)	Lead SNPs independent regardless of distance (r2 < .1 whole chr.)
Extraversion	258	3	255
Agreeableness	39	2	37
Consc..ness	131	0	131
Neuroticism	706	14	692
Openness	127	2	125
Total	1260	21	1239

Grotzinger, A. D., Mallard, T. T., Akingbuwa, W. A., Ip, H. F., Adams, M. J., Lewis, C. M., ... & Nivard, M. G. (2022). Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis. *Nature Genetics*, 54(5), 548-559.

Nagel, M., Jansen, P. R., Stringer, S., Watanabe, K., De Leeuw, C. A., Bryois, J., ... & Posthuma, D. (2018). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature Genetics*, 50(7), 920-927.

2.4 Related to my point above, in order to identify novel SNPs, the authors check whether any of the genome-wide significant SNPs found in previous research are within the 250kb loci that they've identified. This, also, is a very anti-conservative definition of a novel hit. It would be good to see how many SNPs identified in the current study have a squared correlation less than 0.1 with all of the previously identified ones on the same chromosome.

We have now conducted these requested analyses to examine the LD-independence of all novel lead SNPs with previously identified lead SNPs. Overall, there is an extremely close match (99%) between the novel lead SNPs identified by our previous criterion (824 novel variants) and the criterion suggested by the reviewer (819 novel variants). Only 5 novel SNPs are >250kb away on the genome but not LD-independent.

In particular, for extraversion and agreeableness, all previously identified lead SNPs are also LD-independent (for extraversion, max $r^2 = .03$, agreeableness max $r^2 = .02$). For each of conscientiousness and openness to experience, one SNP we identified as novel was greater than 250kb away on the genome but not LD-independent of a previous hit. Finally, for neuroticism, three SNPs we identified as novel were not LD-independent. For full transparency, we now note this in the main text (p. 7), and flag these five SNPs in Supplementary Tables S8, S9, and S10, in the list of novel and replicated lead SNPs.

2.5 It would be interesting to look at what the dosage compensation parameter actually is, whether the heritability due to the X chromosome differs between males and females,

and what the X chromosome male-female genetic correlation is. Some of these analyses might require individual-level genetic data but it would be possible to estimate them in UKB at least for neuroticism.

In the previously submitted X-chromosome GWAS analyses, we only had sex-stratified summary statistics from N ~ 50,000 participants, which is not a substantial enough sample to test the reviewer's questions. Thus, to address these questions, we incorporated new analyses of neuroticism using 168,856 male and 197,927 female participants in the UKB cohort. These analyses were conducted by two researchers who have been added as co-authors on this resubmission, T. T. Mallard and J. D. Tubbs. In the main manuscript (p. 11 and p. 36) and the supplement (pp. 8-9 and p. 58) we describe the methods and results of these analyses. Here, we describe them in brief.

Using PLINK 2.0, we first conducted sex-stratified analyses of x-chromosomal effects on neuroticism in males and females, with effects coded as 0/1 in males and 0/1/2 in females. Using the methods described in Sidorenko et al. (2019) and Lee et al. (2018), we found no sex differences in X-chromosomal genetic effects on neuroticism. X-chromosome-linked h^2_{SNP} did not differ between males (.23%; SE = .04) and females (.18%; SE = .03; $p_{difference}$ = .32), and these genetic effects were correlated near-unity across sex (r_g = .96; 95%CI = [.81, 1.10]). The dosage compensation ratio parameter indicated partial dosage compensation (1.26, SE = .30). However, the confidence interval for this parameter was very wide, limiting our ability to reliably incorporate the estimates into a cross-sex meta-analysis. To determine how sensitive such a meta-analysis is to the coding of male dosage, we conducted cross-sex meta-analyses in the UK Biobank cohort using both zero dosage compensation (males 0/1, females 0/1/2) and full dosage compensation (males 0/2, females 0/1/2). In the model with zero dosage compensation, the mean chi-square test statistic was 1.51, with a max chi-square of 31.07 (corresponding to a p-value of 2.49×10^{-8}). In the model with full dosage compensation, the mean chi-square test statistic was 1.52, with a max chi-square of 29.43 (corresponding to a p-value of 5.79×10^{-8}). That these results are very similar to one another indicates that the choice of dosage compensation coding has a trivial effect on power for cross-sex X-chromosomal meta-analysis. Lee and colleagues (2018) reported similar findings in their tests of dosage compensation in educational attainment.

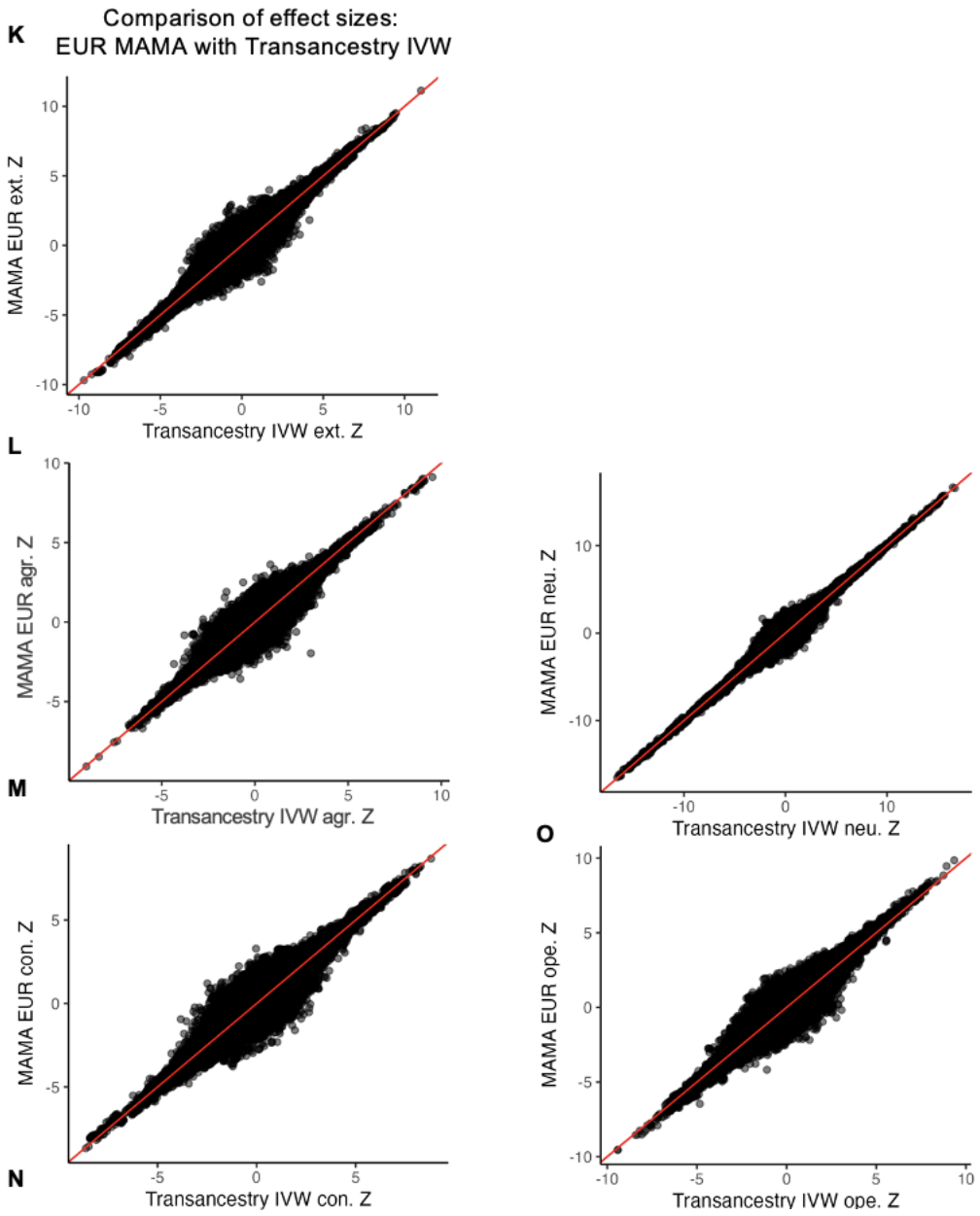
Finally, we incorporated this additional cohort into our X-chromosomal meta-analysis. Combining this data with the previously-included six cohorts led to the identification of three lead SNPs for neuroticism on the X chromosome. Two of these had been previously identified by Luciano and colleagues (2021) and one was novel. We have updated the manuscript to include these three additional lead SNPs.

2.6 The authors apply inverse-variance-weighted meta-analysis to combine the EUR and AFR GWAS. Why not use something more sophisticated designed for this purpose, e.g. MAMA (Turley et al., 2021)?

To address this comment, we have re-estimated the transancestry meta-analysis in MAMA, as suggested by the reviewer. MAMA produces two sets of summary statistics for each Big Five trait: one set of EUR summary statistics boosted by AFR participant data (which we refer to as EUR MAMA), and one set of AFR summary statistics boosted by EUR participant data (AFR MAMA). This is slightly different from our transancestry inverse variance weighted meta-analysis (transancestry IVW), which produced one set of transancestry summary statistics.

EUR MAMA results were nearly identical to the previous transancestry IVW results across many inferential angles, which we present in Supplementary Table 16. This table shows how EUR MAMA SNP z-statistics correlate $r = .98$ on average with transancestry IVW SNP z-statistics. Among SNPs incorporated in both analyses, EUR MAMA mean chi-square values are highly similar to transancestry IVW mean chi-square values. The number of lead SNPs was nearly identical across the two methods. And, as we report in Supplementary Table S16 and visualize below and in Extended Data Figure 1, a regression of EUR MAMA z-statistics on transancestry IVW z-statistics reveals an intercept at almost exactly 0 and a slope near 1, indicating that genetic effect sizes are nearly identical across analyses, especially for SNPs with strong statistical signal.

In contrast, AFR MAMA was considerably underpowered relative to the transancestry IVW GWAS, EUR only GWAS, and EUR MAMA. As was the case when comparing transancestry IVW GWAS to ancestry-stratified GWAS, EUR MAMA had a much larger mean chi-square than EUR only GWAS, and AFR MAMA had a much larger mean chi-square than AFR only GWAS (Supplementary Table S16).



Z-statistics compared between Transancestry Inverse-Variance Weighted Meta-Analysis (Transancestry IVW) and EUR Mixed-Ancestry Meta-Analysis (MAMA EUR). This figure is now included in Extended Data Figure 1.

In the supplementary text, we now describe some unique discoveries from MAMA analyses relative to our previous IVW analyses. In particular, EUR MAMA analyses of openness to experience identified 7 novel lead SNPs independent of all others across the genome, which were not significant in IVW analyses (these SNPs were typically just short of the genome-wide significance threshold in transancestry IVW analyses; mean $p = 1.1 \times 10^{-7}$, and are now just beyond this threshold; mean $p = 3.8 \times 10^{-8}$). We have added these SNPs to Supplementary Table S10 alongside the other openness lead SNPs. However, to avoid inflating our number of

reported lead SNPs in the paper, we do not count these SNPs towards our openness lead SNP total.

We believe that transancestry IVW analyses and MAMA analyses can serve as useful complements. Given space limitations of the main text, and the high degree of similarity between these results, we now note in the results section of the main text that we supplemented our transancestry IVW analyses with MAMA analyses, and we describe the results in full in the supplementary text (section 2.8).

2.7 In Section 2.7 of the SI, the authors say "we aligned SNPs to the 1000 genomes 3v5 EUR reference panel". I am not sure what "aligning" means here. Do they mean the summary stats were clumped using EUR data? If so, why not use the combined EUR and AFR reference data that they've generated to identify the novel SNPs?

As the reviewer correctly mentions, we used 1KG EUR LD patterns in a first stage for clumping when identifying novel transancestry lead SNPs, which we now clarify in the supplement. However, we then proceed to use transancestry LD patterns in a second stage. We chose the 1KG EUR reference for the first stage for two reasons: first, ~95% of the sample is EUR, and 5% is AFR, biasing sample LD structure heavily towards EUR. Second, patterns of LD are generally stronger in EUR than AFR data, leading to an overall conservative estimation of novel lead SNPs in the transancestry analyses.

Importantly, as the reviewer references, and as we now clarify, we then evaluated in a second stage whether these novel lead transancestry SNPs were independent ($r^2 < .10$) by creating combined EUR and AFR local LD reference data using SNPs found in both reference panels. We indicate this in Supplementary Table S11, where we show that only eight of 111 lead SNPs identified in the first stage of these transancestry analyses were not independent ($r^2 \geq .10$) according to transancestry-weighted LD computed in the second stage.

Overall, as we now describe more completely in the supplement, this two-step process ensures that the identification of transancestry lead SNPs is slightly conservative (owing to stronger LD, and thus fewer independent SNPs, in 1KG EUR reference data used in the first stage), confirms that these lead SNPs are indeed LD-independent (as we test independence using transancestry weighted local LD matrices in the second stage), and is computationally efficient and replicable (we provide code for users to estimate these local transancestry LD matrices on their own).

Notably, for the transancestry analyses, we included SNPs in the 1000G 3V5 (1KG) EUR reference panel and excluded SNPs that were solely included in the 1KG AFR reference panel. Again, 95% of the sample is EUR, and 5% is AFR. In AFR participant analyses, we test the significance of these 1KG AFR SNPs. This strategy is more inclusive than filtering SNPs that appear in both the EUR and AFR 1KG reference panels (for results using SNPs in both reference panels, see MAMA results in response R2.6).

2.8 It would be nice to see the ancestry-specific MAF and INFO scores for the SNPs in SI Table 10.

We have now added these INFO scores as requested.

Environmental confounding

2.9 The LDSC intercept for the neuroticism meta-analysis is reported to be less than 1. Similarly, many of the cohort-level LDSC intercepts in Supplementary Tables 1 and 2 appear to be below 1. This often happens because the effective sample size is lower than the reported sample size (e.g. when the GWAS is run with mixed linear models). The authors say that they have re-calculated effective sample size of the meta-analyses because of this. A few questions and suggestions about this:

a. Was this not done at the cohort level before running LDSC? Why are so many intercepts still below 1?

Although mixed linear models can deflate LDSC intercepts below 1, many of the individual cohorts with intercepts below 1 in Table 1 were estimated with standard linear GWAS models, and we attribute their nonsignificant intercept deflation to statistical noise. Indeed, LDSC may not work perfectly in small samples, because like most statistical methods, it is only asymptotically unbiased.

As the reviewer notes, for population-level meta-analysis, we estimated Nhat from N at the meta-analytic level, *after* combining data across cohorts. The reported neuroticism LDSC intercept is 0.987 (SE .018), which is within one standard error of 1. To follow up on the reviewer's comment, we re-estimated Nhat at the cohort level *before* meta-analyzing results. When doing this, the LDSC intercept was nearly identical (.986, SE = .017). Thus, calculating Nhat before or after meta-analysis has no measurable effect on our GWAS of these continuous-scale personality traits.

b. Why is the neuroticism intercept still below 1 even though it was estimated with effective N?

This is a good question. Importantly, as we write above, the intercept is not significantly below 1; we attribute this deviation from expectation to statistical noise.

c. It would be more informative to see the LDSC ratios rather than the intercepts in all cases to understand how the confounding compares to signal.

We have provided these ratios alongside the intercepts and slopes in all relevant supplementary tables.

2.10 Why have the authors not used the Tan et al.(Tan et al., 2024) within-family neuroticism GWAS instead of Howe et al. given the worries about the low INFO score filter applied? Isn't the Tan et al. GWAS better powered as well?

We agree that INFO > .99 is a critical consideration in these within-family analyses. We filtered SNPs from Howe et al. to only include those with INFO above this threshold.

We did not use data from Tan and colleagues (2025) because, with the exception of the HUNT cohort, all cohorts included in Tan and colleagues were also individually included in this study's within-family meta-analysis (UK Biobank, Minnesota Center for Twin and Family Research, Generation Scotland, Finnish Twin Cohort, Netherlands Twin Registry, and QIMR). The HUNT cohort, which was not included in our meta-analysis, is unfortunately omitted from Tan and colleagues' publicly shared summary statistics. In their data sharing note, they have written *"We will update the publicly available summary statistics with the HUNT summary statistics following publication of relevant HUNT studies."* Thus, our analyses incorporate all currently available within-family neuroticism data.

2.11 SI Section 6.4 states the QC procedure mirrored the full-population analysis except for the stated modifications. Does that mean no MAF filter was applied here, either?

We applied the same filter, which includes SNPs with MAF ≥ .01, across all analyses. In the online methods section that we have added in revision, we make this clear when describing within-family cohort QC.

2.12 The authors find a within-family GWAS LDSC intercept significantly higher than 1, and also not lower than the population GWAS LDSC intercept. This is surprising and requires some discussion. It would also be useful to see the ratio statistics rather than intercepts here. Do the LD scores not match the GWAS, or is it the genetic architecture of these traits that don't match LDSC assumptions?

We are happy to discuss these findings further, and we have added the LDSC ratio statistics here (and to the LDSC results reported in Supplementary Tables 1, 2, 4, 46 and 47) to facilitate this discussion. In general, one would not expect the LDSC intercept of within-family GWAS to be higher than population GWAS, because the intercept indexes confounding, which is lower in within-family than population-level analyses. As the reviewer states, within-family GWAS intercepts (mean = 1.037, mean ratio = .49) are indeed higher than the full meta-analytic population GWAS intercept (in the k = 46 samples, mean = 1.012, mean ratio = .02). Crucially, though, when comparing intercepts in the matched set of cohorts (k = 12) that contributed to both population-level and within-family analyses, the intercept is lower in the within-family analyses, as expected. In this set of matched cohorts, the mean population LDSC intercept is 1.064 (mean ratio = .37), compared to 1.037 for the within-family intercept.

These comparisons suggest that for each of the Big Five, the within-family intercept is lower than the population-level intercept (Table S47), indicating that within-family analyses indeed reduce confounding as expected. We now mention this comparison of intercepts across matched cohorts in the main text (p. 24) and describe this comparison in the online methods section.

For further information on inflation in population-level and within-family intercepts, one can also compare population-level LDSC intercepts derived from standard GWAS (Supplementary Table S1) to population estimates from methods designed to explicitly separate effects into within- and between-family components (snipar and the howe within-family pipeline; Supplementary Table S45). For example, in the NTR cohort, the average population-level LDSC intercept estimated using traditional methods is 1.012, whereas the average population-level intercept using snipar is 1.039. This suggests that 1.00 may not be the appropriate expectation for null confounding in LDSC analyses of models that decompose genetic effects into population-level and within-family components. Establishing and benchmarking expectations for intercepts of these models is, we believe, beyond the scope of the present investigation.

2.13. Were the within-family - population-level genetic correlations estimated by LDSC so that population stratification is accounted for? Since the authors claim there is negligible confounding in the GWAS, how do they interpret the correlations for agreeableness and neuroticism being significantly less than 1?

Yes, these genetic correlations were estimated by LDSC. We now mention this specifically in the added Online Methods section.

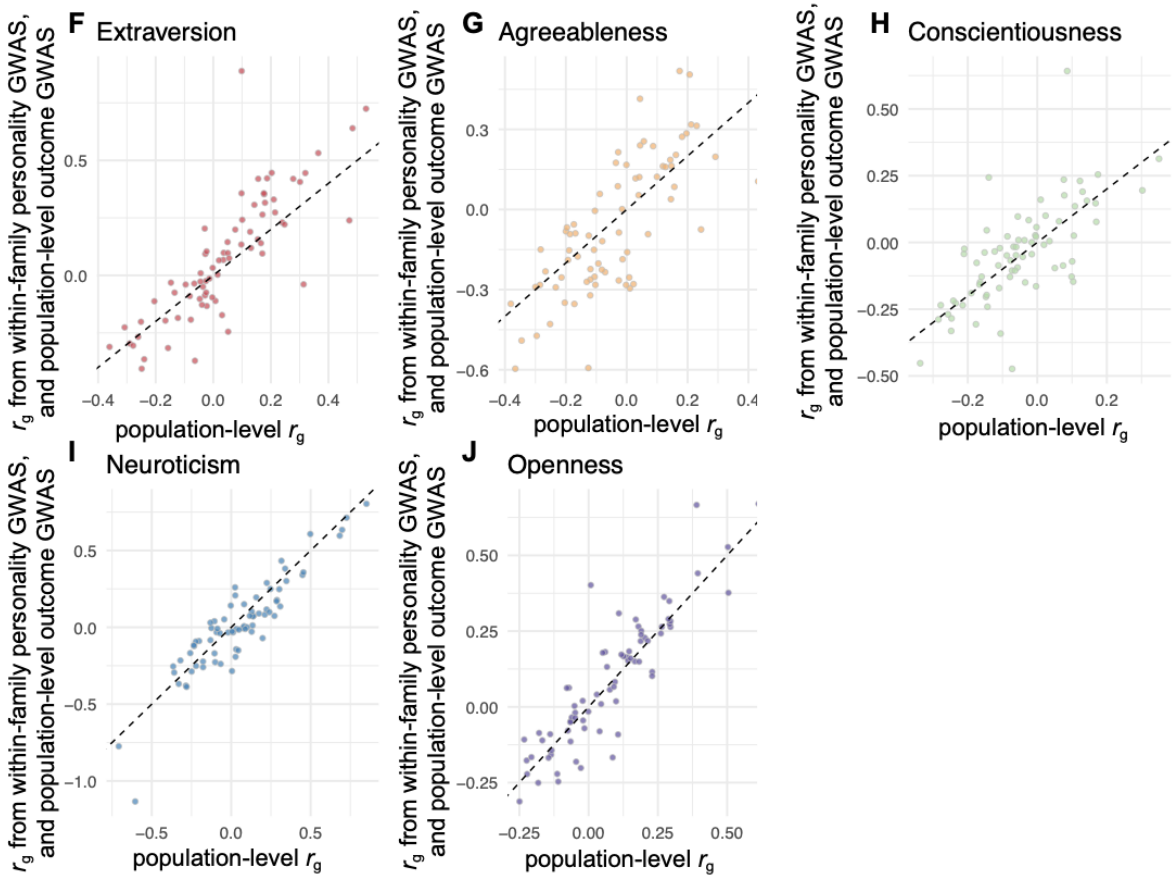
We would like to clarify here and in the text that we are not arguing that there is *no* confounding, but rather that confounding is *relatively minimal*. Across the totality of the evidence (within-family PGI analyses, within-family GWAS, population-level and within-family genetic correlations (see R2.14 below), and assortative mating analyses), results indicate very little familial confounding, but some individual parameters from within-family GWAS tests do provide some evidence for familial confounding. In the discussion section (p. 27), we summarize results of all tests that provided evidence for confounding: within-family versus population-level GWAS (where openness h2SNP decreased, and genetic correlations for agreeableness and neuroticism were less than one), parent-offspring PGI analyses (where there was a small but significant contribution to prediction from non-transmitted conscientiousness alleles), and assortative mating analyses (where effects were significant but very small in magnitude. We also mention that future research can continue bolstering sample sizes for within-family analyses to enable more precise quantification of this confounding.

2.14 It would be interesting to see the genetic correlations of the within-family GWAS with the traits in Figure 3, especially for neuroticism and agreeableness.

We agree – in revision, we have added genetic correlations between within-family Big Five GWAS data and external outcomes. We have done this two ways.

First, we estimated genetic correlations between within-family personality GWAS data and the full set of 86 population-level external phenotypes, which can be compared directly to the results in Figure 3. As we now note in the manuscript (p. 25), these estimates were similar to fully population-level genetic correlational estimates (mean r = .83 between the two sets of r_gs). For

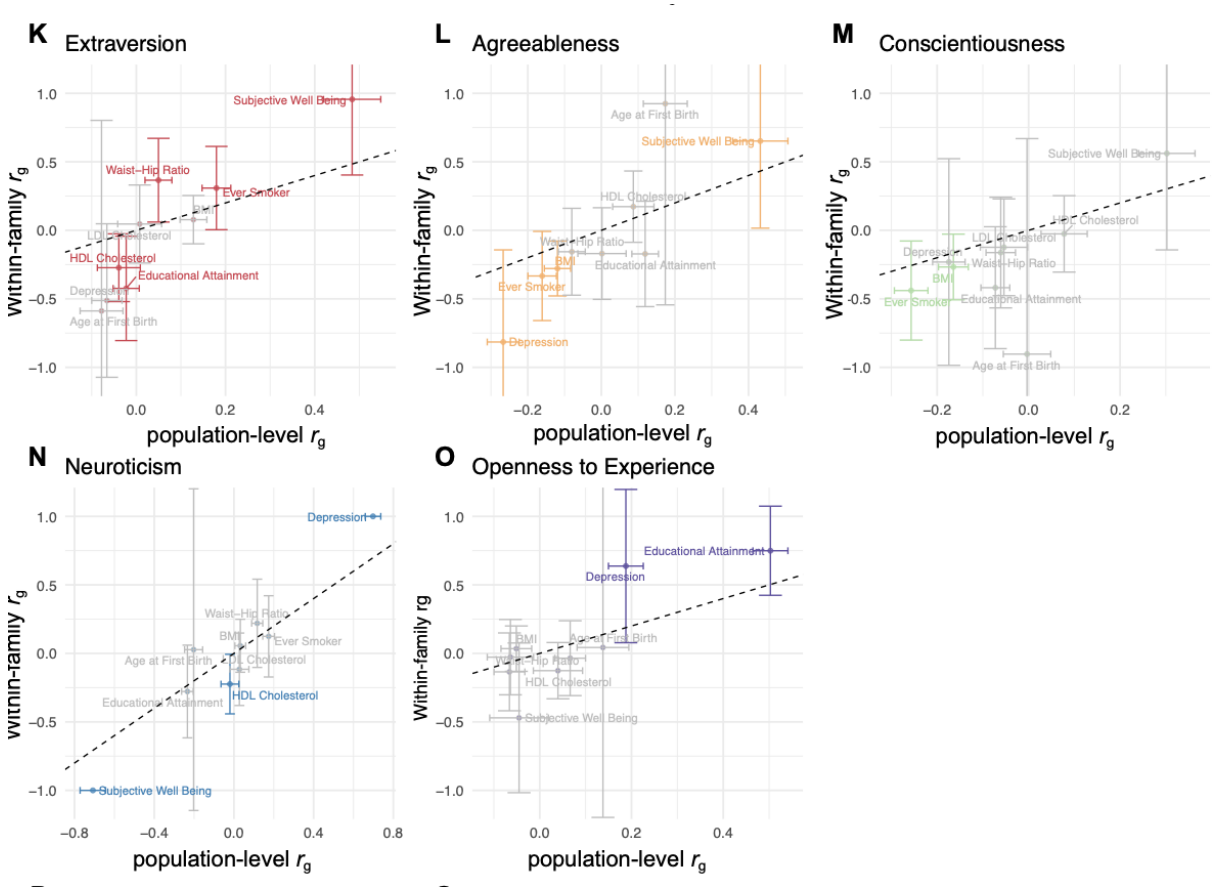
agreeableness (*r* = .75) and neuroticism (*r* = .91), the two specific traits mentioned by the reviewer as having the greatest confounding, correspondence was also similar between these two methods. Below, and in Extended Data Figure 10, we compare these estimates with a scatterplot and regression:



In this scatterplot, genetic correlations between within-family personality GWAS results and population-level external phenotype GWAS results are presented on the Y-axis, and fully population-level genetic correlations are presented on the X axis. The diagonal line (y=x) represents the expectation of equivalence and can be used visualize the difference between estimates; estimates below the line are stronger in population-level analyses, and estimates above the line are stronger in within-family analyses. In these plots the scatter closely surrounds this line of equivalence. In other words, population-level correlations do not appear to be appreciably directionally biased.

We also estimated a new set of *fully within-family* genetic correlations, correlating within-family Big Five GWAS summary data with within-family GWAS data from the 18 external outcomes from Howe and colleagues (2022); (Notably, because Tan and colleagues excluded the HUNT cohort from their publicly-available summary statistics, their GWAS summary data had a much smaller sample size than Howe and colleagues). These estimates were also quite similar to fully

population-level genetic correlations (mean r = .87 between the two sets of r_gs). Below, and in Extended Data Figure 10, we visually compare these estimates with a scatterplot:



In this scatterplot, fully within-family genetic correlations are presented on the Y-axis, and fully population-level genetic correlations are presented on the X axis. As fully within-family correlations are estimated with much less precision, we include 95% confidence intervals and plot significant within-family correlations in color. As with the figure above, there is notable scatter both above and below the diagonal dashed line of equivalence (y=x), indicating that population-level correlations do not appear to be appreciably directionally biased.

Overall, we appreciate this request from the reviewer, which has allowed us to further compare the genetic architecture of personality across population-level and within-family methodologies.

2.15. I think some discussion on the difference between the within-twin polygenic prediction and direct/indirect genetic effects analyses is missing. As it stands, it's unclear why these two different analyses are conducted because they both aim to get at the same result. The way they are presented as if they are complimentary is misleading. The authors do not mention that the PGI effects obtained using the within-twin analysis may be biased by indirect genetic effects from siblings. Imputing the genotypes of the parents and controlling for parental PGIs is a way to deal with this. I am not sure why the within-twin prediction has been done at all. It would be better to run the analysis

controlling for parental PGIs in two cohorts and meta-analyze the results. If the authors do have a reason why they think these two analyses are complimentary and we get something out of the within-twin prediction that we don't get out of controlling for parental PGIs, they should make it clearer in the text.

This comment specifically contrasts the analytic approach used for our within-twin analyses (in the TEDS and NTR cohorts) with that used for parent-offspring analyses (in the deCODE cohort) The reviewer correctly notes that while both approaches control for population stratification, dynastic effects, and assortative mating, the parent-offspring analyses additionally control for indirect genetic effects from siblings. Both are very robust controls over population-level PGI prediction, and in the absence of sibling effects on each other, the two approaches are nearly identical from an inferential standpoint.

We believe that the implementation of these two highly related, albeit technically different approaches is a feature rather than a limitation. Specifically, because we arrive at the same inference of little/no confounding in the within-twin comparisons (which may be subject to indirect effects between siblings) and the parent-offspring analyses (which are not biased by indirect effects between siblings), comparison of these effects across models provides evidence that there are likely minimal sibling effects in PGI prediction of personality. There are several additional reasons that we believe these within-twin analyses are worth retaining. First, many studies of other phenotypes have been conducted using within-twin/within-sibling PGI prediction (e.g., Howe et al., 2022; Selzam et al., 2019; Tanksley et al., 2025), and these analyses allow direct comparison to those published results. Second, conducting within-family analyses across three cohorts permits a much stronger test of replicability and generalizability than solely analyzing the deCODE cohort. There may be particulars of ascertainment or personality measurement in deCODE that restrict generalization of results, but obtaining near-identical effects across deCODE, TEDS, and NTR supports a much stronger inference that these effects are robust.

In the revised text, we now acknowledge the differences between these two sets of tests (p. 22). First, we present the within-twin PGI analyses, and then we present the parent-offspring analyses, which we now write are a complementary method that is unbiased by effects siblings may have on one another. We also now state that the similarity of parent-offspring results to the within-twin results suggests that there is negligible confounding from intra-sibling effects.

Selzam, S., Ritchie, S. J., Pingault, J. B., Reynolds, C. A., O'Reilly, P. F., & Plomin, R. (2019). Comparing within-and between-family polygenic score prediction. The American journal of human genetics, 105(2), 351-363.

Tanksley, P. T., Brislin, S. J., Wertz, J., de Vlaming, R., Courchesne-Krak, N. S., Mallard, T. T., ... & Harden, K. P. (2025). Do polygenic indices capture "direct" effects on child externalizing behavior problems? Within-family analyses in two longitudinal birth cohorts. Clinical Psychological Science, 13(2), 316-331.

2.16 The terminology in this section seems somewhat imprecise. The authors refer to the coefficient of the non-transmitted parental PGIs sometimes as "indirect effects" and sometimes "non-transmitted genetic effects". Given that this coefficient captures biases due to population stratification and assortative mating in addition to genetic effects from parents, calling it an "effect" is a misnomer. I realize this terminology has been used before in other papers but it becomes especially misleading when it gets used interchangeably with "genetic nurture" later on in this section (also in the Figure 4B title) when the finding that the non-transmitted coefficient is non-significant is interpreted as there is negligible genetic nurture. It is also not clear to me that this interpretation follows from the finding. Although not too likely, it is possible that genetic nurture operates in the reverse direction (e.g. growing up with neurotic parents makes the child less neurotic as a reaction?) and it cancels out the inflation in predictive power due to population stratification such that there appears to be no confounding. The fact that heritability estimated using the within-family GWAS appears to be close to the one estimated using the population GWAS is evidence against this to the extent that LDSC accounts for population stratification because then the difference is attributable to genetic nurture, but the authors need to synthesize these findings better in their discussion.

We agree with the reviewer's point that this parental PGI captures more than indirect genetic effects; we now term this coefficient "familial confounds", which we write captures effects of parental genetic variants not transmitted to the offspring, population stratification, and assortative mating (p. 22). Further, we have removed the term "genetic nurture" from the manuscript, as indeed these analyses that encompass multiple potential mechanistic pathways are not informative as to genetic nurture effects, specifically. Rather than speculate on the possibility and intergenerational direction of genetic nurture in the discussion, we restrict our treatment of these analyses to the role of confounding.

Replication

2.17. I'm having a hard time understanding why the replication was done with overlapping samples. It seems straightforward to exclude the Gupta et al.(Gupta et al., 2024) samples from the meta-analysis and try to replicate their findings in independent data. If that is not possible for whatever reason, it's possible to analytically subtract the Gupta et al. GWAS from the meta-analysis. Am I missing something? As it stands, this cannot be called a replication analysis and is not meaningful at all.

The reviewer's confusion is warranted, as we did not explain this well in the initial manuscript submission.

Unfortunately, it is not a straightforward task to exclude the Gupta and colleagues' samples from our meta-analysis and then examine replication, as Gupta and colleagues included summary statistics from Genomics of Personality Consortium (GPC) 1 and 2 analyses, which partially overlap with the current sample size in difficult-to-quantify ways. For example, GPC2 included 6,000 NTR participants, but due to continued data collection, our study includes 13,000 NTR

participants. Removing all partially overlapping samples from Gupta and colleagues to test replication would therefore lead to a severely underpowered test in our meta-analysis. Furthermore, our study was not designed with a held-out replication cohort to independently test replication of effect sizes: all data are included in the discovery cohort. We therefore test the replicability of the genetic signal in three alternative ways.

First, we index similarity of genetic effects by splitting our meta-analytic sample in two balanced sets of cohorts, submitting both split-halves to GWAS, and estimating the similarity of genetic effects between the two halves. As shown in Supplementary Table S14, we find a mean genetic correlation of .94 across these two halves, indicating strong replicability, and the average cross-trait intercept across these two halves is .01 (SE = .01), indicating that indeed the halves represent split sets of participants. We have rewritten the section on replication of effects in the main text to focus on the split-half replication (p. 9).

Second, we test the replicability of effects by comparing the new data added to Gupta and colleagues (the MVP cohort) to the other 45 cohorts included in our study. This ensures no sample overlap but is overly conservative, as it excludes GPC1/2 participants from Gupta and colleagues' analyses. In these tests, we find a mean genetic correlation of .80 between MVP and the remainder of our meta-analytic sample, indicating a medium amount of genetic overlap (Extended Data Figure 6; Table 1). (This correlation is also downwardly biased due to unique features of the MVP sample, such as their abbreviated personality questionnaire and the composition of the sample, which is entirely American military veterans). We note in the supplementary text that these tests provide a conservative test for replicability (p. 19).

A third method we use to examine replicability is testing the similarity of genetic signal between *population-level* GWAS effects and *within-family* GWAS effects in non-overlapping cohorts. These tests are especially stringent given that, in addition to testing replicability of effects across non-overlapping datasets, within-family analyses also account for confounds not accounted-for in the population-level analyses. We find an extremely close replication of effect sizes between the population-level and within-family data, among SNPs with $p < 1e\text{-}5$ in the population-level data. Specifically, within-family betas predict population-level betas as strongly as population-level betas predict within-family betas (mean ratio of replication is 95%). We now add language to the results section of the main text that situates this analysis as a test of replication (p. 26).

These three tests indicate high replicability of genetic effects on personality, mitigating the impossibility of directly testing the replicability of Gupta and colleagues' effects. Importantly, across these tests, we focus on the similarity in genetic signal across samples (e.g., using effect sizes and genetic correlation), rather than tallying the number of overlapping lead SNPs across samples. This simple test of replication is not possible given complicated and inseparable participant overlap between the two GWASes, and, due to the winner's curse and attenuated power when splitting data for replication, we would expect poor replication of tallied lead SNPs due solely to statistical reasons (e.g., conditioning replication on the initial identification of a significant effect), even in the presence of perfectly congruent genetic signal.

Following these tests, we have deleted our comparison between our full meta-analytic sample and the full Gupta and colleagues' meta-analytic sample, which we acknowledge is a suboptimal method for indexing replicability of effects.

Heritability differences across groups/ancestries

2.18. The analyses looking at heterogeneity in heritability estimates across different groups of samples is interesting, but I feel like the results are somewhat glossed over in the main text. The authors discuss only mean genetic correlations across the big-5 between groups, which do not reflect the interesting differences at the trait level. For example, the genetic correlation between the old and young for neuroticism is 0.33, which is quite striking. Moreover, even if we just look at mean rg's, the conclusion that I would draw from a mean rg of 0.8 across age groups would not be that genetic effects are highly similar. An rg of 0.8 for the same trait indicates quite some heterogenetiy. I think these findings need more emphasis for this analysis to be informative.

We agree that an *r*g of .80 indicates some heterogeneity in associations (especially given how genetic correlations can frequently be estimated at 1.00), and we do not want to communicate that the signal is identical across groups. We now clarify this in the main text, writing that correlations are "highly consistent but not identical across groups" (abstract) and that findings provide some evidence for heterogeneity in effects across each set of subgroups (p. 10)

Unfortunately, we do not have available space in the main text to discuss specific traits for which cross-group genetic correlations are less than unity, and so we instead provide much more information as requested by the reviewer in an expanded section of the supplementary text (section 3.4). Specifically, we now discuss reasons why agreeableness may be variable across cultures, expand considerably on the age group genetic correlations, and present three instances where genetic architectures were significantly differentiated across groups (r_g over 2 SE < 1.00): neuroticism across ages, conscientiousness across perspectives, and cross-trait agreeableness-neuroticism correlations across veteran status. We hope that this balances the reviewer's request for further discussion about cross-group correlations without venturing into overly speculative territory; indeed, many of these cross-group correlations with lower point estimates have wider confidence intervals (and so we now present standard errors for all the estimates we discuss to make this imprecision clear to readers).

2.19 As an explanation for the heritability difference between EUR and AFR GWAS, have the authors considered the set of SNPs used to estimate heritability? Are these analyses all based on HapMap3 SNPs? Might using a different/larger set of SNPs lead to higher heritability in AFR?

These analyses are indeed all based on HapMap3 SNPs, as we now make clearer in the added online-only methods section. As the reviewer suggests, increasing the set of common SNPs would potentially increase h^2_{SNP} among AFR participants if there are common SNPs that aren't well tagged by HapMap3; this might be the case among AFR participants due to lesser LD in this superpopulation.

To test this, we used POPCORN analyses, which incorporate a much larger set of SNPs (around 3.5 million) than the standard HapMap3 set (which incorporates around 1.3 million SNPs). In these POPCORN analyses, which we had previously run (Supplementary Table S18), all POPCORN h^2_{SNP} estimates were slightly smaller than their LDSC counterparts. (Note the smaller standard errors for POPCORN, as estimates are based on a larger number of SNPs). We now have included an additional Supplementary Table to describe this comparison (Supplementary Table S20), which we reproduce below:

	POPCORN			LDSC		
Trait	N SNPs	SNP h2	SE	N SNPs	SNP h2	SE
Extraversion	3387335	.025	.012	638279	.042	.019
Agreeableness	3365461	.003	.012	628876	.031	.018
Conscientiousness	3360991	.016	.013	627436	.017	.018
Neuroticism	3895607	.043	.013	801199	.051	.021
Openness	3404822	.008	.011	634483	.031	.018

This finding overall provides evidence that lower h^2_{SNP} in AFR participants is not explained by HapMap3 SNPs in particular, which we now mention in the supplementary text (p. 14); we retain our original speculation that this relatively lower h^2_{SNP} may be due to differences in ancestry, and thus LD patterning, between GWAS participants and the LD reference sample.

Prediction

2.20 I'm surprised by the predictive power of the PGI for neuroticism. The PGI from the second release of Polygenic Index Repository(Alemu et al., 2025) has a predictive power of 3.31% in HRS, and it is based only on UKB, 23andMe(Lo et al., 2016) and GPC(de Moor et al., 2015) with a sample size less than half of the current one. I would expect a much higher predictive power from this PGI. One difference between the two is the set of SNPs (HapMap3 vs ~2.9million SNPs(Lloyd-Jones et al., 2019)) but I don't think that's enough of an explanation. Have the authors looked into why this is the case?

We were also initially surprised by this difference across studies. First, as mentioned by the reviewer, one clear point of difference between these two estimates is in the set of SNPs used to generate weights. To test whether this explains the differences, we re-estimated PGI weights using the set of 2.9 million SNPs included in Alemu et al., 2025, and we compared PGI prediction using these weights to our initial models, which used the set of 1.3 million HapMap3 SNPs in weight generation. We found no improvement in prediction when expanding the set of SNPs included in PGI weights, in either the GSOEP-IS cohort or the HRS cohort. Specifically, comparing our previously-obtained estimates to these estimates, we find that the average prediction was nearly equivalent across methods, with the 1.3 million SNPs actually slightly outperforming the 2.9 million SNPs on average.

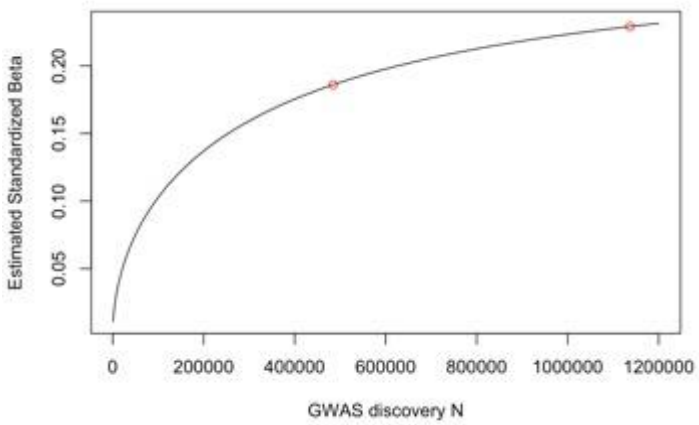
Predictive *B* of PGIs across two cohorts and two methods:

Trait	HRS (1.3m)	HRS (2.9m)	GSOEP-IS (1.3m)	GSOEP-IS (2.9m)
Extraversion	.229	.208	.190	.172
Agreeableness	.109	.092	.070	.060
Conscientiousness	.180	.147	.151	.158
Neuroticism	.182	.179	.176	.181
Openness	.170	.175	.139	.149
Mean	.174	.160	.145	.144

We present these new results in the supplement (pp. 26-27), as well as Supplementary Table S35, where we also provide a potential explanation for the lack of increase in prediction. Specifically, in these new analyses, the number of SNPs estimated to have effects on the personality trait did not increase in proportion to the number of SNPs included in the generation of PGI weights. For example, in the GSOEP-IS cohort, 15% of SNPs were estimated to have an effect on extraversion in the original set of SNPs, whereas only 8% were estimated to have an effect in the expanded set (using the same SBayesR priors on SNP effect distribution). In other words, the HapMap3 SNPs may constitute a set of common genetic variants that are particularly useful (compared to the larger set of 2.9 million variants) for predicting the Big Five personality traits. This lack of improvement may in part be due to imputation quality; HapMap3 SNPs are typically very well-measured (mean INFO score for 1.3m set of openness SBayesR SNPs = .97) in comparison to the expanded reference (mean INFO score for 2.9m set of openness SBayesR SNPs = .92).

The second explanation for differences in predictivity between this sample and Alemu and colleagues is that there are differences in sample composition between the two studies. Indeed, the three cohorts used by Alemu and colleagues (2025) to generate neuroticism PGIs were among those with the highest h^2_{SNP} in our GWAS, leading to an overall highly heritable estimate for neuroticism: 12% (driven by estimates of 11.6% in 23&me and 11.4-12.2% in their split UK Biobank cohorts). Across the 45 neuroticism cohorts we include, h^2_{SNP} is 8.5%, incorporating many cohorts with lower heritability: these cohorts likely do not contribute as much signal-per-participant as the three used by Alemu and colleagues.

The observed prediction of neuroticism in HRS by Alemu and colleagues is $\beta = .182$, and our observed prediction in this sample, with a smaller estimated h^2_{SNP} but a larger sample size, is $\beta = .181$. Using the equations provided by Alemu and colleagues to estimate PGI predictivity, we estimated predictivity if the sample size in Alemu and colleagues was increased to our sample size, holding heritability and polygenicity constant from that study. Under those conditions, expected $\beta = .229$. We visualize this in the figure below, which highlights the decreasing gains in prediction with increased GWAS sample size. Indeed, further increases in sample size would not be expected to increase predictivity much beyond the value in Alemu and colleagues:



Estimated PGI prediction of neuroticism in the HRS cohort. The curve depicts the estimated association between PGI sample size (X-axis) and standardized prediction (Y-axis), holding heritability and polygenicity constant. The red dots correspond to Alemu and colleagues (2025) at their observed sample size (left) and at the sample size of our GWAS study (right). Due to space limitations, this figure is not presented in the re-submitted manuscript but rather is solely included in response to this reviewer comment.

Overall, in our opinion, the difference between the expected value in Alemu and colleagues' data, extrapolated to our sample size (β = .229) and our observed value (β = .181) is not especially large, and can be attributed primarily to the asymptotic gains in prediction that come with increases in sample size of discovery GWAS used to derive PGI weights, especially as the samples added in our GWAS tended to have lower h^2_{SNP} and thus contributed less additional signal per participant. We now describe this comparison of EUR HRS neuroticism prediction in the supplementary text (p. 27). In the main text, we now refer to the predictivity of the neuroticism PGI estimated by Alemu and colleagues when contextualizing our PGI results.

2.21. The authors have used the EUR GWAS weights to make PGIs for the AFR samples in HRS and Add Health. I am curious why they haven't used the trans-ancestry meta-analysis. Is it because it would be complicated to obtain LD estimates reflecting the ancestry composition of the combined sample to use with SBayesR? It would still be informative to see if there is an improvement in predictive power for AFR samples using a method like PRS-CSx(Zheng et al., 2024) to incorporate both EUR and AFR meta-analyses.

We appreciate the reviewer's request for applying transancestral GWAS estimates to improve PGI prediction. As requested, we applied PRS-CSx (Ruan et al., 2022) to the EUR and AFR GWAS discovery samples to test PGI prediction in AFR participants in the HRS cohort. We found that PGIs constructed using PRS-CSx did not improve prediction: the mean prediction β was .046, compared to β = .051 in analyses predicting using EUR-derived PGI weights (**Supplementary Table S35**). This lack of improvement when using PRS-CSx may be attributable to the relative ancestral composition of our GWAS sample. As we write on p. 27 of

the supplement Only 5% of participants had African-like genomes, leading to minimal gains in PGI prediction when incorporating their data into weight construction.

2.22. Are the PGI associations in Panel B of Figure 2, or the genetic correlations in Figure 3 are corrected for multiple testing? Apologies if this information is reported, I couldn't find it anywhere.

We use a *p*-value of .01, which corrects within test dividing .05/5 for the Big Five. We did not correct *p*-values otherwise. We note this in the online methods section and in each of the figure captions.

2.23. In the residential analyses, the authors control for 5 PCs. I was surprised by this choice; it seems very few PCs especially for an analysis like this. Does controlling for more PCs change the results?

We appreciate the reviewer's request to re-run analyses with additional PCs. Based on this feedback, we planned to conduct these analyses controlling for 10PCs. Unfortunately, the transition of UK Biobank data to the Remote Analysis Platform (RAP) precluded these analyses, as geographic data are not currently available on the RAP.

To alleviate the reviewer's concern, though we only control for 5 PCs in these analyses, we also include another, especially stringent, layer of control for potential confounding: within-sibling analyses. Specifically, the results we present in the main text (p. 19) comprise associations between personality PGI and residential mobility controlling for sex, age, place of birth, and 5PCs, filtered to results that are directionally consistent in within-sibling analyses. This filtering reduces potential effects of population stratification considerably. (We note that because the power of this analysis is restricted to siblings that vary in genotype and residential movement, we use within-sibling findings as a filter on population-level results, rather than as our primary analysis, which would be underpowered).

2.24 In SI section 6.2 about the parent-offspring PGI analysis in deCODE, the authors state there were N=34506 probands with at least one genotyped parent. Were the missing parents imputed here?

The missing parents were not imputed – this can be seen in the supplementary tables, where mother and father N are presented separately.

Assortative mating
2.25 I couldn't see any information about how the spouse pairs in UKB were identified.

We apologize for this omission. On p. 35 of the supplement, we now write:

"Because spouse information is not directly available in UK Biobank, spouse pairs were inferred using criteria closely following previous studies (Howe et al., 2019; Rawlik, Canela-Xandri & Tenesa, 2019; Robinson et al., 2017). Individuals were classified as spouses if they shared

residential coordinates (100m resolution), reported living with their spouse, and reported the same number of household occupants and vehicles, the same length of residence at the current address, the same rental status and type of accommodation, the same assessment centre, the same household income, and the same number of smokers in the household, while differing in the age of at least one of their parents. Pairs genetically related up to the second degree were excluded. In the case of multiple matches, the pair sharing the most coordinates across time points was selected, with unresolved cases removed. Restricting the sample to pairs in which both individuals were of European-like ancestry resulted in 33,972 spouse pairs."

Editorial

2.26I find it a bit confusing that the biological follow-up paragraph on page 10 cuts in between the discussion on heritability. I think it would be better if this was a section on its own that comes after "Characterizing Common-Variant Heritability".

We agree that the manuscript reads better in that order. We have moved this section as suggested (pp. 11-12).

Referee #2 (Remarks on code availability):

I cannot seem to access the code using the OSF link above. I get an error message that says "Failed to fetch file metadata from WaterButler. Received response"

We have now made all code available at the OSF link:
https://osf.io/hgnsm/overview?view_only=faa526bbeabc4615b5d6c0fbc52e10ba

Referee #3 (Remarks to the Author):

Review of "Robust inference and widespread genetic correlates from a large-scale genetic association study of human personality"

This manuscript represents an extraordinary scientific contribution, offering the most comprehensive analysis to date of the genetic architecture of the Big Five personality traits. The authors leverage over one million participants across 46 cohorts to generate a robust set of GWAS findings, enriched by within-family analyses, polygenic score prediction, and Mendelian Randomization (MR). Their work is remarkable not only for its scale but also for the care taken to replicate findings across subpopulations and to explore the genetic correlates of personality across a wide range of outcomes and to test causal hypotheses using MR. This is a landmark study and is, in my view, appropriate for publication in Nature pending revisions addressing the points below.

We appreciate Reviewer 3's enthusiasm!

Major Comments:

3.1. A central claim of the manuscript is that genetic associations with personality are largely free from shared environmental confounding. While the within-family and parent-offspring results are compelling, sample sizes are modest (N ≈ 30k–50k; see p. 17–18), and the ability to detect indirect genetic effects (e.g., from parental genotype or shared environment) is limited. Null findings here should be interpreted in light of modest statistical power. The authors should focus on the confidence intervals of the indirect effect estimates (or the ratio of direct/(indirect+direct)) and revise their interpretations accordingly. This concern arises in multiple places, including the Abstract, pp. 6 and 18–19.

We appreciate the reviewer's request to add inferential specificity to the manuscript. We have made changes throughout the manuscript and in the requested locations to address these requests.

First, to increase power for comparing population-level versus within-sibling PGI prediction, we have meta-analyzed prediction estimates across NTR and TEDS to produce a single pair of population-level versus within-family estimates for each Big Five trait, which we reproduce below and in Supplementary Table S36. As we describe in the main text, these meta-analyzed estimates are on average extremely similar to one another (.165 vs. .158), a ratio of 96%. Because the SEs of the population GWAS betas are so small, their estimates are virtually error-free, and so the major point of inferential comparison here concerns the confidence intervals of the within-family estimates (which are wider). As can be seen below, the population-level and within-family prediction estimates are extremely similar, for each trait.

Trait	Population-level B	SE	Within-family B	SE
Ext	.173	.008	.201	.020
Agr	.117	.007	.125	.019
Con	.169	.008	.150	.020
Neu	.169	.008	.141	.020
Ope	.196	.008	.173	.020
Mean:	**.165**		**.158**	

Furthermore, as we show visually in Figure 4 Panel A, this lack of difference between population-level and within-family PGIs replicates across both the TEDS and NTR cohorts (albeit with wider confidence intervals for individual cohorts).

Secondly, to more accurately convey the precision of parent-offspring analyses, we have added confidence intervals to the analyses of indirect effects in Figure 4 Panel B, as well as in our presentation of these models in the main text. These are extremely highly-powered tests on 34,506 family dyads/triads. This test is powered to rule out indirect effects (which we now term

familial confounds, see R2.16) greater than B ~.015 at $p < .05$. Demonstrating this high power, we find significant familial confounding for conscientiousness in parent-offspring analyses, even though this confound is only 8% of the magnitude of the total PGI effect. We mention this on p. 22 of the main manuscript and in the discussion section (p. 27).

Overall, adding confidence intervals to these two PGI analyses more clearly delineates that these tests are highly powered to identify any familial confounding.

We agree with the reviewer that there is one major point of inferential uncertainty: in comparisons between population-level and within-family GWAS, which have much less inferential precision. This point was also raised in R2.13. We agree with the reviewer here that caution is necessary to appropriately contextualize within-family GWAS results. In these tests, specifically, we use $p < .05$ as the criterion for identifying differences between population-level and within-family estimates, to minimize the type-II error rate of missing a true difference between the model estimates. In the main text, we now provide confidence intervals for all point estimates comparing population-level GWAS estimates to within-family GWAS estimates (p. 25), and in the discussion section we mention that some of these tests provided evidence for the presence of familial confounding (p. 27). We have also added notes in the results (p. 25) and discussion (p. 27) regarding the need for future collection of larger within-family GWAS samples to more precisely compare effects.

We hope that these changes have satisfied the reviewer's request for full disclosure about inferential specificity, which is high in within-family PGI analyses and much lower in within-family GWAS analyses.

3.2 Equation 14 in the Supplement uses a family fixed effects model but does not include both individual deviations from the family PGI mean and the family mean PGI itself as separate predictors. The approach used by the authors estimates only the within-family effect and precludes estimation of the between-family component. The comparison to "matched population-level GWAS in cohorts that contributed to within-family GWAS" (p. 19) would be more informative if the same exact individuals were used to estimate both within- and between-family effects (i.e., including both within- and between-family PGI predictors). If this was not done, please justify this decision and consider re-estimating the model accordingly.

We conducted two regressions in each cohort: one regression produces a population-level prediction estimate, defined as the prediction of phenotypic variation in personality traits from variation in genotype in the population. The other produces a within-family prediction estimate, defined as the prediction of phenotypic variation in personality traits from variation in genotype, controlling for the genotypes of one's family members. We then compared the two prediction estimates with each other. In the added online methods section, we explicitly define these population-level and within-family effects. We agree with the reviewer that one way to compute population and within-family estimates would be to fit a single model that includes terms for the within- and between-family components, and then sum the within and between components to

produce a total population effect. However, the approach that we take computes the population effect directly in a separate model and has the benefit of directly producing a standard error for the population effect. It also has the benefit of more straightforwardly corresponding to the approach that is taken in a standard population-level analysis. Both approaches are expected to produce extremely similar results.

In the supplement (p. 39), we now clarify that, in the TEDS, population-level models are estimated using all DZ and MZ twin pairs, and within-family models are estimating using just DZ twin pairs (as MZ twin pairs have no variation in genotype). In the NTR, population-level models are estimating using all DZ and MZ twin pairs, as well as parents and other relatives of twins, and within-family models are estimated using DZ twin pairs. The reviewer requests that we consider estimating within- and between-family effects using the exact same individuals. As we now write on p. 40 of the supplement, *"As a sensitivity test, we also estimated population-level PGI prediction in NTR among only the DZ twins who contributed to within-family PGI prediction analyses (i.e., we pruned other NTR family members, such as MZ twins, from the population-level prediction sample). Population-level prediction in this restricted sample was nearly identical in magnitude to prediction in the full sample (mean β = .174 versus .172 in the full sample;* ***Supplementary Table S35****). Thus, this analytic decision does not impact conclusions of population-level versus within-family PGI comparisons. As PGI prediction in this restricted sample had larger standard errors (mean of .014 versus .009), we focus on population-level analyses with the full NTR sample in the main text."*

We also point the reviewer to within-family and population-level prediction in the deCODE cohort, where we estimated within-family PGI prediction by directly controlling for one's parents' genotypes, which provides a complementary estimate of within-family effects. These estimates and the twin-pair within-family regressions lead to the same inference of minimal confounding, providing evidence that the current approach to twin-pair within-family PGI prediction is appropriate.

3.3 The inferred genetic correlations between spouses (Figure 2C; lines 310–326) are modest, but readers will want to know how these compare to phenotypic spousal correlations and may not understand that spousal rg's are expected to be lower than spousal phenotypic correlations. Given the non-unity heritability of personality, the rg between spouses should be expected to be lower than the rp under phenotypic assortment. The authors could provide inferred rp estimates from rg values (by correcting for the heritability assuming phenotypic homogamy) and compare them to known phenotypic spousal correlations (e.g., Horwitz et al., 2022). This would clarify how the observed genetic rg's relate to reported rp's.

We appreciate the suggestion for comparison of genetic AM estimates with phenotypic AM estimates. To facilitate this comparison, we now present observed phenotypic correlations between deCODE mothers and fathers' personality traits in Extended Data Figure 8, which can be compared to meta-analytic genetic correlations and directly to genetic correlations in the deCODE sample (Supplementary Table S40). As the reviewer predicts, the r_g between spouses

is indeed lower than the r_p. Importantly, phenotypic AM estimates are also quite low compared to published estimates for other traits (r = .07-.12, and openness to experience r = .18). As in past phenotypic research, openness to experience has a greater AM estimate than the other Big Five traits (Horwitz et al., 2023).

As requested by the reviewer, we then used these estimates to estimate whether genetic AM estimates are higher than would be expected based on these phenotypic AM estimates, by multiplying sqrt(h^2_{SNP}) * spousal rg * sqrt(h^2_{SNP}) and comparing this value to phenotypic AM. We did this two ways: first by comparing meta-analytic values for each of these traits (using meta-analytic phenotypic AM estimates from Horwitz et al., 2023), and second, using the values specifically in the deCODE cohort, for which we have estimates for each of phenotypic AM, genetic AM, and h2SNP. We present this table below and in the supplement (Table S41).

Sample	Trait	Phenotypic AM	Genetic AM	H2SNP	Inferred genetic AM from pheno AM * H2SNP	Difference (Genetic AM - inferred Genetic AM)
Meta-analytic	Ext	0.08†	.027	.093	.007	0.019
	Agr	0.11†	.017	.048	.005	0.012
	Con	0.16†	.011	.071	.011	0.000
	Neu	0.1†	.018	.083	.008	0.010
	Ope	0.21†	.045	.074	.016	0.029
deCODE spouses	Ext	.072	.024	.124	.009	0.015
	Agr	.120	.032	.097	.012	0.020
	Con	.093	.008	.117	.011	-0.003
	Neu	.095	.030	.091	.009	0.021
	Ope	.180	.040	.148	.027	0.013

Note: † = estimate comes from Horwitz et al. (2023) meta-analysis

We now present these findings in the main text (p. 20), where we write that *"When we compared genetic with phenotypic assortative mating, the former were slightly smaller than expected, based on h^2_{SNP} and phenotypic spousal correlations, at the meta-analytic level and in deCODE participants specifically. This indicates that a portion of the (minimal) observed assortative mating on personality traits might be explained by assortment on other heritable phenotypes that are genetically correlated with personality (e.g., some observed assortment on openness may be attributable to education (Okbay et al., 2022)). This was not the case for*

conscientiousness, for which phenotypic estimates of assortative mating were fully accounted for."

3.4 The MR analysis is underreported in the main text. Two short paragraphs on pp. 16–17 do not do justice to the depth of the results. I recommend including a dedicated figure and/or expanded main-text table listing the significant MR results and the method used (e.g., CAUSE, weighted median). In addition, please report whether corrections for multiple testing were applied.

We agree with the reviewer that the MR analysis is important, but we are precluded from dedicating this section more than two paragraphs in the main text (pp. 20-21) by manuscript length limitations. To highlight these results as strongly as possible, we have moved the table summarizing MR results from the supplemental tables excel spreadsheet to Extended Data Figure 9, alongside the path diagram that describes the MR method. This table lists the significant MR results across all three methods used, so that readers can see the estimates for each and get a sense of their triangulation. In the supplement, where there is no space limitation, we now describe MR and our inferential criteria in more depth (section 5.5).

As we write in the methods section and in the figure caption for this Extended Data Figure, we only interpret results as putatively causal if they are directionally consistent across estimators (each of which relies on different assumptions) and the 95% confidence interval of the effect estimate excludes zero across two or three of the estimators used. We believe that these criteria, in addition to our use of negative controls to identify potentially spurious effects (birth weight, number of sisters, and number of brothers, which are three outcomes that cannot be caused by one's personality), provide a stringent method for identifying significant effects and discounting false positives.

3.5 The decision to exclude SNPs only when INFO < 0.40 is unusually permissive. The authors note (p. 6 and supplement) that more stringent thresholds were applied elsewhere but do not clarify when and where. Please provide more detailed justification and clearly indicate the imputation reference panels and procedures used across cohorts. This should be included in the main text or, at minimum, expanded in the supplement.

In review, we have added an Online Methods section to comport with Nature's formatting guidelines. Across this section, we now clearly delineate the INFO thresholds we use for each analysis. To respond directly to the reviewer, INFO < .40 is the minimal initial threshold applied at the level of the cohort-specific QC, and more restrictive INFO thresholds were used in downstream analyses. For instance, we use INFO ≥ .90 for all population-level LDSC analyses (as suggested by Bulik-Sullivan et al., 2015), we use the extremely stringent INFO ≥ .99 for all within-family analyses (as suggested by Tan et al., 2025), and we use INFO ≥ .80 for MAGMA analyses. We also clarify these thresholds together in the supplementary text (p.6). Finally, as requested by Referee #2, we have included INFO for genome-wide significant lead SNPs

(Supplementary Tables 6-10). As expected, nearly all lead SNPs have very high INFO scores (e.g., the mean INFO for extraversion lead SNPs is .97).

One potential concern is that including SNPs with a low INFO threshold might downwardly bias PGI predictivity, as there is not explicitly a more strict INFO threshold for PGI analyses. Reassuringly, however, PGIs were estimated using HapMap3 SNPs, which are generally directly measured or very accurately imputed. The mean INFO score among EUR Neuroticism HM3 SNPs, for example, is .98, which we believe alleviates potential concerns about PGI attenuation due to imputation imprecision. We now note this in the supplementary text where we describe PGI prediction (pp. 25-26).

3.6 Some rg results in Figure 3 are counterintuitive. For example, openness shows positive genetic correlations with ASD and OCD, and negative correlations with COVID-19 infection and sports participation. Conscientiousness negatively impacts educational attainment in MR (p. 16). The authors should help the reader interpret these surprising findings, perhaps in a new paragraph in the Discussion.

We wish that we had an entire manuscript's worth of space to discuss these interesting and sometimes counterintuitive genetic correlation results, which can inform theory and future research. Given the limited space to properly contextualize the entire set of genetic correlations, we have now added an extensive section to the supplementary text (section 5.1, pp. 28-30) where we describe genetic correlations between the Big Five and each set of outcomes, focused on outcomes that are counterintuitive or that inform of personality theory. We welcome the reviewer's feedback on our interpretation of these results.

Specific Comments:
3.7 Abstract: Final sentence is awkward—reword "The genetic architecture of personality… is fundamental to being a human."

We have now revised this sentence.

3.8 Page 4: Define what is meant by "assemble" in "assemble data from 46 cohorts." Clarify cohort selection, harmonization, and QC steps.

We have replaced the word "assemble" with "meta-analyze" and we now clarify cohort selection, harmonization, and QC steps in the Online Methods section that we have added in revision. We also provide more information about harmonization of rsID labels across contributing cohorts on pp. 7-8 of the supplement.

3.9 Page 4 (WF GWAS): When stating "little to no shared environmental confounding," provide confidence intervals for estimates of indirect effects or IGE magnitudes.

Unfortunately, we cannot provide a single point estimate or confidence interval for indirect effects here, as this brief phrase in the introduction serves to summarize across a series of tests

that estimate different values (e.g., the ratio of direct effects to indirect confounds, the equality of h^2_{SNP} across population-level and within-family GWAS, etc.). We now state, more clearly, that *"a suite of within-family analyses indicate that genetic effects on personality traits are only minimally confounded by population stratification, dynastic effects, and assortative mating."*

In the within-family portion of the results section, we added confidence intervals to all parameter estimates to make measurement precision maximally transparent.

3.10 Page 5: Supplement the Roberts & Yoon (2022) citation with a more foundational reference (e.g., McCrae & John, 1992).

We have added this reference as requested.

3.11 Page 5 ("predict"): Clarify that predictive relationships of Big Five traits with outcomes are phenotypic, not genetic.

Thank you. We have clarified this.

3.12 Page 6 ("k = 46"): Define "k" explicitly as the number of cohorts per meta-analysis.

We have now defined this.

3.13 Page 6: Fix typo: "across social relevant" -> "across socially relevant"

We have fixed this typo.

3.14 Table 1: Excellent and informative table. Consider adding average cross-cohort LDSC rg values (possibly weighted by the geometric mean of the pairs of cohort sizes). This can help in interpreting RE-meta vs. fixed effect meta-analysis results.

Following the reviewer's suggestion, we estimated average cross-cohort LDSC r_g values for each trait, across a total of 1,781 unique pairings of cohorts. We then meta-analyzed these correlations for each trait, with each estimate weighted by inverse variance to properly account for differences in sample size across cohort pairs. Overall, these results paint a similar picture to the general cross-group genetic correlation results, where cross-cohort correlations are lower for agreeableness than the other Big Five:

Extraversion weighted mean r_g = .71
Agreeableness weighted mean r_g = .45
Conscientiousness weighted mean r_g = .72
Neuroticism weighted mean r_g = .73
Openness to experience weighted mean r_g = .67.

We have added these values to Table 1. We also note in the supplement that these correlations

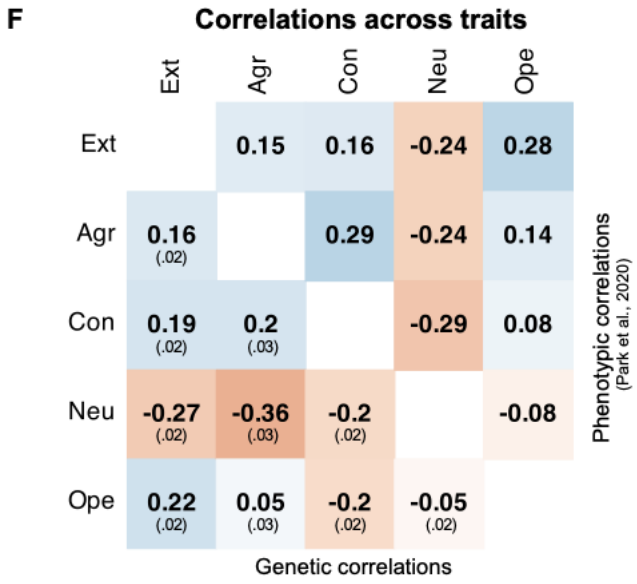
across cohorts are weaker than correlations between grouping variables (e.g., across measurement instrument), which demonstrates that aggregation of signal within groups increases generalizability of genetic effects.

3.15 Page 7: Explain why conscientiousness shows lower disattenuated h²_SNP relative to other traits.

We apologize for the confusion in our previous presentation of estimates; conscientiousness does not show lower disattenuated h^2_{SNP} values. The meta-analytic h^2_{SNP} estimate is 7.1%, h^2_{SNP} disattenuated for typical measurement error is 9.6%, and h^2_{SNP} disattenuated for all measurement error is 11.2%. These increases in h^2_{SNP} with disattenuation are similar to what is seen for the other Big Five traits (Table 1).

3.16 Page 8: The claim that Big Five rg's roughly equal phenotypic correlations (Soto & John, 2017) could be tested formally via matrix comparison of rg vs. rp using maximum likelihood.

Unfortunately, phenotypic correlations among the Big Five were not available from each cohort that contributed to this meta-analysis, so we cannot directly compare genetic correlations to phenotypic correlations. However, we were able to compare the results of this study to meta-analytic intercorrelations among the Big Five estimated by Park and colleagues (2020), the largest study dedicated to this topic. We now include these estimates above the diagonal in Figure 1 panel F, which we have reproduced below. Estimates are highly similar across the two methods, which can be easily seen using in the correlational heatmap:



Given the extreme precision of these estimates, formal comparison of equality across methods would certainly show differences, even those of a very small magnitude (and perhaps due to

differences in participant ascertainment and measurement between meta-analytic samples). We therefore believe that presenting these matrices as is and allowing readers to see the holistic similarity between estimates is the ideal method of comparison.

3.17 Figure 1: Consider adding SNP h² estimates to diagonal of the rg matrix.

We currently present these h^2_{SNP} estimates in Table 1. In Table 1, we estimate this quantity multiple ways to display how estimates of h^2_{SNP} are sensitive to instrument unreliability and meta-analytic methodology. In the interest of parsimony, and to direct reader's eyes to the finding that heritability varies across estimation methods, we do not also present them on the diagonal of the r_g matrix in Figure 1.

3.18 Page 10 (MAGMA): Consider rerunning the gene-set analysis excluding neuroticism to examine its influence on shared signal.

Each of the MAGMA gene-set analyses was run independently of the others (i.e., there were five analyses for five traits), so neuroticism's signal was not incorporated into gene-set analyses of the other Big Five traits. We now clarify this in the online methods section.

3.19 Page 10, Negative Selection: I think that the evidence purported to be in support of negative selection (enrichment in protein truncating variant intolerant gene sets) is pretty tenuous. Perhaps this can be omitted, or perhaps other types of evidence can be brought to bear on this issue.

We agree that this evidence, while intriguing, is not definitive. As the reviewer suggests, we now omit this finding from the introduction. We now solely mention this finding briefly in the results.

3.20 Page 11: Address power limitations in the 20K other-rated Estonian Biobank subsample when interpreting rg = .84 with self-report.

In the main text, we now provide the mean standard error for all cross-group comparisons, so that readers can properly interpret the results in terms of their measurement precision. The EBB subsample mean cross-perspective r_g is .84, with a mean SE of .12 (p. 10).

To further contextualize results, as we describe in our response to Reviewer R2.18, we have added text to the supplementary materials (section 3.4) noting particular traits for which cross-group genetic correlations were less than 1. In the case of these cross-perspective comparisons, the correlation between self-reported and other-reported conscientiousness is r_g = .68, SE = .11; for other Big Five traits, cross-perspective genetic correlations were within 2 standard errors of 1, indicating nonsignificant deviation from unity.

3.21 Figure 2A: AFR PGIs: Report predictive performance when using AFR-derived PGIs (if available). Otherwise, acknowledge this omission.

Unfortunately, we did not have a large enough GWAS discovery sample of AFR participants to predict the Big Five using PGIs derived solely from AFR data.

To further expand AFR PGI analyses in revision, we conducted additional analyses using PRS-CSx (Zheng et al., 2024), which integrates AFR and EUR participant data into PGI analyses. However, these analyses did not improve prediction accuracy in AFR participants. For complete information about these new PRS-CSx analyses, see reviewer response R2.21.

3.22 Figure 2E: Convert effect size into a correlation or similar metric for easier interpretation.

Because associations between PGIs and outcomes are strongly affected by measurement error in the PGI, even standardized correlations between PGIs and residential mobility are very small and not informative. In other words, PGIs are useful for identifying the presence and sign of associations between a GWAS trait (such as personality) and external outcomes (such as residential mobility), but PGIs are not in this case appropriate for applied prediction of outcomes for specific individuals (e.g. in clinical or educational contexts).

For this reason, we test genetic associations using genetic correlations whenever possible and we have added effect sizes (r_gs) to our description of genetic correlation results across pp. 15-18. In the PGI section, we now explicitly state that the purpose of these PGI predictions is to substantiate connections between personality traits and life outcomes that have yet to be submitted to GWAS (p. 19). We also recommend that PGI associations should not be used for the prediction of outcomes for a specific individual (p. 19), and as such, we refrain from emphasizing the magnitude of prediction, but we provide labelled x axis ticks in Figure 2B to denote effect sizes, and we report exact estimates in Table S9. In Figure 2E we plot the z-statistic of effects, which communicate the direction of effect and the magnitude of evidence that it is nonzero. These show that PGIs are linked positively and negatively to certain characteristics of residential mobility across the lifespan, which is our analytic intent.

In the main text, we now better clarify the utility of these PGI-external outcome predictions, mentioning their theoretical utility and practical limitation. See response to reviewer R1.5 for additional details contextualizing PGI-outcome associations.

3.23 Lines 203–207: Clarify the “M_Beta” notation (clearer would be a bar of beta) and confirm that coefficients are standardized.

In revision, we have replaced these mean estimates with meta-analytic estimates, so that we now report the average standardized PGI predictivity across cohorts (Supplementary Table S36). We use the β symbol to depict this estimate, which we confirm is standardized.

3.24 Fig 3: Age-at-first-sex: Analyze potential sex interactions. This outcome may be differentially associated with personality by gender.

Although we appreciate the reviewer's hypothesis for gene-by-sex interactions in the prediction of age at first sex, we unfortunately are not able to test this possibility in the current meta-analytic GWAS. Individual autosomal GWAS contributed by each cohort were not sex-stratified, and so we cannot test moderation by sex. However, in revision we did conduct new analyses for X-chromosomal effects on Neuroticism using UK Biobank data, which indicated very strong cross-sex genetic correlations (*r*g = .96). See response to Reviewer 2.5 for full details.

3.25 Page 16 and Fig. 2: I understand how relatives are used in addition to spouses to estimate the signatures of assortative mating, but this will be confusing to most readers. Consider adding a few sentences here explaining this in better detail.

In revision, we have added an online-only methods section to the main text (p. 40), in which we now devote a paragraph to describe this assortative mating model in detail. We also now foreground the path diagram used to estimate this model in Extended Data Figure 7 Panel B, and we provide further detail in section 5.4 of the supplemental note. Finally, lavaan code to estimate this model is included in the online R code, which has now been made available.

3.26 Page 16 (MR): Report correction for multiple testing and list the full set of exposure-outcome pairs tested.

In the results section, where we introduce the MR methodology, we write that we considered associations reliable only if they were directionally consistent across all three estimators with 95% confidence/credibility intervals that excluded zero across at least two of the three (see Supplementary Text for complete information on MR methodology). Due to space constraints, we list the full set of exposure-outcome pairs tested in the supplementary text (where we include rationale for including each outcome) and provide all phenotypes in Supplementary Tables S42 and S43. We now summarize the significant trait-outcome pairings in Extended Data Figure 9, along with an MR model.

3.27 Page 19 (Figure legend): Clarify what "104%" and similar values mean. Specify units and estimation approach.

We now clarify that these values present the ratio of direct effect to confounding effect prediction estimates; as with other ratios, they are unitless. To provide further clarity on estimation approach, we now include 95% confidence intervals for confounding effect estimates in Figure 4 panel B, which properly communicate the precision of these findings.

3.28 Supp. p. 57: Add missing reference for Tucker-Drob (2017).

Thank you for reading in such detail!

3.29 Supp. Tables PDFs: Some PDFs failed to render; rely on Excel files for clarity.

We apologize for this error. We have ensured that the PDFs now render properly.

3.30 Data Availability: Clarify whether neuroticism sumstats will be available both with and without UKB. Currently ambiguous. Certainly hope it is both (and if not, justify).

Neuroticism data with UKB included and left out will be made publicly available upon publication (UKB participants only contributed neuroticism data and not data on the other Big Five traits). We now clarify this in the data availability statement.

Referee #4 (Remarks to the Author):

4.1 I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Thank you for your co-review!

Comments from AE:

Your manuscript, "Robust inference and widespread correlates from genetic associations with personality", has now been seen again by 4 referees. You will see from their comments below that while they find your work strengthened, some important points are still raised. We are interested in the possibility of publishing your study in Nature, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication. We therefore invite you to revise and resubmit your manuscript, taking into account the points raised by the reviewers. Please highlight all changes in the manuscript text file.

We thank you for the opportunity to respond to these remaining concerns, and we have revised the manuscript in response. Changes to the text are highlighted in yellow in the manuscript.

Beside the remaining comments made by the reviewer #3, we also want to highlight that you please must make sure that all code is made accessible. Like reviewer #2, we were unable to open several of the files (error message "You do not have permission to perform this action.").

We detail how we have addressed each of the remaining comments in the responses below. This includes performing several checks to ensure that all code deposited on OSF is accessible.

Referees' comments:
Referee #1 (Remarks to the Author):

I have no changes to recommend. The authors have done a spectacular job addressing my prior concerns.

Thank you!

Referee #2 (Remarks to the Author):

I sincerely thank the authors for their thorough and thoughtful revision. They have addressed all of my comments and added extensive additional analyses, greatly strengthening the manuscript. I have no further suggestions.

Thank you!

Referee #2 (Remarks on code availability):
I'm still having trouble accessing the code. The initial OSF page opens, but if I click anything on there, I get a "You Don't Have Permission To Perform This Action" error.

We have now confirmed that the files at the reviewer read-only link included with the paper, https://osf.io/hgnsm/overview?view_only=faa526bbeabc4615b5d6c0fbc52e10ba, are viewable by visitors who are logged in or out of OSF, visitors affiliated and unaffiliated with the project page, and visitors from both US and European ISPs.

Please note that due to a quirk of the OSF website, to view the wiki page for the project, one must fully enlarge the window and click “Wiki” on the left, rather than “read more” below the Wiki. Also please note that although the Code folder itself indicates a file size of 0 B, it is not empty. Clicking on the “Code” folder will open it and allow you to access files containing the code.

Referee #3 (Remarks to the Author):

This manuscript represents an extraordinary scientific contribution, offering the most comprehensive analysis to date of the genetic architecture of the Big Five personality traits. The authors leverage over one million participants across dozens of cohorts to generate a robust and highly replicable set of GWAS findings, complemented by within-family analyses, polygenic score prediction, and Mendelian Randomization (MR). The scale and scope of the work are exceptional, and the manuscript has been substantially improved in response to prior reviews. I am satisfied with the majority of the authors' revisions, although some interpretive issues would benefit from further tempering or clarification. Overall, this remains a landmark study that is, in my view, appropriate for publication in Nature pending minor revisions.

Thank you!

R3.1. The authors have improved the presentation and discussion of the within-family analyses, particularly by emphasizing confidence intervals and the detectable magnitude of indirect effects. These revisions are appreciated and address my prior concerns to a large extent. That said, I would still encourage slightly more cautious framing of conclusions regarding the limited role of confounding. The current results more strongly rule out large sources of confounding than they do smaller (but potentially meaningful) effects, and a brief acknowledgment of this would strengthen the interpretation.

We now frame these results more cautiously, in line with the reviewer’s requests.

On p. 24, when describing results of within-family PGI analyses, we had previously written *“these tests reveal a negligible role of environmental confounding…”* – we now write *“these tests rule out large or moderate effects of familial confounding in genetic effects on personality; they indicate that such effects are negligible to small in magnitude.”*

On p. 27, when describing replication of population-level effect estimates in within-family data, we had previously written that these estimates *“provid[e] evidence for minimal uncontrolled*

confounding in population-level GWAS," and we now write that these estimates *indicate that "uncontrolled confounding in population-level GWAS is small in magnitude."*

Finally, in the discussion, on p. 28, we write that *"even small confounding effects can be meaningful*" before summarizing evidence for confounding.

R3.2. The revisions have improved the section on genetic correlations among spouses. However, I wasn't aware of how spouses were detecdted until reading the authors' response to reviewer 2. Previous approaches to detecting spouses in the UKB have major problems - not just that over half of likely spouses were not included, but also that the remaining spouses were ascertained in odd and unintended ways that could bias findings (Kelly, Horwitz, & Keller, 2026). The Kelly 2026 paper explains this issue and provides a simple algorithm to detect spouses in the UKB that is much more accurate and leads to 89K – 93K spouses. We suggest authors reanalyze using this approach.

As suggested by the reviewer, we have now run updated models of spousal genetic correlations using spouse pairs identified using the more accurate Kelly and colleagues (2026) algorithm. Like Kelly et al., we identify over 92K spouse pairs. Restricting to pairs where both members identified as White British and were genotyped results in a total of 78,236 spouse pairs available for genetic analysis of assortative mating. In the table below, we present the updated results for spouse pairs next to the previous results.

Genetic correlations between spouses In the UK Biobank across two spousal identification methodologies

	ReGPC Original Submission	Kelly et al. (2025) Approach
Extraversion	.027 (.011)	.025 (.008)
Agreeableness	.024 (.018)	.019 (.011)
Conscientiousness	.026 (.013)	.007 (.009)
Neuroticism	.009 (.013)	.017 (.008)
Openness to Experience	.068 (.013)	.051 (.009)

Note: Standard errors in parentheses.

It can be seen that these updated estimates are very similar to the previous estimates. Additionally, when they are meta-analyzed with the estimates for other relationship dyads and deCODE cohort dyads, assortative mating estimates for each Big Five trait are identical, to the hundredths place, to those reported previously.

Following these changes, we have revised p. 20 of the manuscript with updated sample sizes, Supplementary Tables S40 and S41 with revised exact estimates, and p. 35 of the supplementary text with updated methodology for spouse pair identification.

R3.3. The MR analyses are extensive but remain somewhat underrepresented in the main text relative to their scope. While I appreciate the authors' need to streamline the MS, I recommend modestly expanding the presentation of these results, either via a summary figure or a more explicit table of key findings. In addition, it would be helpful to clearly state whether multiple testing correction was applied and, if so, how.

In addition to a full one-page treatment of these results in the main text, we currently devote a full extended-data figure to these MR results. However, as we are currently at the maximum text length and maximum number of figures and extended-data figures, we cannot dedicate further space to these results.

To address the reviewer's request for clarity on inferential criteria, in the results section, we now clarify that *"We found plausible causal effects for 33 exposure-outcome pairings, where effects directionally replicated across MR estimators and multiple 95% confidence intervals excluded 0."* In the online Methods section, we write that *"To identify replicable effects in the context of multiple tests, we considered an effect to be significant only if estimates were directionally consistent across all three MR methods and with unadjusted 95% confidence/credibility intervals that excluded 0 across at least two methods."* In other words, to guard against spurious results, we identify effects that replicate across MR estimators and we do not adjust *p*-values for multiple testing.

Comments on Prior Review Points
I appreciate the authors' careful and thorough responses to prior concerns. Most issues have been addressed satisfactorily.
R3.4 The revisions to the within-family analyses and their interpretation are appreciated, although, as noted above, I encourage continued caution in drawing strong conclusions about the absence of confounding.

Thank you – as described above in **R3.1**, we now interpret these findings with more caution.

R3.6: It is satisfying to read the corresponding supplementary text explaining these interesting correlations. I do not believe ascertainment bias fully explains the negative correlation between conscientiousness and educational attainment. Exploring the genetic correlation between conscientiousness and cognitive ability might clarify the picture a little better. However, I recognize this is not the primary focus of the current paper, and I am perfectly satisfied with the authors' current approach.

We appreciate the reviewer's satisfaction with our current approach. We now make it more clear in the supplement (pp. 29-30) that this interpretation is just one possible explanation for the unexpected negative genetic correlation between conscientiousness and educational attainment. As suggested by the reviewer, we have added the following second possible explanation (pp. 29-30):

“A second possible explanation for the unexpected negative association between conscientiousness and educational attainment may relate to how educational attainment is associated with both cognitive and noncognitive factors, each of which may be differentially related to conscientiousness. In line with this explanation, Demange et al. (2021) found that conscientiousness was negatively associated with the cognitive component of educational attainment and positively associated with the noncognitive component of educational attainment.”

We agree with the reviewer that, as this is not the primary focus of the current paper, future research that is specifically tailored to this question will be valuable (p. 30 of the supplement).

R3.7. Thank you for providing the comparison table between genetic correlations and phenotypic correlations. I agree with the authors that most are consistent. However, it is intriguing that the genetic correlation between openness and conscientiousness operates in the opposite direction of the phenotypic correlation. Do the authors have a brief interpretation for this divergence?

We agree that this deviation is interesting. In the supplement (p. 13) we now describe two reasons for this potential divergence:

“The negative genetic correlation between conscientiousness and openness to experience (r_g = -.20) contrasts with the small positive correlation that has been obtained in phenotypic research (r = .08 in Park et al., 2020). One possible explanation for this divergence is ascertainment bias (van Alten et al., 2025). As education is positively correlated with both conscientiousness and openness to experience, this can induce a negative correlation between the traits in educated samples (see Lamp & MacKinnon, 2025). A second possible explanation for this divergence is that environmental factors that are partly independent of genotype may affect openness and conscientiousness in the same direction (both trait increasing or both trait decreasing) whereas genetic factors, on average, may act in opposing directions (e.g., trait increasing for conscientiousness and trait decreasing for openness, and vice versa). For instance, greater schooling may increase both conscientiousness and openness, even if genetic variants that are associated with higher conscientiousness tend to be associated with lower openness. This would result in a case where phenotypic correlations are more positive than genetic correlations. Further research is needed to evaluate these and other possible explanations.”

However, as with the negative genetic correlation between conscientiousness and educational attainment, speculations about underlying mechanism go beyond the scope of the present study, and we do not feel comfortable stating them in the main text, even surrounded by caveats. Such a hypothesis would need to be tested using other research designs such as randomized experiments, natural experiments, and or longitudinal schooling data.

R3.8. The final sentence remains somewhat awkward and could be rephrased for clarity.

We have rephrased this sentence so that it now reads *“Given its pervasive relevance, researchers across scientific fields will benefit from incorporating personality when predicting and explaining biopsychosocial outcomes.”*

R3.9. "Western geography" — This phrase is ambiguous. I recommend using a more precise geographic or demographic descriptor.

We now use the phrase “*four Western country clusters*.” We describe these clusters in the main text (p.11) – United States, Continental Europe, Nordic Europe, and UK/Australia.

R3.10. p. 21: "β .158, compared to β .165" — Missing equality signs (e.g., β = .158, compared to β = .165).

We have added these equality signs.

R3.11. p. 24: "The ratio of direct to confounding estimates is presented at the bottom of the corresponding bar (Supplementary Table S44)" — To clarify, should this ratio be the direct effect over the sum of the direct and confounding effects (direct / [direct + confounds])?

Correct. We have corrected this.

R3.12. p. 25: "mean r between rg estimates = .83" — I suggest revising this to "mean correlation among r_g estimates" for greater clarity.

This is a helpful suggestion; we have revised in this manner.

R3.13. p. 25: "within-family estimates replicated population-level estimates (from independent cohorts that did not contribute within-family data) at 95% of the magnitude that the population-level estimates from the same within-family cohorts replicated the population-level estimates (from the independent cohorts that did not contribute within-family data; Supplementary Table S47)." — The phrasing here is dense and confusing. I recommend breaking this into two distinct sentences to clarify the comparison.

To clarify, we now write:

“We separated population-level cohorts into those that contributed to the within-family GWAS and those that did not, and we submitted both to GWAS. We then separately regressed the within-family estimates and population-level estimates from these within-family-contributing cohorts on the population-level estimates from cohorts that did not contribute within-family data. The ratio of these two regression coefficients was .95, indicating that for significantly and suggestively associated SNPs, within-family effect size estimates replicate population-level estimates at 95% the magnitude, on average (Supplementary Table S47). This indicates that uncontrolled confounding in population-level GWAS of personality is small in magnitude and

provides further evidence that genetic effects on personality replicate across non-overlapping sets of participants."

R3.14. p. 27: "For agreeableness and neuroticism, genetic correlations between within-family and population GWAS were imperfect." — I suggest replacing "imperfect" with a more neutral and precise term, such as "not identical" or "attenuated."

We now write *"For agreeableness and neuroticism, within-family and population GWAS effect estimates were genetically correlated at less than unity ."*

Referee #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Thank you for your re-review!

Comments from AE:

please accept my apologies for the delay in our decision on your manuscript (we had issues with emails landing in reviewers' spam folders), "Robust inference and widespread correlates from genetic associations with personality", which has now been seen again by our referees. In the light of their advice I am delighted to say that we can in principle offer to publish it. First, however, we would like you to make some editorial changes to your paper so that it is as brief as possible and complies with our Guide to Authors. No peer reviewed data should be removed altogether when making these changes.

Thank you very much for this positive evaluation of our revision. We have revised the manuscript to respond to these final formatting changes, without removing any data. Changes to the text are highlighted in yellow in the manuscript.

We note one final small change to the manuscript – in between the most recent submission and this submission, we identified and corrected a minor error in the polygenic index pipeline for the Health and Retirement Survey dataset. We had inadvertently excluded some SNPs from polygenic index analyses. With this error corrected, one result changed. The PGI for openness to experience now predicts openness to experience in AFR participants (B = .047, SE = .017, p = .006). Previously, this prediction was non-significant (B = .033, SE = .017, p = .052). Correcting this error had no other substantive effects on results. This change is highlighted in yellow on p. 11.

As part of the revision, we would encourage you to follow and fill out the PRISMA checklist for meta-analyses (https://www.prisma-statement.org/prisma-2020-checklist). We realize that not all fields will be relevant for the study here, but it will be useful for members of the psychology community, who are familiar with non-genetic meta-analyses, and it will help with reporting in any case.

We appreciate this suggestion – we have added a PRISMA checklist (https://osf.io/sh8gr). We now mention this checklist in the Online Methods section.

In this regard, we also want to ask you to be very careful not to interpret null results (Bayesian or equivalent tests may be necessary to support such interpretations).

We have reviewed the manuscript and do not interpret null results of null hypothesis significance tests as the absence of an effect.

1. Please reduce the length of the title to 75 characters (with spaces) or fewer, so that it fits on two lines in the final layout.

We have reduced *"Robust inference and widespread correlates from genetic associations with personality"* (76 characters) to *"Robust inference and correlates from genetic associations with personality"* (64 characters)

2. Please provide the manuscript in .docx format. Currently it is in pdf format.

We have made this change.

3. Please add references to the abstract.

We have now done so.

4. The Chief Editor has approved 4,800 words for your paper. As you are aware, the current manuscript exceeds this limit by quite some margin. We ask that you make the necessary revisions to bring the article to budget. We recommend that you aim for more economical writing and to move more technical and/or peripheral descriptions and discussions of results to Supplementary Notes.

We have now made revisions to bring the article down to below 4,800 words, to 4,666 words. These changes were made by cutting text from the introduction and discussion so that they comply with Nature's length guidelines, by introducing more economical writing, and by moving ancillary results to the Supplementary Note.

5. Please remove the main figures from the article file and re-supply them individually in an acceptable format such as EPS, AI, PS, PDF, PPT, PSD or XLS (for graphs) with editable vector files.

We have made this change.

6. Please ensure that the text size in all figures is at least 5 pt Arial, keeping in mind that main figures cannot exceed 18cm width by 17cm height. It may be necessary to move panels from the main figures to the ED section and/or re-arrange panels to meet these requirements.

We have edited figures to comply with these text size and overall dimension requirements.

7. Please reduce subheadings to 40 characters (with spaces) or fewer.

All subheadings now comply with this formatting guideline.

8. You have more than one account on EJP. Please contact Nature Manuscripts (naturemanuscripts@nature.com) to confirm all accounts which belong to you, and your primary email address.

I have now linked all accounts under my primary email address (tedschwaba@gmail.com)

9. We have noticed that not all corresponding authors of the manuscript are ORCID linked. As part of our efforts to improve transparency in authorship, we now request that all authors identified as

'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS) prior to acceptance. For more information please visit http://www.springernature.com/orcid. If you have an alternative account which is ORCID linked, please contact johannes.black@springernature.com

We have now linked all three corresponding authors to their ORCID accounts.

10. There are potential third party rights issues in the figures. Please check the sources of all illustrations and clarify whether permissions are needed to adapt or reproduce them. Please make sure to include the relevant details in third party rights table when you resubmit (more information below). If Biorender or a similar software has been used, please also ensure to provide relevant licenses. In particular please check figure/s: main figure 2c. Please make sure to provide a third party rights table (more information below) when you resubmit.

We have edited figure 2c so that it no longer uses a third-party illustration (we now supply the illustration ourselves).

11. Please re-supply the Extended data figures individually in EPS, JPEG or TIF format.

We have re-supplied all Extended data figures as individual .TIFF files

12. The funding section is provided along the acknowledgment section, please provide it as separate section.

We have retitled this section "Funding" as it contains no acknowledgements.

13. The data availability statement must also refer to the individual-level data. The GWAS summary statistics must be live at the time of final resubmission. The sentence "This includes EUR, AFR, and trans-ancestry meta-analytic population-level summary statistics, polygenic index weights, and EUR within-family summary statistics, neuroticism summary statistics with data from UK Biobank participants excluded, and country-, age-, perspective- and questionnaire-stratified summary statistics." is lacking in clarity. Please can you explain why UKB was excluded?

We apologize for the ambiguity here. We now clarify that neuroticism summary statistics are available *"both including and excluding UK Biobank participants,"* so that researchers using these data can either include this cohort or hold it out for analyses as they prefer.

14. The name(s) of the IRB(s) that approved the study should be mentioned in the methods section (incl. study number if applicable), and a statement that research participants consented to the research is required.

At the beginning of the online methods section, we now write, *"As this meta-analysis used de-identified summary-level data, it was deemed exempt from ethics board approval by the University of Texas at Austin Institutional Review Board (STUDY00001941); ethics approval for individual cohorts that collected data and contributed to the meta-analysis is presented in the Supplementary Text. All participants consented to the research, and this research was performed in accordance with all relevant guidelines and regulations."*

15. Since your author list includes consortia, please consult the attached formatting instructions and revise as needed.

We have now added the consortia COGA collaborators and Spit for Science working group to the end of the manuscript, in compliance with the attached formatting instructions.