“How robust are cross-population signatures of polygenic adaptation in humans?” by Refoyo-Martinez et al is an important and timely study. It builds on work previously done in Sohail et al 2019 and Berg et al 2019 among other studies, by analyzing multiple existing biobanks and multiple traits, and assessing polygenic adaptation using polygenic scores constructed in the 1000 genomes phase 3 populations. They further assess different approaches for running a GWAS in the UK Biobank data and, in particular, the population stratification introduced by the meta-analysis approach.

I don’t have any major concerns with this study. Below I have some questions, comments and a few additional analyses I would like to see added.

- The authors say that: “Starting from 805, 426 genotypes variants across the genome, we restricted to SNPs with a minor allele frequency (MAF) < 5% globally.” Do they mean MAF > 5% globally?

- They say in the text “P < 0.05/n, where n is the number of assessed GWAS.” However, in the caption for figure 2, n and m are defined differently. It is generally not clear why there are two Bonferroni corrected P values and how they are used. Wouldn’t it make more sense to use one corrected threshold, corrected by number of assessed GWAS times number of traits? The authors should clarify and justify in the text, and use consistently throughout.

- The LD score regression ratio – can the authors define what this measures in the text when they describe the result and how should the reader interpret it relative to LD score intercept which is more widely considered?

- In the meta-analysis of UKB, were PCs included in the analysis of each cohort before the meta-analysis? Were they included during the meta-analysis? Please clarify this in the text. The GIANT study appears to have different cohorts that had different kinds of correction for population structure (some with PCs, some with self-identified ancestry, and some likely with none). Can the authors comment on this and how it relates to their meta-analysis and ability to the compare the two?

- All the mixed model analyses (british, white and all ethnicities) show that that Africans have the highest polygenic score for height. How to interpret this? Can the authors comment on this in the text?

- Please add error bars for figures 3 and 5 and related supplementary figures.

- In figure 6, the ordering of rows of the different association methods is not following any specific coherent order. Could the authors re-order these to make the figure easier for the reader to interpret?

- Population stratification. Can the authors provide more intuition in the text for the particular way used to demonstrate population stratification to help the reader follow along? For example, why may we expect to see stratification in FINRISK or in UKB along the dimension of Eurasian stratification (GBR-CHB allele frequency differences)?

  I think it would clearer and more intuitive for the research community and wider readership to see analyses of stratification (along different axes) within each GWAS for

different traits. One way of doing this would be to plot the correlation of 1000g SNP PC loadings with effect size estimates for PCs 1-20 (see fig 2a of Sohail et al eLife) in each of the GWAS considered. This would show effect of stratification along different axes (not only European). Another way would be to show, similarly to the authors do, allele frequencies in two 1000g populations, but as two separate axes for each SNP colored by effect size (see fig. 2c of Sohail et al eLife). This would result in a single plot for each GWAS for a given trait, and can be made to observe different kinds of stratification using allele frequency comparisons between different populations. For example, within each GWAS, they can look at “within-population” stratification, the same way they do with European stratification, as well as that reflecting stratification in or with other populations. Ideally they would do this for a range of traits, and at least the ones deemed significant for polygenic adaptation in the authors’ and other published analyses. Adding these analyses to the manuscript would make it’s characterization of population structure in different GWAS more comprehensive and helpful, and allow it to serve as a reference for assessing kinds of population stratification in each GWAS with respect to different traits.

- Educational attainment has been the focus of studies of polygenic adaptation (Racimo et al 2018) and differential polygenic scores (Abdellaoui et al 2019), using both meta-analysis GWAS (which the authors show is an approach likely to generate population stratification) as well as UKB GWAS. Since height has already been shown to be affected by stratification in previous studies, I think the authors would add further novelty to their work, as well as address an important issue by assessing polygenic adaptation and population stratification in different GWAS (used by previously published studies) for this trait, and add it to their figures and narrative. I believe it warrants at least a comment in light on their analyses of the meta-analysis approach for GWAS, and its use in these previous studies.

- The authors should at least comment in the manuscript on different LD approaches used to compute polygenic scores, and give a discussion based in the literature of how to think about the different approaches with respect to their results, and what approach makes sense when comparing across different biobanks (If LD is different in different biobanks, is their choice of using LD windows to compute polygenic scores well suited and why? Were LD windows picked from the European panel for all biobanks or from different panels for different biobanks (should be described in manuscript)? It has been argued (Chen et al AJHG 2020) that the 1700 LD window approach of computing polygenic scores is most affect by stratification issues, compared to a clumping approach. Ideally, they would do some analyses to check and show the effects of these choices in computing polygenic scores and discussion of those results as well.

- The authors should give some more discussion of how to think about the results for height in the different biobanks, in light of their stratification analyses. Are they concluding that the only biobanks that show overdispersion in Qx for height do so because of stratification? If the meta-analysis approach induces stratification, can they say this more explicitly, and make clearer for the reader what this means given that GIANT and PAGE study different populations, one only European, and the other a range of diverse cohorts? What is going on with studies such as FINRISK and APCDR?