Reviews

I'm grateful to the authors for their very detailed reply. The manuscript is clear and concise in synthesizing quite a lot of complex results, and I found it quite interesting to read. I have a few very minor suggestions, detailed below, and I will defer to the authors as to whether/how to make any further changes to the manuscript in response to these suggestions.

We thank the reviewer for his helpful comments.

1. The authors write "We obtained GWAS summary statistics from five large-scale biobanks a GWAS meta-analysis and a mega-analysis." This sentence seems like it needs a comma or two

We have now added the missing comma in the text. "We obtained GWAS summary statistics from five large-scale biobanks, a GWAS meta-analysis and a mega-analysis"

2. The authors wrote "Because we are using the exact same population panels to obtain population allele frequencies in all tests, the source of the inconsistencies must necessarily come from differences in the effect size estimates in the different GWAS" It might help to clarify in this line that the set of SNPs also differs between most of the comparisons performed, as the authors have mentioned elsewhere.

We have now expanded this sentence:

"Because we are using the exact same population panels to obtain population allele frequencies in all tests, the source of the inconsistencies must necessarily come from differences in the effect size estimates in the different GWAS, or be due to different SNPs passing the significance threshold in different GWAS."

3. The authors write "This suggests differences in scores are likely not driven by a biological signal and are instead driven by population stratification in GIANT and/or PAGE." I'm not sure I understand this sentence fully. I agree that the direction of the polygenic score difference between populations computed from estimated effects should be correlated with the "true" score difference between populations, but because these polygenic scores are computed from a subset of putatively causal SNPs that explain only a small percentage of the heritability, it seems like there is no guarantee that the estimated score difference between populations will be consistent in direction between studies, even if the differences are due to unbiased effect size estimates at true causal alleles. E.g., if we had only a small number of estimated effect sizes, one could easily infer a positive score difference between population A and B when the true difference using all of the unobserved effect sizes would be negative. I'm happy to be corrected here if I am misunderstanding the authors' point.

This is a good point. In that paragraph, we now have changed the word "likely" to "perhaps", and also emphasized the evidence for stratification in PAGE and GIANT (PCA alignment figures per Sohail et al. and regression figures per Berg et al.) in support of the argument for

this not being a biological signal, rather than this solely being because the scores are over-dispersed in PAGE and GIANT alone, or inconsistently dispersed when comparing against each other.

We have also removed this sentence: "Furthermore, in the case of height, the distribution of genetic scores when using GIANT estimates and when using PAGE estimates are not consistent."

4. "While modeling the individual effect of each of these on the inflation of the QX statistic is beyond the scope of this study, we note that all of these factors may be influencing the differences we observe among score sets." It might make sense to refer back to the partial attenuation of score differences reported when using a single set of SNPs here (i.e. Figures S8/S9).

We have added this sentence at the end of that paragraph: "Indeed, when controlling for one of these factors (the set of SNPs included in the score), we see an attenuation of differences between scores (Figures \ref{PS_height_05_ascUKB} and \ref{PS_height_08_ascUKB})."

Dear authors,

I commend you on the new analyses, revisions to the text, and clarification. All my comments have been addressed, and the manuscript will stand as an important reference for many researchers in the field.

We thank the reviewer for her kind words and helpful comments.

If the authors see value in this as well, I think it would be useful to see analyses like Figure 5 of population stratification for educational attainment as well to help interpret their newly added section comparing polygenic scores for this trait using different GWAS. This would of course be for the GWAS that they considered for this trait, and along a few axes of potential stratification (as they present a global analysis). I suggest this as, as a reader, an open question that remains for me with their new section is how much are the different GWAS that show different over dispersion values afflicted by stratification or not for this trait. That, is stratification as much a concern for educational attainment as they have shown it is for height?

We have now added the following figure to the supplementary figures:
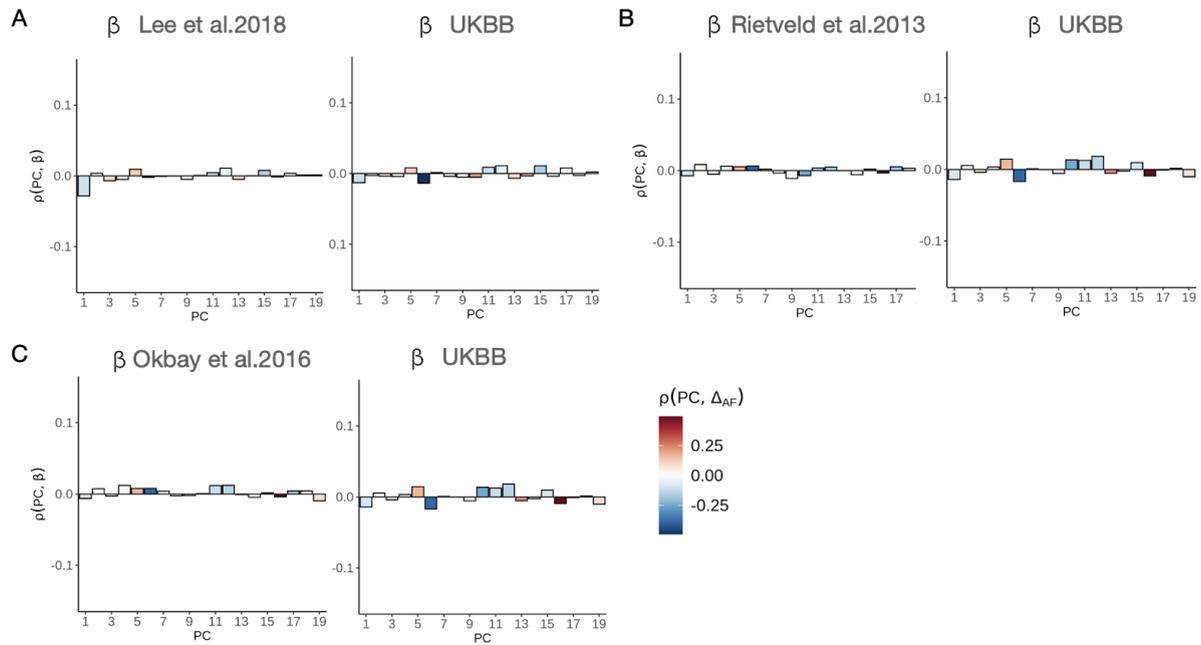
Pearson correlations between 20 PC loadings and educational attainment effect size estimates from a non-UKBB GWAS, compared to the same correlation using effect size estimates from the UKBB GWAS, for different choices of the non-UKBB GWAS. The correlations were computed using SNPs that are present in both the UKBB and non-UKBB GWAS cohorts, and in the 1000 Genomes Project. The barplots are coloured by the correlation between each loading and the allele frequency difference between GBR and TSI. A) Lee et al. 2016 vs. UKBB. B) Rietveld et al. 2018 vs. UKBB. C) Okbay et al. 2016 vs. UKBB.

We have also added this sentence to the Results section:

"Even though evidence for population stratification is weaker in these meta-analyses than in the height GWAS meta- and mega-analyses (Figure \ref{EduYearsPopst}), the axes of stratification still differ across studies, which might help to explain the differences in score distributions."

This revised version of the manuscript "How robust are cross-population signatures of polygenic adaptation in humans?" addressed, in my opinion, all points raised by the reviewers in the previous round of reviews. It addresses the important topic of how interpretations of polygenic risk score dispersion across populations can be misleading. The authors look at many traits for which there is at least two other GWAS apart from the UKBB, but they focus mostly on height and educational attainment - the former because of having multiple publicly available GWAS and previous claims for evidence of polygenic selection in Europe and the latter due to its high interest in the media and great potential for misappropriation by far-right hate groups.

The authors show unequivocally that measures of polygenic risk score dispersion across populations:

1) depend on the set of SNPs used

2) depends on whether SNPs were ascertained (chosen for the PRS) in the large single ancestry cohort (aka UKBB in this study) or in the non-UKBB ancestries, regardless of which effect sizes are subsequently used

3) depends on how homogenous (ancestry-wise) the GWAS is - the more homogenous, the less overdispersion is observed

4) and whether the inference comes from a single GWAS vs meta-analysis (even of a single ancestry). To illustrate this the point they split the UKBB into sub-cohorts and then meta-analyzed them, finding increased overdispersion of PRS mimicking that seen for GIANT and not seen in the UKBB single-cohort analysis

5) is inconsistent for educational attainment depending on the GWAS chosen, resulting in different ancestries having higher PRS values

6) in brief, it is increased by having multiple ancestries in the GWAS (e.g. the PAGE study) and/or multiple sub-cohorts composing a meta-analysis (e.g. GIANT)

7) the patterns seen for meta-analyses are independent of the method (standard error or sample size based)

Their findings strongly suggest that population stratification in GIANT and PAEGE are driving these findings, although the possibility remains that more diverse cohorts such as PAGE are better at capturing true biological signals that are overcorrected for in the UK Biobank.

Overall I think this is a great contribution to the field and an important methodological manuscript on the caveats and biases involved in polygenic selection studies/interpretation.

The authors used publicly available data and provided a link for a repository containing the scripts needed to reproduce this analysis.

Finally, I want to enthusiastically commend the authors for plainly pointing out how this kind of study has enormous potential for misinterpretation (e.g. educational attainment). I would like to see this become way more prevalent in the literature.

We thank the reviewer for her kind words and helpful comments.

Very minor comments:

1. for clarity: please make sure to emphasize which 'score' authors are talking about in different parts of the paper: the PRS or the Qx.

We have now addressed this comment.

2. for the educational attainment GWASs, how did the authors handle the sample overlap?

We took the summary statistics of each educational attainment GWAS "as is", without modeling the sample overlap, though we caution in the Methods section that there is indeed sample overlap between Lee et al., Okbay et al. and the UK Biobank GWAS:

"A meta-analysis of 1,131,881 individuals \citep{lee2018gene} (71 cohorts in total). Note that this study includes the samples from the \cite{okbay2016genome} study and the UK Biobank as well."

We have also clarified this below that sentence:

"We took the summary statistics of each educational attainment GWAS "as is", without modeling sample overlap between cohorts."

3. table S2: Since the N for GIANT is variable across positions, authors should clarify that this value is the maximum N.

We have now clarified this in the table caption:

"Full list of all traits tested and the total number of individuals (N) is shown for the GWAS in which this data was available. For GWAS summary statistics with variable N across positions, we list the maximum N for that study."

4. tables S5-S6: I imagine these refer to height but the captions should say it.

We have addressed this now:

"Pairwise Pearson correlation coefficient between height effect size estimates from the UKBB GWAS and from another GWAS. The SNPs used were determined based on their P-value in the non-UKBB study. n = number of SNPs used to compute the correlation."