"How robust are cross-population signatures of polygenic adaptation in humans?" by Refoyo-Martínez et al examines the evidence for polygenic selection on trait-associated variation using a single population reference (26 populations from the TGP) and summary statistics obtained from several different GWAS. They apply the Qx framework (Berg & Coop, 2014), finding contrasting/conflicting results across different GWAS and analysis approaches (e.g. meta-analysis vs single-cohort GWAS, or BOLT-LMM vs Plink's LM). They find generally weaker evidence for polygenic selection on height-associated loci than many previous studies that used heterogeneous cohorts and meta-analyses. These results add to a growing literature that cautions against the overinterpretation of polygenic scores, which may be subject to numerous biases. While previous papers (Berg et al 2019, Sohail et al 2019) found qualitatively similar results by comparing just 2 GWAS (UKBB and GIANT) or 3 GWAS (UKBB, GIANT, and BBJ -- Chen et al 2020), the current manuscript goes further in assessing potential sources of bias in summary statistics and compares across several more GWAS. They argue that meta-analysis can introduce substantial biases, as has been suggested but to my knowledge not thoroughly examined, and that more homogeneous samples seem to result in less stratification. While the latter finding is predicted from population genetic theory, this paper adds to the growing empirical support for this claim.

Overall, I think this is a very nice analysis that adds a lot to an important and rapidly developing area of research. I do have some questions relating to the interpretation of some of the findings. In particular, the authors argue that the results are strongly suggestive of uncorrected stratification biases in several of the datasets. While my personal bias would be to agree with them, in the absence of any model-based simulations I think some of the interpretation may need additional caveats or supporting analyses. I also have several minor comments about methods and suggestions for clarity.

Please note that although this is a relatively long review, the length reflects my high level of interest in the questions pursued by the authors, rather than an overall critical take on their manuscript.

Main comments:

A primary finding of the paper is that several of the analyses are suggestive of residual population stratification that inflates/biases effect size estimates in GWAS, and leads to spurious selection signals. The idea seems to be that if there were no stratification biases in each of the GWAS, then the effect size estimates would be essentially the same across the studies, and the results would be consistent regardless of which study is used for the GWAS step (e.g., the authors write in the discussion that “the distribution of genetic scores when using GIANT estimates and when using PAGE estimates are not consistent, suggesting differences in scores are likely not driven by a biological signal that is not being picked up by the biobank-based tests.”)

This would certainly be true in limit of 100% statistical power in each GWAS, additionally assuming that effects are not environmentally dependent or population specific. My concern here is that in general the authors have limited the analyses to sets of SNPs for which there were significant associations within each study (although they have performed some limited tests

using sets of SNPs that do not rely on this conditioning -- see below). Conditioning on p-values introduces several possible effects on the analyses, and in the current draft it seems that the authors have not fully elaborated on these effects, as detailed in the main comments below.

1. In comparing the Qx calculations across various GWAS run in different populations, it's not entirely clear how much agreement we should expect even if the individual GWAS do not have any stratification. The null model in a single instance of a Qx analysis is clear -- the population history is specified, and the null corresponds to neutral drift of trait-associated alleles. In a comparison across multiple different Qx analyses using different sets of associated alleles, the null is less clear, because each individual GWAS imposes strong conditioning on the set of SNPs included in the analyses, as well as their estimated effect sizes. For example, we expect the SNPs associated with the phenotype in a European cohort to explain more variance in Europe than other populations, because they need to have large enough allele frequencies in Europe to be detected. Hence it is not obvious that we should expect Qx to return the same results for each GWAS, especially if some of the GWAS are underpowered. When we do a GWAS in Europeans and find evidence for polygenic adaptation but we don't find it when we do a GWAS in Japan -- is it a false negative in the BBJ or a false positive in UKBB? Moreover, if stabilizing/negative selection acts on some of the trait-related variation, it will affect statistical power by constraining large effect alleles to lower frequencies, and make large effect alleles more population specific.

In principle, one way to approach this is to simulate polygenic selection on the population history inferred from the 1000 genomes populations, run a GWAS, and then assess the effect on Qx. In practice, this would be a massive effort and I do not necessarily recommend it for this study, especially given the complexity of the population history of 26 TGP populations. However, in the absence of this simulation, it would help if the authors could expand their discussion of their expectations under the null and caveats of their interpretation. They also could perform some of the analyses by conditioning on p-value in one GWAS cohort, and then using this single set of SNPs for all of the GWAS. The authors have employed this approach in some of their stratification analyses (Figs. S19-S23), but not to the Qx/polygenic scores analyses as far as I can tell. Of note, this suggestion is very similar to the analysis in Chen et al 2020, who ascertained the effects in BBJ to infer selection in UKBB (published in AJHG -- the biorxiv version of this paper is cited by the authors). A very similar analysis was also applied in Tucci et al 2018 (Science): https://science.sciencemag.org/content/361/6401/511.abstract

2. Figure 3 shows the polygenic scores across TGP populations using summary stats from various height GWAS. When the analysis is performed in GIANT, Europe shows up as having elevated scores, whereas African populations have elevated scores in PAGE.

Similar to my question above, I'm not sure that stratification in GIANT/PAGE is the only way to interpret this result. The sets of SNPs used to compute these scores are not the same, and in principle we should have better power to find SNPs with large contributions to the phenotypic variance in populations from Africa when we include individuals of African ancestry in the GWAS. Stratification is certainly a potential (and in my opinion, likely) explanation. But could we not be overcorrecting population stratification in the more homogeneous samples, or getting more complete representation of the SNPs that contribute to the African polygenic score in

PAGE? If the authors can rule this possibility out, it would help to spell out how, and if not it would help to add these caveats.

3. The authors write "Berg et al. (2019) looked for latent population stratification by studying the relation between allele frequency differences in two GWAS and their difference in effect size estimates. Presumably, if neither GWAS is affected by population stratification, there should not be a relation between these two variables." One minor comment, it seems that what was actually plotted in Berg et al (Figure 4) is not the difference in frequency between the GWAS samples, but differences in frequency between human populations (TSI and GBR) as compared to the difference in effect size estimates between two GWAS, which is the same as what the authors plot in their Figure S19.

More broadly, the authors state that we expect no relationship between frequency difference (e.g., between GBR and TSI) and effect size difference (e.g., between GIANT and UKBB) if there is no stratification. However, by conditioning on p-value, we expect systematically larger unsigned effects in the GWAS that we used to compute the p-values than another GWAS, due to the winner's curse (e.g., bigger expected effect size in GIANT than UKBB when conditioning on a low GIANT p-value). The magnitude of this bias depends on the variance explained by the SNP -- SNPs that fall close to the threshold for detection in the GWAS will tend to have larger biases. To me, the question then becomes whether the magnitude of this winner's curse bias can vary with frequency differences between populations for reasons other than stratification.

It's not obvious to me if/when this could happen, but it seems plausible that polygenic adaptation could do this, by systematically causing the magnitude and sign of the true effect size to covary with population frequency differences. In any case what I'm proposing may be a non-issue, but my overall point is that the analysis here involves some complex conditioning that depends on allele frequencies, LD, statistical power for detection, effect size inflation due to the winner's curse, and the underlying evolutionary genetic process that "generates" the observed data. Although stratification is a likely mechanism, I don't feel altogether comfortable with the degree of weight the authors have put on it.

In Uricchio et al (2019, Evolution Letters) we did a qualitatively similar analysis, but we compared SNPs based on p-value rank (Fig S6-B). We also found a substantial difference between GIANT and UKBB, which we interpreted as likely due to stratification. However, we could not rule out overcorrection of effect sizes in UKBB on the basis of our analysis alone. The lack of signal in sib-based effects seems to be strongest argument in favor of stratification at the moment, as far as I can tell.

4. All of the meta-analysis simulations the authors performed on UKBB result in inflated p-values (Figure 6). However, the all-ethnicities UKBB LM analysis, which is not a meta-analysis, also resulted in inflated p-values but a slightly lower Qx. The authors wrote "we created an artificial meta-analysis on the entire UKBB cohort" -- hence this meta-analysis includes all of the individuals included in the all-ethnicities analysis? If so, it is unclear to me whether the inflated p-values are mostly due to the meta-analysis itself, or just the more heterogeneous cohort. In fact, the polygenic scores reported for the "all ethnicities linear model" in Figure 5 look almost identical to the scores reported for all of the meta-analyses. Moreover, can the authors suggest

any mechanism? It was not clear to me why the random partitions should result in inflated p-values, given that there should be no stratification in random partitions.

Minor Comments (I will often start these comments with quotes, which are taken from the manuscript):

1. I realize that Qx is often described as a polygenic adaptation test, but is it really? It seems more like a neutral-null violation test, without regard to the particular source of violation. E.g., could overdispersion be driven by negative selection alone? Or stabilizing selection alone?

2. "They looked for overdispersion in the frequencies of trait-associated variants across populations, relative to a neutral null model." I think this is a slightly narrow way to describe these previous papers, as some also looked for directional changes, not just overdispersion. Turchin et al (e.g.) also finds that height-increasing SNPs have higher frequencies in northern Europe.

3. "variants with a minor allele frequency lower than 0.01 and those classified as low confident variants whenever this information was available in the summary statistics file" Please clarify whether the exclusion is on a per-population basis (e.g., frequency is 1% in population X, or total allele frequency across all sampled populations). Same issue for the low confidence variants, and it would help to have information about the specific threshold used to assess confidence.

4. "which serves as a fairer comparison among studies" could use a reference.

5. "In order to build an empirical genome-wide covariance matrix (F-matrix) with non-associated SNPs, we extracted all SNPs with a P-value larger than 5e−8 and then sampled every 20th "nonassociated" SNP across the entire genome. " Does this mean every 20th SNP regardless of LD in the region? Were any additional filters on quality or allele frequency applied, as was done for the associated SNPs?

6. Much of the text in the "Robustness of signal of selection and population-level differences" section is duplicated from the methods, I would suggest trimming it.

7. "we observe little notable differences in P-values" this is a bit awkwardly phrased, I believe this means that the P-values are similar across the various approaches but it could be rephrased to be clearer.

8. The LD blocks used herein are derived from a subset of human populations -- is it possible they are not representative of some of the populations?

9. Fig. S18 – the only population with a very large sample is very similar in terms of ancestry to UKBB (GIANT). While it certainly seems true that sample size should have large influence here, it's not obvious that trend would persist without this outlier included.

10. "Approaches based on tree sequence reconstructions along the genome appear to be a fruitful avenue of research towards the development of methods that can properly control for some of

these confounders." I agree. Can the authors expand on this point? Why do they think this is true, and how will such approaches improve on the state of the art?

11. Chen et al, which the authors have cited as its biorxiv version, is now published in AJHG: https://www.sciencedirect.com/science/article/abs/pii/S0002929720301610

12. "As an example, Figure S1 shows the distribution of effect size estimates of 1700 approximately independent SNPs for height ($P < 1e{-}5$)". I understand that there are ~1700 blocks, but are there also ~1700 SNPs? In Figure S5 it looks like there are only about 1100 in UKBB after conditioning on p-value, and fewer in the other studies.

13. The authors employ a sign-randomization approach to computing p-values for Qx, which sometimes returns very different p-values from the other approaches based on frequency matched SNPs. Do the authors have any idea why? My first thought was that sign randomization may affect latent LD between the associated SNPs, but I'm not sure this is likely given the LD pruning approach they have used. A second possibility might be that random SNPs are not matched for background selection strength -- associated SNPs are likely to be in regions of substantial background selection, perhaps exacerbating their drift (see Torres, Szpiech, & Hernandez 2018 for example). I will note that we used the same LD blocks in our analysis in Uricchio et al 2019, so if latent LD is a problem here then it may also be a problem in our paper, although we only used Europeans.

14. "Importantly, those SNPs that also have a significant P-value in the non-UKBB GWAS in each comparison (colored in red in Figure 4) show a higher correlation than the rest of the SNPs." As the authors likely know, this effect is expected due to the winner's curse. The effect will be exacerbated when one study has a much smaller cohort (due to lower power for intermediate effect SNPs) and also potentially exacerbated by distinct LD patterns when the study populations are not closely related. Berg et al 2019 briefly discusses this effect. Some mention here might help for clarity.

15. In the "Evidence for population stratification" section, I believe all the effect size estimates correspond to height, but this is not specified in this section's text or in the figure legends.

16. Several comparisons with PAGE result in large numbers of SNPs with ~0 frequency difference but very large effect size differences. Do the authors have any sense for why this occurs? Perhaps differences in LD, or population-specific effects?

17. "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." I agree, but I think it is also important to note that the specific set of SNPs varies between cohorts, because of the conditioning on p-value.

18. "To try to avoid stratification issues, recent studies have proposed to look for evidence for polygenic adaptation within the same panel that was used to obtain SNP effect size estimates, i.e. avoiding comparisons between populations that might be made up of individuals outside of the GWAS used to obtain effect size estimates, e.g. (Liu et al. , 2018b)." This seems a strange

suggestion to me, and exactly opposite of the analysis in Chen et al 2020 (AJHG), which ascertains the effects in one cohort and then applies those effects in a "distantly" related population. I would think that recycling SNP effects within populations is exactly the problem that caused the spurious selection inferences to begin with. Perhaps the authors can clarify how this would work or help.

19. I was pleased to see a note about the potential misuse of polygenic scores by hate groups at the end of the paper. Thank you, we should all be doing this.

Thanks for the opportunity to review your very interesting manuscript. I want to emphasize again that while I had many comments, I think most can be handled by minor alterations to the text or simply correcting me if I was mistaken about anything. I am happy to correspond with the authors should any comments need clarification.

Sincerely,
Lawrence Uricchio