Hi Kelsey,

This paper reports a substantial and impressive body of work. I really like it. The simulated and theoretical results support each other very well.

I have only minor comments, most of which relate to readability (Sharon gets the same type of comments from me when I edit our co-authored papers for which she is the first author). Many of these comments are suggestions for eliminating words, because eliminating unnecessary words can often increase a sentence’s crispness and impact. You can decide whether to incorporate these suggestions depending on whether the suggested revision sounds better to your ear.

The only big picture comment I have is that in your future read-throughs of the manuscript it would be good to be on the lookout for material that could be made more concise or moved to the Supplemental Information.

My comments are listed below.

Best regards,

Brian

-------------

Excellent title. It is informative and grabs your attention.

Abstract. Some words do not seem to be necessary.

* “more PCs than may be actually necessary” [“actually” could be deleted]
* “However, in this paper we show that adjusting for extraneous PCs can actually induce spurious associations.” [“However” and “actually” could be deleted]
* “atypical linkage disequilibrium (LD) patterns” [“patterns” could be deleted]

General comment. I won’t comment on future uses of “actually” in the paper. If you agree that the sentences in the abstract read better with “actually” deleted, then I suggest doing a global search for uses of “actually” and deciding whether the word is ~~actually~~ needed in other sentences.

Abstract. The sentence beginning “Investigating three samples of African American individuals” is tough to digest because of its length. I suggest breaking this into two sentences: one sentence describing the studies analyzed, and one sentence describing the results.

Abstract. “a model-based estimate of genetic ancestry”.

* Would it be clearer (and still correct) to say “an estimate of genome-wide ancestry proportions”?

General Comment. The adjective “model-based” is used several times in the paper to describe the estimated global ancestry proportions. As a non-statistician, I find the term “model-based” a bit confusing because I don’t see why emphasizing that the estimates come from a model is important here. How critical is it to highlight that the estimates are “model-based”? If it is not critical, you could consider replacing “model-based” with “estimated”.

P. 3. Here are some more words that I think could be deleted:

* “… has been observed in many studies of admixed populations such as African Americans and Hispanics/Latinos” [“many” and “such as African Americans and Hispanic/Latinos” could be removed].
* “It has been widely documented that heterogeneous global ancestry, …” [“It has been widely documented” could be deleted. Deleting this phrase has the added benefit of changing passive voice to active voice].
* “as one (among many) reason” [“(among many)” could be removed.]

P. 3 “This fixed effects adjustment for global ancestry is used extensively throughout the literature, …”

* Consider including a few representative citations here. The phrase “throughout the literature” could be replaced with “in published studies”.

P. 4 Words that could be deleted:

* “(PCA) is a widely-implemented unsupervised approach …” [ “widely-implemented” could be deleted.
* “… but in practice it is not uncommon to see applications …” [“not uncommon” could be replaced by “common”]
* “and/or small regions of the genome” [I think “and/or” could be replaced with “and” without changing the meaning of the sentence]

General comment. Is it possible to renumber the citations in order of appearance in the text? This would have the benefit of replacing several non-consecutive reference numbers in the citations with a simple, hyphenated range of reference numbers, which will make it easier for reviewers to find the cited lists of referenced papers in the bibliography. If it is not easy to do this in LaTeX, then the journal can reorder citations when the journal typesets the article.

P. 5, “The above-cited suggestions regarding LD pruning and filtering are not universally implemented …”

* I would be inclined to replace “implemented” with “followed” or “practiced” (I tend to use “implemented” to refer to implementing an algorithm in a computer program).
* Add a comma after the first independent clause.

P. 6, Table 1 legend.

* “… or otherwise unusual patterns of linkage disequilibrium (LD)” [“patterns of” could be deleted]
* “This list of regions was generated on the basis of an extensive literature review.” [This sentence could be deleted. The numerous references in the table show that you have scoured the literature].
* “Also available for download …” [I suggest converting this sentence fragment into a complete sentence by replacing “Also available” with “This list is available”]

P. 8, “We used bcftools to remove indels and otherwise restrict our analyses to biallelic single nucleotide variants (SNVs).”

* Consider removing “remove indels and otherwise” so that the sentence reads “We used bcftools to restrict our analyses to biallelic single nucleotide variants (SNVs).”

P. 8 “which implements the iterative procedure recommended in 54”.

* There are a couple of ways to fix the superscript preceded by a space. You could change “54” to “reference 54” (with 54 as normal text, not superscripted), but I think it would be better to replace “implements the iterative procedure recommended in 54” with “implements the Conomos et. al iterative procedure54”

P. 10. The “T” symbol for matrix transpose looks unusual to my eye. I’m used to seeing a Transpose symbol that has a thicker, serif font (as found, for example, in <https://testbook.com/learn/maths-transpose-of-a-matrix/>). In my experience, journals like the AJHG do not change the fonts in math equations when they typeset the article – it is up to the author to choose the math font.

P. 10, typo in “gentoypes”

P. 10, Consider replacing “Here we focus on results using an …” with “Here we use an …”

P. 10, I’m not sure what “literature-based filters” means. If you are excluding regions of unusual LD identified in Table 1, you could say that directly.

General comment. There are several URLs embedded in the main text. Before submitting the manuscript, I suggest checking whether the journal’s instructions to authors explicitly permit or prohibit URLs in the main text. If URLs are prohibited, they can be included in the electronic resources section.

P. 11, “We identified the 10 variants on each chromosome with the highest absolute SNP loading for each of the first four PCs.”

* I don’t understand this sentence as it is currently written. I can make sense of it if there 10 variants selected for each of the first four PCs, but not if 10 variants are selected in total (because I had thought there was one loading value per (PC, SNP) pair).

P. 13, Is $\gamma$ in the displayed equation a vector when $w\_i$ is a vector. If so, should “$\gamma w\_i$” be a dot product? If the current notation in the displayed equation (without a dot product) is a standard notation, then ignore this comment.

P. 14, “This ancestral heterogeneity motivates the need to carefully adjust for global ancestry in genome-wide association studies in these, just as in other, admixed samples.”

* I don’t think the verb “motivate” fits with the direct object “the need”. Perhaps replace “motivate” with “shows”.
* The word “carefully” could be deleted.
* “, just as in other,” could be deleted.

P. 15, “As we see in Figure S1, in African American samples the first principal component seems to be capturing global ancestry, whereas later PCs are not.”

* Could “seems to be capturing” be replaced with the more precise “is highly correlated with”?

P. 15, “This is evident upon inspection of SNP loadings, which represent the contribution of each variant to each principal component, or in investigating the correlation between principal component scores and the original genotypes.”

* Consider using parallel wording in the two clauses (e.g. changing “upon inspection of” to “from inspection of” and changing “in investigating” to “from investigation of”).

P. 16, “The peaks in these genotype-PC correlation plots indicate that those principal components are primarily capturing variation at a handful of positions along the genome…”

* I was surprised to see this described as a “handful of positions”, because elevated correlation is seen across most of some chromosomes in Figure 2 (unless I am misunderstanding something).

P. 18, “Similar patterns are observed in JHS and COPDGene.”

* If these results are not included in the paper, it can be helpful to include a “(data not shown)” at the end of the sentence.

P. 20, I was puzzled how equations (2) and (3) were derived until I read the last paragraph in Results on p. 26. Consider pointing the reader to the Supplemental Information sections that give the derivations for equations (2) and (3) at the point where each equation is introduced.

P. 26, “Our theoretical work (Equation 2) explicitly demonstrates the factors that impact the magnitude of the bias incurred by GWAS models that fail to adjust for global ancestry, …”

* The first independent clause could be condensed. For example, “Our theoretical work (Equation 2) identifies the factors that determine the bias incurred by GWAS models that do not adjust for global ancestry, …”
* The full sentence is rather long. You could consider breaking the two independent causes into two sentences.

General Comment: There are some places where I think it would be clearer if “capture” were replaced with the more precise “is correlated with”. One of these instances is on p. 27: “In particular, if a PC captures the genotype of multiple variants, …”

P. 28, Figure 7 legend. The symbol “y” in the figure is not defined in the legend.

P. 30, “that a stricter threshold (e.g., r 2 = 0.1) is needed for LD pruning in admixed populations than is often suggested in the literature (r 2 = 0.2):”

* The end of this sentence could be rewritten as “than the r2=0.2 threshold that is often suggested in the literature:” or as “than the commonly-used r2=0.2 threshold:”

P. 30, “… we suggest that GWAS models adjust for model-based estimates of global ancestry rather than principal components”

* It looks like the best results in Figure 5 are obtained with 4 PCs combined with strict LD pruning. I don’t know if I’m misinterpreting Figure 5, or if something more needs to be said here.

P. 31. “WHI SHARe genotype data and TOPMed whole genome sequence data are available for analysis upon request and application. Visit study sites and dbGaP for more information.”

* I suggest saying that the data is available from dbGaP and give the database accession numbers. The journal can add hyperlinks to the web pages for the accession numbers.