

Within-family heritability estimates for behavioural and disease phenotypes from 500,000 sibling pairs of diverse ancestries

Corresponding Author: Professor Loïc Yengo

Version 0:

Decision Letter:

2nd March 2026

Dear Professor Yengo,

Thank you once again for your manuscript, entitled "Within-family heritability estimates for behavioural and disease phenotypes from 500,000 sibling pairs of diverse ancestries," and for your patience during the peer review process. Please accept my sincere apologies for the delays in the peer review process, I am so grateful for your patience and understanding.

Your manuscript has now been evaluated by 3 reviewers, whose comments are included at the end of this letter. Although the reviewers find your work to be of interest, they also raise some important concerns. We are very interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these concerns in the form of a revised manuscript before we make a decision on publication.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within two months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

- **EXTENDED DATA FIGURES**

When re-submitting your manuscript, please ensure that any supplementary figures and tables that are crucial to the manuscript's conclusions are converted into Extended Data figures and tables to increase visibility of these data. Extended Data figures and tables are online-only (present in the online PDF and full-text HTML versions of the paper), peer-reviewed display items that provide essential background to the article but are not included in the main article due to space constraints. A maximum of ten Extended Data display items (figures and tables) is permitted.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,

[Redacted]

[Redacted]

Nature Human Behaviour

Reviewer expertise:

Reviewer #1: Behavioural genetics

Reviewer #2: Population genetics; quantitative genetics

Reviewer #3: Statistical genetics; behavioural genetics

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

Yengo et al. analyze a sample of ~500,000 sib pairs with molecular genetic data and data on 14 traits. For each pair, they compute the identity-by-descent (IBD) statistics and use Full Sibling IBD Regression (FSIR) to compute the heritability of each trait. For each trait, they also compute the share of trait variation that is due to the common environment (but those estimates can be biased by other factors such as assortative mating and non-additive genetic variance, as they write). With their enormous sample sizes, the authors report significant and precisely estimated heritability estimates for most traits. They also find that heritability estimates don't appear to vary much across ancestry. And importantly, the authors develop a method to transform estimates on the observed 0-1 scale to liability scale estimates.

This is a short, sweet, but excellent paper: the data is simply fantastic; the analysis is well executed; and overall the paper makes an important contribution.

I only have one main comment and a few rather minor comments.

1. My only not-so-minor comment is that the authors should devote at least some space in the main text (where they currently allocate almost no space) to introducing and explaining their method to convert h^2 and c^2 estimates from the observed to the liability scale. I believe this is also an important contribution of the paper, but the authors undersell it. The authors should explain that the commonly used Robertson/Dempster-Lerner formula is an approximation, explain why this is an approximation (I remember finding the literature opaque on this point – if I remember well it is because Dempster and Lerner assume a linear relation between the genetic value of an individual on the observed scale and that on the liability scale), and explain that their method is based on exact derivations (and computational approximation). They should also mention in the main text (not just in “Code Availability” at the end) that the package to use the method is available online.

2. The authors do not show the results for birth weight and no. of children in Figure 1. I'm guessing the reason for that is that these h^2 estimates are not significant (after correction for multiple hypothesis testing), but I think they should still be shown in the figure.

3. In the note to Figure S1, they should clarify what “the linear transformation classically used” is. The authors must be referring to the Dempster and Lerner formula, but if so they should say so (or else say what they are referring to).

4. Figure S2 (and similarly for S4), if I understand well, this is meant to be similar to Fig. S1, but for c^2 instead of h^2 . But if so, why not show c^2 on the liability scale on the x-axis and h^2 in the legend (and not vice-versa), and show the dotted line red for the Dempster-Lerner formula as they do in Figure S1 (since in principle that formula is also applicable for c^2).

5. What are the sample sizes (n) reported in ST1? They don't match those in the paper and in the other supplementary tables.

Signed: Jonathan P. Beauchamp

Reviewer #2 (Remarks to the Author):

The paper is an extension of the heritability analyses of Sidorenko et al. 2024, which presented very similar analyses for height and BMI, applied to other traits. They have increased the sample size for the two traits examined in that paper by over five fold, which has reduced the uncertainty in the estimates. For some of the other traits, the sample sizes are quite large as well, providing robust estimates of narrow-sense heritability.

The paper is concise, clear, and lays out the estimates. A real strength of the paper is its analysis of non-European ancestry populations. In presenting estimates of heritability with fewer assumptions and possible biases than other methods, the paper will likely be of general interest, and provides a point of reference against which to compare estimates of GWAS variance explained, polygenic prediction accuracy, and similar.

- The authors nicely lay out possible confounding issues or other factors that can bias the estimates of heritability and c^2 . Among those noted are non-additive effects and gene-environment interactions. The supplemental note 2 outlines the effects of non-additive variance on heritability and c^2 in a clear way, which will be helpful for readers to see. A bit more could be done on the general point of factors that influence the estimates:

The authors have noted that epistatic effects influence additive and c^2 estimates when unmodelled (supplemental note 2), and this is treated generally as non-additive variance in the discussion. They note that unmodelled epistatic variance will increase additive variance estimates and underestimate c^2 estimates (to some degree). Beyond epistatic effects, to the extent that there is a correlation of realized additive and dominance relatedness, then this would be expected to bias the estimate of heritability. In the authors' original Visscher 2006 PLoS Genetics article, there is a strong correlation between the additive and dominance relationships, and that paper noted that larger sample sizes will be required to distinguish additive from dominance variance. In Sidorenko et al., they estimated dominance variance, and argued that millions of individuals are needed because of the large uncertainty in the estimates.

I understand why the authors aren't computing the full realized epistatic relationship matrix, but the realized dominance relationship matrix is something they can estimate with the data in hand. In Sidorenko 2024 they estimated dominance variance, and the SEs of those estimates for height and BMI were on the order of 0.1 (supplemental table 12 there), but they have ~5 times the sample size here, which will reduce the SEs substantially. This seems like a good opportunity to address the issue with largest dataset available, and a dataset that very few others will have access to. If the estimates are really 0, that itself is interesting, and worth reporting at least in a supplemental table.

- Among the points at the end, gene-environment interaction point is another area of interest that they have not explored with as much detail as epistatic variance impacts. It would be nice to have a similar treatment of the gene-environment interaction impacts.

- Table 3 is great as a comparison to other published estimates, but the exact types of analyses included as comparators in Table 3 is restrictive. There are many twin studies available and extended twin family designs for many traits. There are such estimates available for some of those here in Table 2 but not in Table 3 (e.g., alcohol consumption, metabolic disease), but the authors usefully refer to "twin designs" as a comparison (line 165). Expanding Table 3 to the rest of the traits would be helpful, with a clear indication in Table 3 of the design of the comparison study. The discussion points about potential causes of the discrepancies is useful, and this should be expanded with additional traits.

Very minor points:

- Fig. 1: The colors are difficult to parse with the color scale used. It would help to make the colors a bit more distinct.
- L285: Best to specify that L_G is the length of the diploid genome.
- There was no link provided for zenodo to the R function.

Reviewer #2 (Remarks on code availability):

The zenodo script had no link to it, only a citation without a link.

Reviewer #4 (Remarks to the Author):

This study undertakes an extensive, well-executed effort to estimate the extent to which DNA statistically accounts for differences between siblings across 14 morphometric, behavioral, and health traits. The methodology is strong and the sample is impressively large and diverse. While the precise statistical approach is outside my primary expertise, it appears rigorous and executed by experts in the field.

My main concerns center on clarifying the conceptual framing, articulating the novelty, and improving accessibility for an interdisciplinary audience.

1) Clarify the gap, aims, and take-home message

The key gap, aims, and novel contribution are not fully clear, especially for the broad NHB readership. As written, the abstract's concluding sentence could be read as restating established insights from decades of twin studies (e.g., Polderman et al., 2015 meta-analysis; Turkheimer's "first law of behavior genetics that everything is heritable") and the previously discussed gap between family-based heritability and GWAS-based estimates.

Why do we need more precise estimates of direct genetic effects on complex traits (e.g., to define the upper limit of polygenic index prediction for concrete applications)? Why are such estimates not available to date (limitations of twin/family designs, indirect genetic effects in GWAS, assortative mating, PGI portability, etc.)? What does this study add beyond prior literature (e.g., design features that isolate direct effects, cross-ancestry consistency, trait breadth)?

2) Define the sample and discuss its novelty

The diversity and size of the sample may be the most distinctive aspect of the study across methodologies used to estimate heritability (including GWAS, twin studies that often underrepresent non-WEIRD samples). It is indeed notable that the within-sibling, DNA-based h^2 estimates are so consistent across ancestral groups. However, the introduction fails to give a sense of who these people are. How global is this pooled sample, what generations are they from? What ancestries are included?

3) Trait-wise interpretation in regards of gene-environment interplay

Moreover, the discussion lacks an in-depth examination of the results. What do the results tell us about gene-environment interplay in morphometric vs behavioral vs health traits? Those trait categories differ in the degree of h^2 and c^2 estimates. For health traits with lower h^2 or higher c^2 , do findings reiterate the primacy of environmental context relative to traits like height?

You note that "missing heritability" is smaller here - why might that be in a within-sibling, direct-effects framework? Are these estimates less affected by gene-environment interactions that contribute to A in twin models?

4) Explain or avoid genetic determinism terminology

Some language (e.g., the first sentence of the Introduction) refers to genetic determinism without explaining what that means in this context for the various traits. The Results seem to underscore both direct genetic and shared environmental contributions. It would help to preview in the Introduction that within-family DNA can inform shared environmental influences—this currently appears only in the Results.

Please add brief framing, caveats or remove evolutionary terminology, for instance equating "number of children" with "reproductive fitness". For demographers, parity is tightly linked to education, labor markets, and policy (and the comparatively lower level of h^2 may support this?). The brief mention of natural selection on complex traits feels underdeveloped or tangential. Either expand with relevance to the present estimates or remove.

5) Add explicit Limitations and Conclusions

A Limitations and conclusion section would strengthen the transparency and guide interpretation.

Minor points

- Clarify the distinction between genetic ancestry and ethnicity in the main text (e.g., 23andMe-derived genetic ancestry vs "Latino" as an ethnicity). These are not parallel categories; please define terms and usage precisely.
- Abstract opening: "direct genetic effects" are not confounded by between-family indirect effects (if that is the intended point).
- Fix typos

Figures

- Figure 2: Panel c is visually larger than a and b; consider harmonizing scales/layout.
- Consider overlaying h^2 estimates from complementary methods (twin, population-GWAS) for context.
- Consider adding a figure (or panels) for c^2 estimates rather than an "environmental blank space" and to parallel the h^2 presentation.

Reviewer #4 (Remarks on code availability):

Accessible

Version 1:

Decision Letter:

Our ref: NATHUMBEHAV-25084207A

15th July 2026

Dear Dr. Yengo,

Thank you for submitting your revised manuscript "Within-family heritability estimates for behavioural and disease phenotypes from 500,000 sibling pairs of diverse ancestries" (NATHUMBEHAV-25084207A). It has now been seen by the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within a couple of weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

[REDACTED]

[REDACTED]

Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

I had reviewed the original submission and had mostly minor comments, which the authors have addressed.

My only remaining suggestion is related to my original comment 2, in which I had mistakenly written that "[t]he authors do not show the results for birth weight and no. of children in Figure 1." What I mean to write is that the authors do not show the results for physical activity and handedness in Figures 2-3. I think it'd make sense for the authors to show these traits as well in those figures, or at the very least explain in the table notes why these are not shown.

Signed: Jonathan P. Beauchamp

Reviewer #2 (Remarks to the Author):

The authors have addressed my comments to the best of their ability, given the 23&Me constraints, which is unfortunate but understandable. The Table 3 expansion is helpful and the additions to Supplemental Note 2 addressed my other questions. It's a very nice paper that will be of general interest to the genetics research community.

Reviewer #4 (Remarks to the Author):

I thank the authors for their thorough and thoughtful responses to my comments. The revisions substantially improve the manuscript, and I am satisfied that my major concerns have been addressed.

The authors have meaningfully strengthened the Introduction. The addition clarifying that findings are broadly concordant with twin studies while providing better quantification across diverse populations — and identifying cases where between-family estimates are inflated — sharpens the take-home message considerably. The explicit naming of the groups (European, East Asian, African American, and Latino) in the introduction also gives the reader early orientation to the sample's scope. The clarification distinguishing genetic ancestry from ethnicity, noting that African American and Latino labels reflect genetically admixed groups, is appropriate. I would recommend adding a reference to the end of this sentence that tags a paper explicitly discussing this relationship. For example: "for discussion on the relationship of genetic ancestry and socially-constructed racial and ethnic identities see XXX". There are many publications on this, eg a consensus on "Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research", <https://www.nationalacademies.org/projects/HMD-HSP-21-14>

The new Limitations and Conclusions section is a substantive addition. The discussion of self-reported phenotypes as a

potential source of downward bias in both h^2 and c^2 is well-motivated, and the eye color example is an effective illustration. The authors are appropriately candid about the constraints imposed by their 23andMe collaboration framework. The conclusion that within-family estimates are on average substantially lower than twin-based estimates, while still demonstrating substantial genetic variation across diverse ancestries, is clear. And the acknowledgment that estimates might be capturing gene-by-shared-environment interactions is important.

The addition clarifying that FSIR can also inform shared environmental contributions — now appearing in the Introduction rather than only in the Results — improves the framing for interdisciplinary readers.

INTERNAL USE ONLY AIP modules required: ["Behavioural Science"]

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

Yengo et al. analyze a sample of ~500,000 sib pairs with molecular genetic data and data on 14 traits. For each pair, they compute the identity-by-descent (IBD) statistics and use Full Sibling IBD Regression (FSIR) to compute the heritability of each trait. For each trait, they also compute the share of trait variation that is due to the common environment (but those estimates can be biased by other factors such as assortative mating and non-additive genetic variance, as they write). With their enormous sample sizes, the authors report significant and precisely estimated heritability estimates for most traits. They also find that heritability estimates don't appear to vary much across ancestry. And importantly, the authors develop a method to transform estimates on the observed 0-1 scale to liability scale estimates.

This is a short, sweet, but excellent paper: the data is simply fantastic; the analysis is well executed; and overall the paper makes an important contribution.

I only have one main comment and a few rather minor comments.

We thank the referee for their accurate summary of our study and for acknowledging that it makes an important contribution to the field. We address below, each of their comments

1. My only not-so-minor comment is that the authors should devote at least some space in the main text (where they currently allocate almost no space) to introducing and explaining their method to convert h^2 and c^2 estimates from the observed to the liability scale. I believe this is also an important contribution of the paper, but the authors undersell it. The authors should explain that the commonly used Robertson/Dempster-Lerner formula is an approximation, explain why this is an approximation (I remember finding the literature opaque on this point – if I remember well it is because Dempster and Lerner assume a linear relation between the genetic value of an individual on the observed scale and that on the liability scale), and explain that their method is based on exact derivations (and computational approximation). They should also mention in the main text (not just in “Code Availability” at the end) that the package to use the method is available online.

We thank the referee for this point. We have now expanded the presentation of our method to convert heritability estimates from the observed to the liability scale. As suggested by the referee, we highlight that previous approaches to convert 0-1 scale estimates were derived by Robertson in the Appendix of Dempster & Lerner and that these approaches are approximations relying on a linear transformation of the 0-1 scale estimate, while our approach uses a method of moments using the expectation of the 0-1 scale estimates under a liability threshold model. We have added the following paragraph in the RESULTS section (lines 51 – 63): *“We analysed binary phenotypes on an observed 0-1 scale and developed a new method to transform estimates to an underlying liability scale as described in **Supplementary Note 1** and **Supplementary Figures 1–4**. Briefly, previous approaches to convert heritability estimates for binary traits from an observed 0-1 scale to an underlying continuous liability scale were derived by Robertson (in the appendix of Dempster and Lerner¹⁰) and further extended by Lee et al.¹¹ to account for ascertainment. Both approaches are first-order approximations largely valid when analysing distant relatives and rely on linear transformation of the 0-1 scale estimates. However, such linear transformation yields biased estimates in the context of FSIR, especially when the underlying heritability is large and the prevalence is small (**Supplementary Figure 1**).^{12,13} Instead, we developed a method of moments matching the expectation of FSIR estimates under the liability threshold model¹⁴ with observed estimates (**Supplementary Note 1**) and implemented this method in a publicly available R script (Code Availability).”*

2. The authors do not show the results for birth weight and no. of children in Figure 1. I'm guessing the reason for that is that these h^2 estimates are not significant (after correction for multiple hypothesis testing), but I think they should still be shown in the figure.

We thank the referee for this comment, but perhaps these results were not visible to the referee as they were, in fact, included in both Fig. 1 and 2.

3. In the note to Figure S1, they should clarify what “the linear transformation classically used” is. The authors must be referring to the Dempster and Lerner formula, but if so they should say so (or else say what they are referring to).

We thank the referee for this comment. We have now clarified that we referring to the Dempster and Lerner formula in the legend of Fig. S1.

4. Figure S2 (and similarly for S4), if I understand well, this is meant to be similar to Fig. S1, but for c^2 instead of h^2 . But if so, why not show c^2 on the liability scale on the x-axis and h^2 in the legend (and not vice-versa), and show the dotted line red for the Dempster-Lerner formula as they do in Figure S1 (since in principle that formula is also applicable for c^2).

We thank the referee for this question. Fig. S2 (and S4) are not analogous to S1 (or S3) as they focus on the relationship between heritability on the liability scale and the residual variance on the observed 0-1 scale, instead of the relationship between heritability on the liability scale and heritability on the 0-1 scale. This way of representing the data helps to better visualise the effect of varying heritability on the liability scale. What the referee is suggesting would lead to superimposing multiple curves and, therefore, reducing clarity.

5. What are the sample sizes (n) reported in ST1? They don't match those in the paper and in the other supplementary tables.

We thank the referee for this question and apologize for the confusion. The sample size reported in ST1 are the total number of individuals with the phenotype available. In the paper, we report sample size as the number of sibling pairs. We have now clarified this in the legend of ST1.

Signed: Jonathan P. Beauchamp

Reviewer #2 (Remarks to the Author):

The paper is an extension of the heritability analyses of Sidorenko et al. 2024, which presented very similar analyses for height and BMI, applied to other traits. They have increased the sample size for the two traits examined in that paper by over five fold, which has reduced the uncertainty in the estimates. For some of the other traits, the sample sizes are quite large as well, providing robust estimates of narrow-sense heritability.

The paper is concise, clear, and lays out the estimates. A real strength of the paper is its analysis of non-European ancestry populations. In presenting estimates of heritability with fewer assumptions and possible biases than other methods, the paper will likely be of general interest, and provides a point of reference against which to compare estimates of GWAS variance explained, polygenic prediction accuracy, and similar.

We thank the referee for their accurate summary of our study and for acknowledging that our contribution is of general interest and will provide a point of reference for future complex traits genetics studies. We address below each point raised by the referee.

- The authors nicely lay out possible confounding issues or other factors that can bias the estimates of heritability and c^2 . Among those noted are non-additive effects and gene-environment interactions. The supplemental note 2 outlines the effects of non-additive variance on heritability and c^2 in a clear way, which will be helpful for readers to see. A bit more could be done on the general point of factors that influence the estimates:

The authors have noted that epistatic effects influence additive and c^2 estimates when unmodelled (supplemental note 2), and this is treated generally as non-additive variance in the discussion. They note that unmodelled epistatic variance will increase additive variance estimates and underestimate c^2 estimates (to some degree). Beyond epistatic effects, to the extent that there is a correlation of realized additive and dominance relatedness, then this would be expected to bias the estimate of heritability. In the authors' original Visscher 2006 PLoS Genetics article, there is a strong correlation between the additive and dominance relationships, and that paper noted that larger sample sizes will be required to distinguish additive from dominance variance. In Sidorenko et al., they estimated dominance variance, and argued that millions of individuals are needed because of the large uncertainty in the estimates.

I understand why the authors aren't computing the full realized epistatic relationship matrix, but the realized dominance relationship matrix is something they can estimate with the data in hand. In Sidorenko 2024 they estimated dominance variance, and the SEs of those estimates for height and BMI were on the order of 0.1 (supplemental table 12 there), but they have ~5 times the sample size here, which will reduce the SEs substantially. This seems like a good opportunity to address the issue with largest dataset available, and a dataset that very few others will have access to. If the estimates are really 0, that itself is interesting, and worth reporting at least in a supplemental table.

We thank the referee of this important question. We agree this would have been a useful addition to the study. However, the framework of our collaboration with 23andMe (which started in 2021) does not allow us to perform additional analyses at this point. Following this referee's comment, we nevertheless approached 23andMe again to find ways to run the suggested analyses, but they were not in a position to commit any resources to our request. Nevertheless, we have expanded our **Supplementary Note 2** to show that other forms of unaccounted non-additive genetic effects such as dominance could also induce biases in both heritability and c^2 estimates.

- Among the points at the end, gene-environment interaction point is another area of interest that they have not explored with as much detail as epistatic variance impacts. It would be nice to have a similar treatment of the gene-environment interaction impacts.

We thank the referee for this point. GxE effects were mainly mentioned as a source of bias for between-family estimates. In the context within-family designs, GxE will only affect heritability estimates when the interacting environment is shared between siblings. We have now expanded **Supplementary Note 2** to specifically address the effect of interactions between genes and shared environment and that of dominance variance on FSIR estimates.

- Table 3 is great as a comparison to other published estimates, but the exact types of analyses included as comparators in Table 3 is restrictive. There are many twin studies available and extended twin family designs for many traits. There are such estimates available for some of those here in Table 2 but not in Table 3 (e.g., alcohol consumption, metabolic disease), but the authors usefully refer to "twin designs" as a comparison (line 165). Expanding Table 3 to the rest of the traits would be helpful, with a clear indication in Table 3 of the design of the comparison study. The discussion points about potential causes of the discrepancies is useful, and this should be expanded with additional traits.

We thank the referee for this point. We have now extended Table 3 from 6 to 12 phenotypes. Unfortunately, we have not been able to find twin-based or family-based studies for "burden of metabolic/heart diseases" (defined in our study as a composite phenotype) and "talk vs listen". Moreover, we now clearly distinguish estimates from twin-based designs and those from family-based designs using more distant relatives. Except for physical activity, we selected recent twin studies where reported (or meta-analysed between sexes) heritability estimates had a standard error lower than 0.1.

Very minor points:

- Fig. 1: The colors are difficult to parse with the color scale used. It would help to make the colors a bit more distinct.

We have changed the colours in Fig. 1 and the shape of the dots to help distinguish the different groups and traits.

- L285: Best to specify that L_G is the length of the diploid genome.

We have now added this clarification.

- There was no link provided for zenodo to the R function.

We apologize for this omission. The link was generated but included in the reference. We have now added the link (<https://doi.org/10.5281/zenodo.10416893>), which largely refers to code generated for our previous publication (Sidorenko et al. Nat. Genet. 2024).

Reviewer #2 (Remarks on code availability):

The zenodo script had no link to it, only a citation without a link.

We apologize for this omission. As mentioned above, we now provide the link the Code Availability section.

Reviewer #4 (Remarks to the Author):

This study undertakes an extensive, well-executed effort to estimate the extent to which DNA statistically accounts for differences between siblings across 14 morphometric, behavioral, and health traits. The methodology is strong and the sample is impressively large and diverse. While the precise statistical approach is outside my primary expertise, it appears rigorous and executed by experts in the field. My main concerns center on clarifying the conceptual framing, articulating the novelty, and improving accessibility for an interdisciplinary audience.

1) Clarify the gap, aims, and take-home message

The key gap, aims, and novel contribution are not fully clear, especially for the broad NHB readership. As written, the abstract's concluding sentence could be read as restating established insights from decades of twin studies (e.g., Polderman et al., 2015 meta-analysis; Turkheimer's "first law of behavior genetics that everything is heritable") and the previously discussed gap between family-based heritability and GWAS-based estimates.

Why do we need more precise estimates of direct genetic effects on complex traits (e.g., to define the upper limit of polygenic index prediction for concrete applications)? Why are such estimates not available to date (limitations of twin/family designs, indirect genetic effects in GWAS, assortative mating, PGI portability, etc.)? What does this study add beyond prior literature (e.g., design features that isolate direct effects, cross-ancestry consistency, trait breadth).

We thank the referee for these comments. The gap that our study addresses is that estimates using the gold standard IBD-based approach for estimating narrow sense heritability are not available for most traits, in most populations. Our introduction underlines that twin-based and GWAS-based designs, which are the most prevalent approaches, are prone to systematic biases (e.g., population stratification or assortative mating) that cannot be corrected without additional and sometime unverifiable assumptions. Our study uses the gold standard IBD-based approach, which is unaffected by these biases, but typically requires large samples to produce precise estimates. This explains, in part why this approach has not been wide used in the past, but our study addresses this limitation with its large sample size. For many traits, our conclusions are indeed consistent with twin studies in that we report evidence of substantial genetic variation underlying complex traits. However, our study provides a better quantification of that genetic variation (across diverse populations) and highlights examples (type 2 diabetes) where between-family estimates have been inflated. We have added the following sentence at the end of the introduction (lines 38 – 42): *"Overall, our conclusions are consistent with twin studies in that we report evidence of substantial genetic variation underlying human traits. However, our analyses provide a better quantification of that genetic variation (across diverse populations) and highlights cases where between-family estimates have been inflated."*

2) Define the sample and discuss its novelty

The diversity and size of the sample may be the most distinctive aspect of the study across methodologies used to estimate heritability (including GWAS, twin studies that often underrepresent non-WEIRD samples). It is indeed notable that the within-sibling, DNA-based h^2 estimates are so consistent across ancestral groups. However, the introduction fails to give a sense of who these people are. How global is this pooled sample, what generations are they from? What ancestries are included?

We thank the referee for this point. We have now expanded the last paragraph of the introduction to highlight that our study sample includes research participants of 23andMe, Inc and named the four genetic ancestry and ethnicity groups used to stratify individuals in our analyses. This paragraph now reads as

(lines 33 – 38): *“In this study, we use genomic data from up to 547,126 full sibling pairs among 23andMe, Inc. research participants of diverse genetic ancestry and ethnicity groups (European, East Asian, African American and Latino) to estimate the heritability (hereafter denoted h^2 , and its estimate $\hat{h}^2$) of 9 quantitative traits and 5 binary phenotypes, including 5 behavioural traits (smoking initiation, alcohol consumption, number of children, handedness and a personality trait) and 2 common diseases with behaviour-associated risk factors (asthma and type 2 diabetes).”*

We also highlight in the first paragraph of the Result section that the age (mean age: 43-52 across groups; average = 50) and sex (~40% males) distribution of all groups is reported in ST1.

3) Trait-wise interpretation in regards of gene-environment interplay

Moreover, the discussion lacks an in-depth examination of the results. What do the results tell us about gene-environment interplay in morphometric vs behavioral vs health traits? Those trait categories differ in the degree of h^2 and c^2 estimates. For health traits with lower h^2 or higher c^2 , do findings reiterate the primacy of environmental context relative to traits like height?

We thank the referee for these comments. It is difficult to draw robust conclusions on significant differences between anthropometric vs behavioural vs health related traits because of the small numbers of traits in each category. Moreover, our estimates may be differentially impacted by measurement errors because phenotypes are self-reported. We now mention among the limitations of the study that phenotypes were self-reported, which might have downwardly biased both our h^2 and c^2 estimates due to measurement error.

You note that “missing heritability” is smaller here - why might that be in a within-sibling, direct-effects framework? Are these estimates less affected by gene-environment interactions than contribute to A in twin models?

We thank the referee for these questions. We have updated the discussion section as follows (lines 186 – 199): *“The estimates from within-family segregation variance are generally much closer to estimates from pedigree or twin designs (**Table 3**) than estimates captured from GWAS data (SNP-based heritability), even when the latter is from whole-genome sequence (WGS) data. For example, the estimates of BMI heritability from within-family and pedigree designs are approximately 0.5 (**Table 3**) whereas the WGS-based estimate from Wainschtein and colleagues²⁰ is only about 0.34 (s.e. 0.01). Interestingly, Wainschtein and colleagues also reported an estimate of BMI heritability of 0.39 (s.e. 0.02) based on a between-family analysis accounting for non-additive genetic effects. Therefore, we cannot rule out that the discrepancy between WGS-based and within-family estimates for BMI could be explained by non-additive genetic effects (for example, gene-by-gene or gene-by-shared-environment interactions) contributing to within-family estimates (**Supplementary Note 2**) but not WGS-based estimates. In theory, it is possible to separate additive from non-additive genetic variation using the within-family design, but much larger sample sizes (of the order of millions of sibling pairs) would be required to separate such effects.^{8”}*

4) Explain or avoid genetic determinism terminology

Some language (e.g., the first sentence of the Introduction) refers to genetic determinism without explaining what that means in this context for the various traits. The Results seem to underscore both direct genetic and shared environmental contributions. It would help to preview in the Introduction that within-family DNA can inform shared environmental influences—this currently appears only in the Results.

We thank the referee for these points. We have replaced “determined” with “explained” in the first sentence of the introduction. We also added the following sentence (lines 15 – 17): *“Importantly, FSIR can also inform the degree to which shared environmental effects explain trait differences between siblings.”*.

Please add brief framing, caveats or remove evolutionary terminology, for instance equating “number of children” with “reproductive fitness”. For demographers, parity is tightly linked to education, labor markets, and policy (and the comparatively lower level of h^2 may support this?). The brief mention of natural selection on complex traits feels underdeveloped or tangential. Either expand with relevance to the present estimates or remove.

We have removed “reproductive fitness” from the text.

5) Add explicit Limitations and Conclusions

A Limitations and conclusion section would strengthen the transparency and guide interpretation.

We have now added a limitation and conclusion section at the end of the manuscript (lines 203 – 233): *“Our study has a few limitations. First, while our study is currently the largest and most genetically diverse of its kind, sample sizes in non-European ancestry groups remain limited. Therefore, future studies might illuminate subtle differences in complex trait heritability between populations that our study was not powered to detect. Second, all phenotypes in our study were self-reported, which might have induced downward biases in both $\widehat{h}^2$ and $\widehat{c}^2$, as result of measurement errors. For example, measurement errors might explain our unexpectedly low estimate of heritability for eye colour (0.45, s.e. 0.13), while twin-based studies have reported estimates of nearly 100% (**Table 3**). Nevertheless, it is worth noting that sibling correlations for height and BMI were highly consistent between our study and previous ones using directly measured phenotypes (Table 1), suggesting, at least for these traits, that the effect of measurement errors might be small. Moreover, our within-family estimates are also largely concordant with that from Markel and colleagues,⁹ which provides an independent replication of our results. Third, as briefly mentioned above, our estimates might be capturing non-additive genetic effects such as gene-by-gene or gene-by-shared-environment interactions (**Supplementary Note 2**). Larger studies including a comprehensive assessment of environmental exposures will be needed to fully separate these effects. Finally, IBD-based heritability estimates are still needed for many more traits than the ones analysed here. For example, there is no consensus in the literature on the magnitude of additive genetic variation associated with cognitive ability and educational attainment (EA). Reported within-family GWAS estimates (analogous to SNP-heritability) for EA are very low (0.04, ref.²¹), but pedigree-based (0.43 s.e. 0.04) and FSIR estimates from Young et al.⁷ in Iceland 0.40 (s.e. 0.15) were much larger. On the other hand, Markel and colleagues⁹ report an estimated EA heritability (unconstrained analysis) of ~0.08 (s.e. 0.09).⁹ Moreover, the estimates from within-family GWAS and within-family IBD are biased downwards because of assortative mating.^{8,17} We could not analyse EA and cognitive traits because of ethics restrictions. However, we emphasise that additional research using the within-family design is needed to inform ongoing debates about the relative importance of direct and indirect genetic effects for such traits.*

In conclusion, our findings are broadly concordant with twin studies yet quantify an ~40% lower heritability, on average, across the 12 phenotypes included in our study. Our results demonstrate, across diverse genetic ancestries, that complex phenotypes are influenced by substantial genetic variation beyond the effects of shared environmental exposures and assortative mating, with most of that genetic variation remaining unexplained by current GWAS.”.

Minor points

- Clarify the distinction between genetic ancestry and ethnicity in the main text (e.g., 23andMe-derived genetic ancestry vs “Latino” as an ethnicity). These are not parallel categories; please define terms and usage precisely.

We thank the referee for this point. These labels were given to us by 23andMe and emphasize the fact that the Latino group contains individuals with a wide range of genetic ancestries (African, Indigenous American and European). We added the following sentence in the main text (lines 49 – 51): *“Although inferred from genetic data, we hereafter refer to African American and Latino groups as ethnicity groups (as opposed to ancestry groups) because individuals in those groups are typically admixed with a range of genetic ancestries.”*.

- Abstract opening: “direct genetic effects” are not confounded by between-family indirect effects (if that is the intended point).

We thank the referee for this point. The sentence refers to the fact that estimates of direct genetic effects can be confounded, but not the effect themselves. We believe this is accurate.

- Fix typos

Figures

- Figure 2: Panel c is visually larger than a and b; consider harmonizing scales/layout.

We deliberately chose to have larger panel c because it displays results for more traits, which helps readability.

- Consider overlaying h^2 estimates from complementary methods (twin, population-GWAS) for context.

We thank the referee for this suggestion. However, we believe this may not improve clarity. There are lots of different methods to estimate heritability from GWAS data and equally many twin studies we can draw from. Table 3 (which we have now expanded) is meant to help compare our estimates with that of twin studies.

- Consider adding a figure (or panels) for c^2 estimates rather than an “environmental blank space” and to parallel the h^2 presentation.

We thank the referee for this suggestion. We have now generated a Fig.3 similar to Fig. 2 but focusing on c^2 estimates. Note that both h^2 and c^2 capture environmental effects such that there is no “environmental blank space” in our initial Fig. 2.

Reviewer #4 (Remarks on code availability):

Accessible

RESPONSE TO REFEREES

Reviewer #1:

Remarks to the Author:

I had reviewed the original submission and had mostly minor comments, which the authors have addressed.

My only remaining suggestion is related to my original comment 2, in which I had mistakenly written that “[t]he authors do not show the results for birth weight and no. of children in Figure 1.” What I mean to write is that the authors do not show the results for physical activity and handedness in Figures 2-3. I think it’d make sense for the authors to show these traits as well in those figures, or at the very least explain in the table notes why these are not shown.

Signed: Jonathan P. Beauchamp

We thank the referee for this comment. We have now displayed estimates of heritability (Fig. 2.) and residual correlations (Fig. 3) for these two traits.

Reviewer #2:

Remarks to the Author:

The authors have addressed my comments to the best of their ability, given the 23&Me constraints, which is unfortunate but understandable. The Table 3 expansion is helpful and the additions to Supplemental Note 2 addressed my other questions. It's a very nice paper that will be of general interest to the genetics research community.

We thank the referee for their suggestions, which have helped improved the quality of our manuscript.

Reviewer #4:

Remarks to the Author:

I thank the authors for their thorough and thoughtful responses to my comments. The revisions substantially improve the manuscript, and I am satisfied that my major concerns have been addressed.

We thank the referee for their suggestions, which have helped improved the quality of our manuscript.

The authors have meaningfully strengthened the Introduction. The addition clarifying that findings are broadly concordant with twin studies while providing better quantification across diverse populations — and identifying cases where between-family estimates are inflated — sharpens the take-home message considerably. The explicit naming of the groups (European, East Asian, African American, and Latino) in the introduction also gives the reader early orientation to the sample's scope. The clarification distinguishing genetic ancestry from ethnicity, noting that African American and Latino labels reflect genetically admixed groups, is appropriate. I would recommend adding a reference to the end of this sentence that tags a paper explicitly discussing this relationship. For example: “for discussion on the relationship of genetic ancestry and socially-constructed racial and ethnic identities see XXX”. There are many publications on this, eg a consensus on "Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research", <https://www.nationalacademies.org/projects/HMD-HSP-21-14>

We thank the referee for this suggestion. We have now added a reference to Lewis et al. Science (2022), which also addressed this question. “_”.

The new Limitations and Conclusions section is a substantive addition. The discussion of self-reported phenotypes as a potential source of downward bias in both h^2 and c^2 is well-motivated, and the eye color example is an effective illustration. The authors are appropriately candid about the constraints imposed by their 23andMe collaboration framework. The conclusion that within-family estimates are on average substantially lower than twin-based estimates, while still demonstrating substantial genetic variation across diverse ancestries, is clear. And the acknowledgment that estimates might be capturing gene-by-shared-environment interactions is important.

The addition clarifying that FSIR can also inform shared environmental contributions — now appearing in the Introduction rather than only in the Results — improves the framing for interdisciplinary readers.

Once more, we thank the referee for acknowledging that our revised manuscript has now addressed all of their major concerns.