

Genetic ancestry and population structure in the All of Us Research Program cohort

Corresponding Author: Dr I. King Jordan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Recommend: accept w/ revisions

Sharma et al. present an analysis of the distribution of race, ethnicity, and genetic ancestry from the All of Us research program. They find an overlap between SIRE and genetic ancestry and a correlation between genetic ancestry and geographic location. Additionally, they find that the diversity of ancestry is associated with age. Overall, I thought the authors presented a thoughtful analysis of the cohort and found interesting associations, especially the association with age. Technical-wise, I have mostly minor comments. However, I would like to see some additional analyses that look more in-depth at some of these presented results (see below).

Major:

- Can the authors do an analysis regarding the similarities/differences between results obtained by UMAP and PCA. Given that the methods model non-linear and linear patterns, it would be an interesting comparison.

- Could the UMAP analyses also be performed on subcontinental ancestry groups as well and then compared with the PCA results? Again, this would be an interesting comparison that I have not seen before.

- Under Supervised genetic ancestry inference, "Participants with 100% African, American, Asian, or European continental ancestry were taken for subcontinental ancestry inference." Please emphasize that all of these estimates are "inferred" ancestry. Additionally, could the subcontinental ancestry analyses be performed within each continental PCA cluster? The thresholding of 100% seems too stringent, especially considering the inherent genetic diversity of the United States.

- Subcontinental analyses for East Asia and South Asia should be separate. The reference datasets contain many samples from individual regions in East Asia, but these are not at all examined.

- I am uncertain about the use of explicitly attaching genetic ancestry estimates to zip code locations, especially when presented in a map format as in Figure 5. Connecting zip codes, which are relatively fine-grain geographic locations, with biological ancestry draws concern given the historical and even recent events of minority populations being targeted by geographic regions, such as historically Black neighborhoods. I understand that the Census has very similar visualizations and information, but I worry that this could further inflate the concepts of race, ethnicity, and genetic ancestry and cause harm given the biological basis of genetic ancestry.

Also, by using admixture proportions in this representation, showing states with 0% or 100% is not accurate since there are likely individuals within these regions with genetic ancestry from these populations, even if they aren't samples. I suggest, breaking the gradient into deciles and the bottom decile not being white.

There are very interesting patterns regarding the distribution of genetic ancestry within states (i.e. by zip codes), but unless the authors are going to thoroughly present a finding about this, I think it does not need to be represented in the map. The map can be shaded by state level. Admixture plots by state would also communicate the findings regarding genetic diversity at the state level.

Additionally, how many ppl are there sampled per county? Unless it's in the hundreds, I would be very apprehensive about stating that this is representative of an entire county; also given the non-random sampling of individuals.

Provide an enlarged insert for Hawaii since the current visualization obscures most information

Could maps also be provided for all continental populations (i.e. East Asia and South Asia) for completeness.

Minor:

- Could the authors provide a brief description of the datasets used in addition to 1KGP and HGDP, especially ones that were used specifically to obtain references for subcontinental populations since these datasets are less well-known.

- For the subcontinental analysis for African ancestry, I'd remove the European ancestry references, as these individuals are not of African ancestry. Because of this, it also looks like PCA is capturing mostly the variance between continental ancestry as opposed to fine-scale African ancestry.

- Is there a reason why zip code data wasn't available for the NE states? I'd recommend analyses be done at the state level instead so as to not exclude such a large portion of the US population in the results and visualizations.

- The authors mention that the race and ethnicity groupings are OMB standards. Can the authors provide a discussion about the potential shortcomings of such limited categorizations. I think that given that All of Us is also a government initiative, a discussion on this is especially prudent.

- The authors should also emphasize how all of these estimates are conditional on the choice of reference panels, methods, etc.

Reviewer #2

(Remarks to the Author)

Overall an interesting and important contribution to the literature, clearly differentiating SIRE and genetic ancestry and describing their distribution in All of Us. I do think there are some points that need to be clarified/added to improve the paper however.

As discussed in NASEM report, likely important to state, even in the abstract, how ancestry is defined (i.e. based on similarity to reference panels 1000G/HGDP).

While this is true- "shared ancestry is typically inferred by some measure of genetic similarity among individuals"¹¹ – the labelling of such clusters is still influenced by social factors (such as use of continental ancestry boundaries, etc) in how reference panels to which a new cohort is compared are defined (for example, in much work deciding that Caribbean and self-identified African American populations within the US belong in the AFR cluster for 1000G- though maybe those groups were excluded here based on Supplementary Table 2). This statement needs to be better contextualized.

Could the downsampling performed for initial PC derivation impact the number of clusters identified in All of Us? I also think the authors should be clearer that number of clusters may well change with future releases. I find figure 2D confusing- I think the clusters need to be presented color coded both by cluster identification from part C and by SIRE (two part figure)- and which color corresponds to which SIRE group should be clarified (use of blue for Hispanic or Latino and for Black or African American is confusing). Also how are individuals who noted more than one SIRE identification listed? Not clear as currently presented.

How were the individuals chosen for this analysis "A similar approach was used to infer subcontinental ancestry for All of Us participants with high levels of African (n=2,934), American (n=235), Asian (n=2,652), and European ancestry (n=73,150; Figure 4). " Why so few African, American, and Asian ancestry individuals? How was "high" defined? I think readers should be further cautioned in the discussion that the conclusions in Figure 4 for American ancestry individuals in particular are based on very small numbers, and that no individuals with Native American self-identification were included in All of Us at this time.

Also, why was the GBR 1KGP grouping included in the African reference panel? Also which All of Us participants are included in Supplementary Figures 5-8- I think just those with high levels of ancestry from one continental population? (similar questions for European grouping in Supplementary Figure 5- I guess this is just showing individuals with "high" African ancestry from All of Us, and European cluster just included for reference? More labelling is needed- it should also be clarified if these are the "cross-All of Us" PCs (I think yes) or ones derived within each "high" ancestry group for these supplemental figures- better/more complete labelling needed).

"Genetic ancestry can interact with the environment to affect health outcomes in a non-additive way"¹⁶⁻¹⁹ This is actually quite a strong claim to my mind, and I'm not sure that it is adequately proven in the papers you cite... not only are you claiming that both genetic ancestry (due to tagging of ancestry differentiated variants which are either of very large effect or all in the same direction of effect) and environment both associated with health outcomes independently (which I think is fair based on cited papers), you are stating definitively that there are statistically significant interactions between them, which I think are less known – gene environment interactions have been hard to identify and replicate in any population, and sample sizes for global ancestry proportion x environment interaction analyses will often be quite small. The association of global ancestry proportions with health outcomes is often a bit inconsistent across cohorts and can be confounded by social factors

(such as discrimination based on skin color, separate from any true disease causal variants with inconsistent frequency across populations). I think this section should be revised/removed.

In figure 3c- why aren't all these ancestry components represented? African, American, East Asian, European, Oceanian, South Asian, and West Asian . Particularly confusing for Asian ancestry panels.

I think the relationship between the data presented here and the ancestry cluster identification information described in https://support.researchallofus.org/hc/en-us/article_attachments/19370367115796 in the All of Us research portal needs to be described. For example, the documentation describes use of an "other" ancestry cluster, which is not described here. In Figure 5- do you mean "Native American (teal)" or "American (teal)"? The second would be more consistent with terminology in the rest of the paper.

I'm not sure I agree with the second part of this statement- I worry this could be taken as inferring that, on the whole, it is ok to assume that SIRE groups generally have the same pattern of genetic ancestry and that SIRE is an ok proxy for genetic ancestry at the individual level, which I think is almost certainly not what the authors mean. "tend to show group-specific patterns of genetic ancestry, which may be relevant to the diagnosis and treatment of disease". Would clarify/reword. The authors should cite/describe differences between their methods and those presented in <https://jamanetwork.com/journals/jamacardiology/fullarticle/2808413>.

Reviewer #3

(Remarks to the Author)

The paper should be a good additional to the literature in order to understand the diversity of the All of Us dataset. I did not identify flaws in the analysis, interpretation and conclusion, but suggested the authors consider speaking to how the use of race/ancestry is affected by the recent NASEM Population Descriptors report. Specific suggestions below.

Lines

49-50: Suggest explaining why Eurocentric bias exacerbate health disparities.

60: Suggest adding "self-reported"

61: Suggest adding characteristic "derived from" the genome

71-72: Suggest speaking to how genetic ancestry is associated with phenotype and thus perceived race/ethnicity which also affects health outcomes

107: Suggest defining PMI: Skip

108: Suggest explaining why data is currently unavailable for AI/AN

161-162: With zip code data not available for states listed, how much of the cohort is affected?

190-191: Sentence confusing -- "and participants"?

216: "there are" is repeated

262: Again suggest adding "derived from"

263: Suggest mentioning NASEM report on Population Descriptors and how it affects the use of race/ethnicity

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Overall, I am satisfied with the revisions provided by the authors. Additionally, I appreciate the authors' extensive re-analysis and revision of Figure 4.

A few points that would benefit the clarity of the manuscript:

- Could a colorbar be added to the heatmap plots, like those in Sup Figure 2C?

- For the subcontinental PCA/UMAP analyses, was dimensionality reduction performed on the entire cohort and then sub-populations are just highlighted? Or was dim reduction performed ONLY on participants within a given inferred continental ancestry group? Either is fine, but it would be helpful to clarify the exact steps.

- On the figure title for Sup Figure 5, what does "coherent" mean in this context? Perhaps a different word would be more clear.

Reviewer #4

(Remarks to the Author)

Comment #2.1

Overall, the authors do a good job at addressing the comments of the previous reviewer (Reviewer 2) and have performed multiple re-analyses that have improved the paper. The narrower focus on population structure and ancestry versus adding ethnicity into the mix is probably a good choice, given the ease of which this complex field can be misappropriated. The paper is short - which makes it a nice and easy read.

There are, however, a few comments that I believe could be addressed to balance this brevity with a bit more clarity to improve interpretability of the current results and especially to clarify methodological and conceptual challenges posed by these kinds of analyses.

Comment #2.2

Lines 50-61. The statement on the “Conceptual, mechanistic, and operational” definitions of genetic ancestry is a positive addition. A few points, however.

“Conceptually, genetic ancestry reflects the geographic origins of an individual’s ancestors” – I’d argue that this is perhaps not a useful way of describing how “genetic ancestry” is being conceptualised here or that this is broadly agreed upon in the field. Even within the paper, ethno-linguistic labels used in tandem with geography for proxy ancestral populations like “Bantu” and “Han”. Human populations have moved a lot over time, especially recently, so you will likely have ancestors that lived in various locations depending on what slice of time you are looking at and at what resolution.

Broadly, in this method, you are making an indirect inference about the sharing of more recent genetic ancestors between an individual and some ethno-linguistic or geographic group relative to some others. Whilst this can certainly be useful for some downstream applications, is not necessarily informative about the geographic origins of a person’s ancestors as a whole. A few tweaks for clarity would probably minimise this potential for misunderstanding.

Comment #2.3

When using the term “surrogates for ancestral populations” I believe it is worth noting (e.g. in Note on ancestry inference or in the introduction) that discrete ancestral populations don’t (and never have) existed. Population boundaries both through time and in the present-day are blurry and genetic ancestry itself doesn’t neatly map onto these PCA clusters or labelled reference populations.

Whilst the fact that “ancestral source populations” model a more complex reality may be obvious to more informed readers, the All Of Us project has a broad user-base, so ensuring this point is drilled home is appropriate I think.

Comment #2.4

Following from the above, convex “ancestry components” as output by the Rye method may require some more careful interpretation.

For example: “All groups show evidence of admixture with multiple ancestry components present in different proportions.” is perhaps a reasonable interpretation in some cases (e.g. this is widely described for African Americans). However, could some of these results not also simply be explained by incomplete sampling of reference populations + spatial genetic structure?

Imagine performing this inference on a member of a population roughly equally dissimilar (central) to multiple reference pops simply through genetic drift / isolation by distance. When you force discretisation, I’d wager you will always see evidence of such “admixture”, which could then just disappear if the reference populations changed or another was added e.g. should we really just interpret cluster 5 as simply the product of recent admixture between Europeans, South Asian and West Asian populations? Is there any explanation for this that maps to historic events? It seems to me that alternatives (such a spatial population structure) could also explain these results.

These issues surrounding potential for misinterpretation should, in my opinion, be discussed more explicitly or even tested for / demonstrated in Supplementary Materials (if possible!) e.g. by simply removing one or more reference populations from a given analysis. Given the method is described as “accurate” elsewhere (line 133; I’m not sure how this can be discerned necessarily unless through simulations for example), I’d recommend instead taking the chance to emphasise what the outputs of the methods are actually telling us, and where modelling things using incomplete / discrete source populations may break down.

Comment #2.5

As an aside to the previous comment, apologies as I didn’t see the initial draft, but almost all individuals in e.g. with any apparent “African” ancestries in Figure 2 also have some evidence of “European” ancestries, and European reference pops are now not included (it appears they may have been before?). Given this, is the ubiquitous “East Bantu” component in Figure 3a just leaking in as the “closest proxy” for European ancestries here?

Given the apparent lack of any evidence for Eastern Bantu related ancestry components in the African American population in the USA in previous research (e.g. see Ongaro et al. 2019, The Genomic Impact of European Colonisation of the Americas), this is an unexpected result that may demonstrate another example of the challenges of reference population curation and potential issues of interpretability described above.

Version 3:

Reviewer comments:

Reviewer #4

(Remarks to the Author)

The authors have done a great job at responding to my questions and adding new analyses quickly and thoughtfully.

I believe the paper should now be accepted for publication.

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/>

Note: redo everything for V7

Reviewer #1 (Remarks to the Author):

Recommend: accept w/ revisions

Comment #1.1: Sharma et al. present an analysis of the distribution of race, ethnicity, and genetic ancestry from the All of Us research program. They find an overlap between SIRE and genetic ancestry and a correlation between genetic ancestry and geographic location. Additionally, they find that the diversity of ancestry is associated with age. Overall, I thought the authors presented a thoughtful analysis of the cohort and found interesting associations, especially the association with age. Technical-wise, I have mostly minor comments. However, I would like to see some additional analyses that look more in-depth at some of these presented results (see below).

Response #1.1: We appreciate the reviewer's positive comments on our manuscript and the constructive suggestions for additional analyses. We have performed all of the additional analyses and revised our manuscript accordingly. We describe each of the analyses we perform and where the new results are presented in the responses below.

Major:

Comment #1.2: Can the authors do an analysis regarding the similarities/differences between results obtained by UMAP and PCA. Given that the methods model non-linear and linear patterns, it would be an interesting comparison.

Response #1.2: As requested, we performed an analysis on the similarities/differences between results obtained from UMAP and PCA. Overall, the results are quite similar (see Supplementary Figure 2). Density-based clustering of UMAP is more granular and thus produces twice as many clusters (14) compared to density-based clustering of PCA (7 clusters; see Figure 2B). Nevertheless, there is a high correspondence between cluster membership for the two methods with respect to both the percentage of participant overlap in the clusters and the SIRE percentages within clusters (see Supplementary Figure 2 and Figure 2D).

Comment #1.3: Could the UMAP analyses also be performed on subcontinental ancestry groups as well and then compared with the PCA results? Again, this would be an interesting comparison that I have not seen before.

Response #1.3: As suggested, we performed UMAP on the same set of global reference population samples and All of Us participant samples used for subcontinental ancestry inference. By way of comparison, the PCA results for these samples, stratified by continental ancestry, are shown in Supplementary Figure 6, and the UMAP results are shown in Supplementary Figure 7.

Comment #1.4: Under Supervised genetic ancestry inference, "Participants with 100% African, American, Asian, or European continental ancestry were taken for subcontinental ancestry inference." Please emphasize that all of these estimates are "inferred" ancestry. Additionally, could the subcontinental ancestry analyses be performed within each continental PCA cluster? The thresholding of 100% seems too stringent, especially considering the inherent genetic diversity of the United States.

Response #1.4: As suggested, we made our approach to identifying clusters of individuals for subcontinental ancestry inference less stringent using PCA clusters. The new PCA-based clusters for All of Us participants, and their relationships to global reference populations, are shown in Supplementary Figure 5. This resulted in larger and more diverse clusters for downstream analysis, so the reviewer's suggestion is much appreciated! We now include 9,291 African, 804 American, 2,457 East Asian, 26,440 European, and 2,792 South Asian individuals in our subcontinental ancestry inference (see Figure 4).

Comment #1.5: Subcontinental analyses for East Asia and South Asia should be separate. The reference datasets contain many samples from individual regions in East Asia, but these are not at all examined.

Response #1.5: As suggested, we performed subcontinental ancestry inference separately on East Asian and South Asian participant clusters (see Figure 5 and Supplementary Figures 5-7).

Comment #1.6: I am uncertain about the use of explicitly attaching genetic ancestry estimates to zip code locations, especially when presented in a map format as in Figure 5. Connecting zip codes, which are relatively fine-grain geographic locations, with biological ancestry draws concern given the historical and even recent events of minority populations being targeted by geographic regions, such as historically Black neighborhoods. I understand that the Census has very similar visualizations and information, but I worry that this could further inflate the concepts of race, ethnicity, and genetic ancestry and cause harm given the biological basis of genetic ancestry.

Response #1.6: To avoid this issue, we have redone the geographic analysis of genetic ancestry at the state level rather than the more fine-grained zip code level (see Figure 5).

Comment #1.7: Also, by using admixture proportions in this representation, showing states with 0% or 100% is not accurate since there are likely individuals within these regions with genetic ancestry from these populations, even if they aren't samples. I suggest, breaking the gradient into deciles and the bottom decile not being white.

Response #1.7: We agree with this important point. This issue is resolved with the new state level analysis shown in Figure 5. It should also be noted that we have substantially higher nationwide coverage given that we analyzed a new All of Us data release (version 7) with 297,549 participants compared to 165,127 participants (version 6) in the previous analysis (see Table 1).

Comment #1.8: There are very interesting patterns regarding the distribution of genetic ancestry within states (i.e. by zip codes), but unless the authors are going to thoroughly present a finding about this, I think it does not need to be represented in the map. The map can be shaded by state level. Admixture plots by state would also communicate the findings regarding genetic diversity at the state level.

Response #1.8: This is a great point! We re-did the analysis as suggested by the reviewer, using admixture plots by state, and it does indeed communicate the findings regarding ancestral diversity at the state level in a more clear and compelling way (see Figure 5).

Comment #1.9: Additionally, how many ppl are there sampled per county? Unless it's in the hundreds, I would be very apprehensive about stating that this is representative of an entire county; also given the non-random sampling of individuals.

Response #1.9: This valid point is addressed by the state level analysis suggested by the reviewer (see previous response and Figure 5).

Comment #1.10: Provide an enlarged insert for Hawaii since the current visualization obscures most information

Response #1.11: We now provide admixture plots for each state, as suggested by the reviewer, including Hawaii. We use a stylized representation of the US map (the geographical grid map commonly used for graphics of this kind), which scales all states the same size and orders them approximately by geography (see Figure 5).

Comment #1.11: Could maps also be provided for all continental populations (i.e. East Asia and South Asia) for completeness.

Response #1.11: As suggested, we now provide admixture plots for each state, including all ancestry components. Figure 5A shows ancestry components for all continental reference population groups for each state and Figure 5B shows ancestry proportions for the non-European continental reference population groups for each state.

Minor:

Comment #1.12: Could the authors provide a brief description of the datasets used in addition to 1KGP and HGDP, especially ones that were used specifically to obtain references for subcontinental populations since these datasets are less well-known.

Response #1.12: As suggested, we provide a brief description of the datasets used in addition to 1KGP and HGDP (see Methods section on supervised genetic ancestry inference).

Comment #1.13: For the subcontinental analysis for African ancestry, I'd remove the European ancestry references, as these individuals are not of African ancestry. Because of this, it also looks like PCA is capturing mostly the variance between continental ancestry as opposed to fine-scale African ancestry.

Response #1.13: We have re-done the subcontinental ancestry analysis as suggested by the reviewer (see comments and responses #1.3-#1.5 above). Since we have lowered the percent ancestry threshold to include more samples, using PCA instead as suggested by the reviewer, the samples are more admixed and thus we need to include the minor component ancestry fraction samples in the PCA (e.g. the European ancestry references in the African subcontinental group) in order to ensure accurate ancestry inference. As suggested, we have removed the European component ancestry fraction samples from the African PCA shown in Supplementary Figure 6 for visual clarity (ie to emphasize the within group PC variance that is used in subcontinental ancestry inference).

Comment #1.14: Is there a reason why zip code data wasn't available for the NE states? I'd recommend analyses be done at the state level instead so as to not exclude such a large portion of the US population in the results and visualizations.

Response #1.14: Those zip code data for the NE states were not available in the version 6 All of Us release. We have re-done our analysis using the version 7 All of Us release. The NE zip code data are

available and the version 7 release, but as suggested by this reviewer, we now show ancestry at the state level as opposed to the zip code level (see comments and responses #1.6-#1.11 above and Figure 5).

Comment #1.15: The authors mention that the race and ethnicity groupings are OMB standards. Can the authors provide a discussion about the potential shortcomings of such limited categorizations. I think that given that All of Us is also a government initiative, a discussion on this is especially prudent.

Response #1.15: The All of Us participant survey includes questions that allow for far more detailed information on population origins within each OMB category. These data will be made available in a future All of Us release, which will allow for more fine-grained analysis of cohort diversity. As suggested, we provide a discussion of this (see Methods section on the “All of Us participant cohort”).

Comment #1.16: The authors should also emphasize how all of these estimates are conditional on the choice of reference panels, methods, etc.

Response #1.16: We have added a new section to the Methods to emphasize all of these important caveats (see Methods section “Note on genetic ancestry inference”).

Reviewer #2 (Remarks to the Author):

Comment #2.1: Overall an interesting and important contribution to the literature, clearly differentiating SIRE and genetic ancestry and describing their distribution in All of Us. I do think there are some points that need to be clarified/added to improve the paper however.

Response #2.1: We appreciate the reviewer’s positive impression of our paper. We address all of the reviewer’s individual points below in an effort to clarify the presentation of our results.

Comment #2.2: As discussed in NASEM report, likely important to state, even in the abstract, how ancestry is defined (i.e. based on similarity to reference panels 1000G/HGDP).

Response #2.2: As suggested, we have added a new section to the Methods, following the suggestions of the NASEM report, to state exactly how ancestry is defined (see Methods section “Note on genetic ancestry inference”).

Comment #2.3: While this is true- “shared ancestry is typically inferred by some measure of genetic similarity among individuals¹¹”. – the labelling of such clusters is still influenced by social factors (such as use of continental ancestry boundaries, etc) in how reference panels to which a new cohort is compared are defined (for example, in much work deciding that Caribbean and self-identified African American populations within the US belong in the AFR cluster for 1000G- though maybe those groups were excluded here based on Supplementary Table 2). This statement needs to be better contextualized.

Response #2.3: We agree that it is important to better contextualize how genetic ancestry is inferred, and we thank the reviewer for pointing this out. We have added a new section to the Methods to explain how genetic ancestry is defined conceptually, mechanistically, and operationally. In this section, we clarify that “The genetic ancestry results reported here are contingent upon the choice of reference

populations, how these reference populations are defined, and the method used to infer genetic similarity between All of Us participants and the reference population samples” (see Methods section “Note on genetic ancestry inference”).

Comment #2.4: Could the downsampling performed for initial PC derivation impact the number of clusters identified in All of Us? I also think the authors should be clearer that number of clusters may well change with future releases.

Response #2.4: We eliminated the downsampling analysis for the revised manuscript and now perform density based clustering of all PCA data using HBDSCAN (see Methods section on unsupervised clustering and Figure 2B). As suggested, we state that “The number of density-based clusters could change with future participant data releases.” (see Results section “Unsupervised: genetic diversity clusters”).

Comment #2.5: I find figure 2D confusing- I think the clusters need to be presented color coded both by cluster identification from part C and by SIRE (two part figure)- and which color corresponds to which SIRE group should be clarified (use of blue for Hispanic or Latino and for Black or African American is confusing). Also how are individuals who noted more than one SIRE identification listed? Not clear as currently presented.

Response #2.5: We completely re-did Figure 2. The color coding by SIRE is now consistent in panels B-D. Individuals who noted more than one SIRE identification are labeled “More than one population” as indicated in the color legends for Figure 2 and Figure 3 and described in the Methods.

Comment #2.6: How were the individuals chosen for this analysis “A similar approach was used to infer subcontinental ancestry for All of Us participants with high levels of African (n=2,934), American (n=235), Asian (n=2,652), and European ancestry (n=73,150; Figure 4). “ Why so few African, American, and Asian ancestry individuals? How was “high” defined? I think readers should be further cautioned in the discussion that the conclusions in Figure 4 for American ancestry individuals in particular are based on very small numbers, and that no individuals with Native American self-identification were included in All of Us at this time.

Response #2.6: As suggested by reviewer #1 (see comments and responses #1.3-#1.5 above), we have changed the criteria by which we select individuals for subcontinental ancestry inference. We now use clusters defined by PCA in an effort to include more individuals per ancestry group, and to separate out East Asian and South Asian groups. The new groups are shown in Supplementary Figure 5 and the results of the new subcontinental ancestry inference are shown in Figure 4. This approach resulted in the inclusion of more individuals within each group: 9,291 African, 804 American, 2,457 East Asian, 26,440 European, and 2,792 South Asian. As suggested, we added that “Results on American subcontinental ancestry should be interpreted with caution given that the data analyzed here do not include participants who identified as American Indian or Alaska Native” (see Results section “Supervised: genetic ancestry inference”).

Comment #2.7: Also, why was the GBR 1KGP grouping included in the African reference panel? Also which All of Us participants are included in Supplementary Figures 5-8- I think just those with high levels of ancestry from one continental population? (similar questions for European grouping in Supplementary Figure 5- I guess this is just showing individuals with “high” African ancestry from All of Us, and European cluster just included for reference? More labelling is needed- it should also be clarified if these are the

“cross-All of Us” PCs (I think yes) or ones derived within each “high” ancestry group for these supplemental figures- better/more complete labelling needed).

Response #2.7: We have re-done the subcontinental ancestry analysis as suggested by reviewer #1 (see comments and responses #1.3-#1.5 above). Since we have lowered the percent ancestry threshold to include more samples, using PCA instead as suggested by the reviewer, the samples are more admixed and thus we need to include the minor component ancestry fraction samples in the PCA (e.g. the European ancestry references in the African subcontinental group) in order to ensure accurate ancestry inference. As suggested, we have removed the European component ancestry fraction samples from the African PCA shown in Supplementary Figure 6 for visual clarity (ie to emphasize the within group PC variance that is used in subcontinental ancestry inference). We have clarified which All of Us individuals were used for this analysis by visualizing them in PCA compared to continental reference populations (see Supplementary Figures 5-7).

Comment #2.8: “Genetic ancestry can interact with the environment to affect health outcomes in a non-additive way¹⁶⁻¹⁹ This is actually quite a strong claim to my mind, and I’m not sure that it is adequately proven in the papers you cite... not only are you claiming that both genetic ancestry (due to tagging of ancestry differentiated variants which are either of very large effect or all in the same direction of effect) and environment both associated with health outcomes independently (which I think is fair based on cited papers), you are stating definitively that there are statistically significant interactions between them, which I think are less known – gene environment interactions have been hard to identify and replicate in any population, and sample sizes for global ancestry proportion x environment interaction analyses will often be quite small. The association of global ancestry proportions with health outcomes is often a bit inconsistent across cohorts and can be confounded by social factors (such as discrimination based on skin color, separate from any true disease causal variants with inconsistent frequency across populations). I think this section should be revised/removed.

Response #2.8: We removed this statement as suggested.

Comment #2.9: In figure 3c- why aren’t all these ancestry components represented? African, American, East Asian, European, Oceanian, South Asian, and West Asian. Particularly confusing for Asian ancestry panels.

Response #2.9: As suggested, we re-did Figure 3 to show all the ancestry components: African, American, East Asian, European, Oceanian, South Asian, and West Asian.

Comment #2.10: I think the relationship between the data presented here and the ancestry cluster identification information described in https://support.researchallofus.org/hc/en-us/article_attachments/19370367115796 in the All of Us research portal needs to be described. For example, the documentation describes use of an “other” ancestry cluster, which is not described here.

Response #2.10: As suggested, we compare the continuous ancestry percentages that we infer in our manuscript to the categorical ancestry predictions provided on All of Us Research Workbench. We provide a description of this new analysis in the Results section Supervised: genetic ancestry inference and the results are shown in Supplementary Figure 5. The categorical ancestry group predictions provided on the Researcher Workbench were inferred independently of our own approach, as discussed in the manuscript, but there is a high correspondence between group membership and percent continental ancestry.

Comment #2.11: In Figure 5- do you mean “Native American (teal)” or “American (teal)”? The second would be more consistent with terminology in the rest of the paper.

Response #2.11: We have corrected this to “American”. Thanks for catching that mistake.

Comment #2.12: I’m not sure I agree with the second part of this statement- I worry this could be taken as inferring that, on the whole, it is ok to assume that SIRE groups generally have the same pattern of genetic ancestry and that SIRE is an ok proxy for genetic ancestry at the individual level, which I think is almost certainly not what the authors mean. “tend to show group-specific patterns of genetic ancestry, which may be relevant to the diagnosis and treatment of disease”. Would clarify/reword.

Response #2.12: We reworded this to state “tend to show group-specific patterns of genetic ancestry consistent with shared geographical origins.” This follows from the conceptual definition of genetic ancestry that we provided in the new “Note on genetic ancestry inference” Methods section.

Comment #2.13: The authors should cite/describe differences between their methods and those presented in <https://jamanetwork.com/journals/jamacardiology/fullarticle/2808413>.

Response #2.13: We cite this relevant paper as suggested. The paper is a very short report that provides only a visual comparison of PCA data stratified by SIRE, rather than any quantitative analysis of genetic diversity, clustering, or ancestry. The visual evaluation of PCA data did not allow for any conclusion about how the data are clustered.

Reviewer #3 (Remarks to the Author):

Comment #3.1: The paper should be a good additional to the literature in order to understand the diversity of the All of Us dataset. I did not identify flaws in the analysis, interpretation and conclusion, but suggested the authors consider speaking to how the use of race/ancestry is affected by the recent NASEM Population Descriptors report. Specific suggestions below.

Response #3.1: We appreciate the reviewer’s positive response to our manuscript. We have added a new Methods section “Note on genetic ancestry inference” to elaborate on how our use of race/ancestry is affected by the recent NASEM Population Descriptors report.

Lines

Comment #3.2: 49-50: Suggest explaining why Eurocentric bias exacerbate health disparities. **Response #3.2:** Done.

Comment #3.3: 60: Suggest adding "self-reported". **Response #3.3:** Done

Comment #3.4: 61: Suggest adding characteristic "derived from" the genome. **Response #3.4:** Done.

Comment #3.5: 71-72: Suggest speaking to how genetic ancestry is associated with phenotype and thus perceived race/ethnicity which also affects health outcomes. **Response #3.5:** We have added a sentence stating this.

Comment #3.6: 107: Suggest defining PMI: Skip. **Response #3.6:** Done.

Comment #3.7: 108: Suggest explaining why data is currently unavailable for AI/AN

Response #3.7: The All of Us program recognizes that Tribes in the United States are sovereign nations, and some Tribes have concerns about taking part in research studies. All of Us has been engaged in a Tribal consultation process to learn about their views and to ensure that research using the program's biospecimens and data from Tribal members is done in a way that is respectful of applicable Tribal customs, culture, and laws. A primary concern of Tribal leaders has been minimizing the risk of harm to Tribal communities once data from self-identified AI/AN participants is shared with researchers. In response to these concerns, the program paused any work proposing to share data volunteered by participants self-identifying as AI/AN in order to consult on the best path forward. See <https://allofus.nih.gov/about/diversity-and-inclusion/tribal-engagement>

Comment #3.8: 161-162: With zip code data not available for states listed, how much of the cohort is affected?

Response #3.8: We no longer use zip code data, and instead show ancestry/admixture results on a state-by-state level (see responses to Comments #1.6-#1.11 from reviewer #1 and new figure 5).

Comment #3.9: 190-191: Sentence confusing -- "and participants"? **Response #3.9:** Typo fixed.

Comment #3.10: 216: "there are" is repeated. **Response #3.10:** Typo fixed.

Comment #3.11: 262: Again suggest adding "derived from". **Response #3.11:** Fixed.

Comment #3.12: 263: Suggest mentioning NASEM report on Population Descriptors and how it affects the use of race/ethnicity.

Response #3.12: As suggested, we have added a new Methods section "Note on genetic ancestry inference" to elaborate on how our use of race/ancestry is affected by the recent NASEM Population Descriptors report.

Reviewer #1 (Remarks to the Author):

Recommend: accept w/ revisions

Comment #1.1: Sharma et al. present an analysis of the distribution of race, ethnicity, and genetic ancestry from the All of Us research program. They find an overlap between SIRE and genetic ancestry and a correlation between genetic ancestry and geographic location. Additionally, they find that the diversity of ancestry is associated with age. Overall, I thought the authors presented a thoughtful analysis of the cohort and found interesting associations, especially the association with age. Technical-wise, I have mostly minor comments. However, I would like to see some additional analyses that look more in-depth at some of these presented results (see below).

Response #1.1: We appreciate the reviewer's positive comments on our manuscript and the constructive suggestions for additional analyses. We have performed all of the additional analyses and revised our manuscript accordingly. We describe each of the analyses we perform and where the new results are presented in the responses below.

Major:

Comment #1.2: Can the authors do an analysis regarding the similarities/differences between results obtained by UMAP and PCA. Given that the methods model non-linear and linear patterns, it would be an interesting comparison.

Response #1.2: As requested, we performed an analysis on the similarities/differences between results obtained from UMAP and PCA. Overall, the results are quite similar (see Supplementary Figure 2). Density-based clustering of UMAP is more granular and thus produces almost twice as many clusters (13) compared to density-based clustering of PCA (7 clusters; see Figure 1C). Nevertheless, there is a high correspondence between cluster membership for the two methods with respect to the percentage of participant overlap in the clusters (see Supplementary Figure 2C). [172-178]

Comment #1.3: Could the UMAP analyses also be performed on subcontinental ancestry groups as well and then compared with the PCA results? Again, this would be an interesting comparison that I have not seen before.

Response #1.3: As suggested, we performed UMAP on the same set of global reference population samples and All of Us participant samples used for subcontinental ancestry inference. By way of comparison, the PCA results for these samples, stratified by continental ancestry, are shown in Supplementary Figure 6, and the UMAP results are shown in Supplementary Figure 7.

Comment #1.4: Under Supervised genetic ancestry inference, "Participants with 100% African, American, Asian, or European continental ancestry were taken for subcontinental ancestry inference." Please emphasize that all of these estimates are "inferred" ancestry. Additionally, could the subcontinental ancestry analyses be performed within each continental PCA cluster? The thresholding of 100% seems too stringent, especially considering the inherent genetic diversity of the United States.

Response #1.4: As suggested, we made our approach to identifying clusters of individuals for subcontinental ancestry inference less stringent using PCA clusters. The new PCA-based clusters for All of Us participants, and their relationships to global reference populations, are shown in Supplementary

Figure 5. This resulted in larger and more diverse clusters for downstream analysis, so the reviewer's suggestion is much appreciated! We now include 9,291 African, 2,457 East Asian, 24,730 European, and 2,484 South Asian individuals in our subcontinental ancestry inference (see Figure 3). [119-123]

Comment #1.5: Subcontinental analyses for East Asia and South Asia should be separate. The reference datasets contain many samples from individual regions in East Asia, but these are not at all examined.

Response #1.5: As suggested, we performed subcontinental ancestry inference separately on East Asian and South Asian participant clusters (see Figure 3 and Supplementary Figures 5-7). [119-123; 217-219]

Comment #1.6: I am uncertain about the use of explicitly attaching genetic ancestry estimates to zip code locations, especially when presented in a map format as in Figure 5. Connecting zip codes, which are relatively fine-grain geographic locations, with biological ancestry draws concern given the historical and even recent events of minority populations being targeted by geographic regions, such as historically Black neighborhoods. I understand that the Census has very similar visualizations and information, but I worry that this could further inflate the concepts of race, ethnicity, and genetic ancestry and cause harm given the biological basis of genetic ancestry.

Response #1.6: To avoid this issue, we have redone the geographic analysis of genetic ancestry at the state level rather than the more fine-grained zip code level (see Figure 4). [223-228]

Comment #1.7: Also, by using admixture proportions in this representation, showing states with 0% or 100% is not accurate since there are likely individuals within these regions with genetic ancestry from these populations, even if they aren't samples. I suggest, breaking the gradient into deciles and the bottom decile not being white.

Response #1.7: We agree with this important point. This issue is resolved with the new state level analysis shown in Figure 4. It should also be noted that we have substantially higher nationwide coverage given that we analyzed a new All of Us data release (version 7) with 297,549 participants compared to 165,127 participants (version 6) in the previous analysis. [223-228]

Comment #1.8: There are very interesting patterns regarding the distribution of genetic ancestry within states (i.e. by zip codes), but unless the authors are going to thoroughly present a finding about this, I think it does not need to be represented in the map. The map can be shaded by state level. Admixture plots by state would also communicate the findings regarding genetic diversity at the state level.

Response #1.8: This is a great point! We re-did the analysis as suggested by the reviewer, using admixture plots by state, and it does indeed communicate the findings regarding ancestral diversity at the state level in a more clear and compelling way (see Figure 4). [223-228]

Comment #1.9: Additionally, how many ppl are there sampled per county? Unless it's in the hundreds, I would be very apprehensive about stating that this is representative of an entire county; also given the non-random sampling of individuals.

Response #1.9: This valid point is addressed by the state level analysis suggested by the reviewer (see previous response and Figure 4). [223-228]

Comment #1.10: Provide an enlarged insert for Hawaii since the current visualization obscures most information

Response #1.11: We now provide admixture plots for each state, as suggested by the reviewer, including Hawaii. We use a stylized representation of the US map (the geographical grid map commonly used for graphics of this kind), which scales all states the same size and orders them approximately by geography (see Figure 4).

Comment #1.11: Could maps also be provided for all continental populations (i.e. East Asia and South Asia) for completeness.

Response #1.11: As suggested, we now provide admixture plots for each state, including all ancestry components. Figure 4A shows ancestry components for all continental reference population groups for each state and Figure 4B shows ancestry proportions for the non-European continental reference population groups for each state.

Minor:

Comment #1.12: Could the authors provide a brief description of the datasets used in addition to 1KGP and HGDP, especially ones that were used specifically to obtain references for subcontinental populations since these datasets are less well-known.

Response #1.12: As suggested, we provide a brief description of the datasets used in addition to 1KGP and HGDP (see Methods section on supervised genetic ancestry inference). [127-129]

Comment #1.13: For the subcontinental analysis for African ancestry, I'd remove the European ancestry references, as these individuals are not of African ancestry. Because of this, it also looks like PCA is capturing mostly the variance between continental ancestry as opposed to fine-scale African ancestry.

Response #1.13: We have re-done the subcontinental ancestry analysis as suggested by the reviewer (see comments and responses #1.3-#1.5 above). Since we have lowered the percent ancestry threshold to include more samples, using PCA instead as suggested by the reviewer, the samples are more admixed and thus we need to include the minor component ancestry fraction samples in the PCA (e.g. the European ancestry references in the African subcontinental group) in order to ensure accurate ancestry inference. As suggested, we have removed the European component ancestry fraction samples from the African PCA shown in Supplementary Figure 6 for visual clarity (ie to emphasize the within group PC variance that is used in subcontinental ancestry inference). [119-123; 213-217]

Comment #1.14: Is there a reason why zip code data wasn't available for the NE states? I'd recommend analyses be done at the state level instead so as to not exclude such a large portion of the US population in the results and visualizations.

Response #1.14: Those zip code data for the NE states were not available in the version 6 All of Us release. We have re-done our analysis using the version 7 All of Us release. The NE zip code data are available and the version 7 release, but as suggested by this reviewer, we now show ancestry at the state level as opposed to the zip code level (see comments and responses #1.6-#1.11 above and Figure 4).

Comment #1.15: The authors mention that the race and ethnicity groupings are OMB standards. Can the authors provide a discussion about the potential shortcomings of such limited categorizations. I think that given that All of Us is also a government initiative, a discussion on this is especially prudent.

Response #1.15: Following the recommendation of the All of Us consortium, we have removed all analysis and comparison of race and ethnicity in the revised manuscript and focus exclusively on population structure and genetic ancestry (see new title). This entailed removing Figure 1 and stratifying groups by genetic similarity, rather than race/ethnicity.

Comment #1.16: The authors should also emphasize how all of these estimates are conditional on the choice of reference panels, methods, etc.

Response #1.16: We have added a new section to the Methods to emphasize all of these important caveats (see Methods section “Note on genetic ancestry inference”). [146-157]

Reviewer #2 (Remarks to the Author):

Comment #2.1: Overall an interesting and important contribution to the literature, clearly differentiating SIRE and genetic ancestry and describing their distribution in All of Us. I do think there are some points that need to be clarified/added to improve the paper however.

Response #2.1: We appreciate the reviewer’s positive impression of our paper. We address all of the reviewer’s individual points below in an effort to clarify the presentation of our results.

Comment #2.2: As discussed in NASEM report, likely important to state, even in the abstract, how ancestry is defined (i.e. based on similarity to reference panels 1000G/HGDP).

Response #2.2: As suggested, we have added a new section to the Methods, following the suggestions of the NASEM report, to state exactly how ancestry is defined in our paper (see Methods section “Note on genetic ancestry inference”). We also describe how ancestry is defined in the revised Introduction section. [146-157; 55-62]

Comment #2.3: While this is true- “shared ancestry is typically inferred by some measure of genetic similarity among individuals¹¹”. – the labelling of such clusters is still influenced by social factors (such as use of continental ancestry boundaries, etc) in how reference panels to which a new cohort is compared are defined (for example, in much work deciding that Caribbean and self-identified African American populations within the US belong in the AFR cluster for 1000G- though maybe those groups were excluded here based on Supplementary Table 2). This statement needs to be better contextualized.

Response #2.3: We agree that it is important to better contextualize how genetic ancestry is inferred, and we thank the reviewer for pointing this out. We have added a new section to the Introduction and Methods to explain how genetic ancestry is defined conceptually, mechanistically, and operationally. In the revised Methods section, we clarify that “The genetic ancestry results reported here are contingent upon the choice of reference populations, how these reference populations are defined, and the method used to infer genetic similarity between All of Us participants and the reference population samples” (see Methods section “Note on genetic ancestry inference”). [55-62; 146-157]

Comment #2.4: Could the downsampling performed for initial PC derivation impact the number of clusters identified in All of Us? I also think the authors should be clearer that number of clusters may well change with future releases.

Response #2.4: We eliminated the down sampling analysis for the revised manuscript and now perform density-based clustering of all PCA data using HBDSCAN (see Methods section on unsupervised clustering and Figure 2B). As suggested, we state that “The number of density-based clusters could change with future participant data releases.” (see Results section “Unsupervised: genetic diversity clusters”). [109-110; 172-173; 177-178]

Comment #2.5: I find figure 2D confusing- I think the clusters need to be presented color coded both by cluster identification from part C and by SIRE (two part figure)- and which color corresponds to which SIRE group should be clarified (use of blue for Hispanic or Latino and for Black or African American is confusing). Also how are individuals who noted more than one SIRE identification listed? Not clear as currently presented.

Response #2.5: Following the recommendation of the All of Us consortium, we have removed all analysis and comparison of race and ethnicity in the revised manuscript and focus exclusively on population structure and genetic ancestry (see new title). This entailed removing Figure 1 and stratifying groups by genetic similarity, rather than race/ethnicity, in revised Figure 2. [190-196]

Comment #2.6: How were the individuals chosen for this analysis “A similar approach was used to infer subcontinental ancestry for All of Us participants with high levels of African (n=2,934), American (n=235), Asian (n=2,652), and European ancestry (n=73,150; Figure 4). “ Why so few African, American, and Asian ancestry individuals? How was “high” defined? I think readers should be further cautioned in the discussion that the conclusions in Figure 4 for American ancestry individuals in particular are based on very small numbers, and that no individuals with Native American self-identification were included in All of Us at this time.

Response #2.6: As suggested by reviewer #1 (see comments and responses #1.3-#1.5 above), we have changed the criteria by which we select individuals for subcontinental ancestry inference. We now use clusters defined by PCA in an effort to include more individuals per ancestry group, and to separate out East Asian and South Asian groups. The new groups are shown in Supplementary Figure 5 and the results of the new subcontinental ancestry inference are shown in Figure 3. This approach resulted in the inclusion of more individuals within each group: 9,291 African, 2,457 East Asian, 24,4370 European, and 2,484 South Asian. Following the recommendation of the All of Us consortium, we removed American subcontinental ancestry analysis to avoid misrepresentation/stigmatization of AI/AN groups and the data analyzed does not include participants who identified as American Indian or Alaska Native” (see Results section “Supervised: genetic ancestry inference”). [119-123; 212-220]

Comment #2.7: Also, why was the GBR 1KGP grouping included in the African reference panel? Also which All of Us participants are included in Supplementary Figures 5-8- I think just those with high levels of ancestry from one continental population? (similar questions for European grouping in Supplementary Figure 5- I guess this is just showing individuals with “high” African ancestry from All of Us, and European cluster just included for reference? More labelling is needed- it should also be clarified if these are the “cross-All of Us” PCs (I think yes) or ones derived within each “high” ancestry group for these supplemental figures- better/more complete labelling needed).

Response #2.7: We have re-done the subcontinental ancestry analysis as suggested by reviewer #1 (see comments and responses #1.3-#1.5 above). Since we have lowered the percent ancestry threshold to include more samples, using PCA instead as suggested by the reviewer, the samples are more admixed and thus we need to include the minor component ancestry fraction samples in the PCA (e.g. the European ancestry references in the African subcontinental group) in order to ensure accurate ancestry inference. As suggested, we have removed the European component ancestry fraction samples from the African PCA shown in Supplementary Figure 6 for visual clarity (ie to emphasize the within group PC variance that is used in subcontinental ancestry inference). We have clarified which All of Us individuals were used for this analysis by visualizing them in PCA compared to continental reference populations (see Supplementary Figures 5-7).

Comment #2.8: “Genetic ancestry can interact with the environment to affect health outcomes in a non-additive way¹⁶⁻¹⁹ This is actually quite a strong claim to my mind, and I’m not sure that it is adequately proven in the papers you cite... not only are you claiming that both genetic ancestry (due to tagging of ancestry differentiated variants which are either of very large effect or all in the same direction of effect) and environment both associated with health outcomes independently (which I think is fair based on cited papers), you are stating definitively that there are statistically significant interactions between them, which I think are less known – gene environment interactions have been hard to identify and replicate in any population, and sample sizes for global ancestry proportion x environment interaction analyses will often be quite small. The association of global ancestry proportions with health outcomes is often a bit inconsistent across cohorts and can be confounded by social factors (such as discrimination based on skin color, separate from any true disease causal variants with inconsistent frequency across populations). I think this section should be revised/removed.

Response #2.8: We removed this statement as suggested.

Comment #2.9: In figure 3c- why aren’t all these ancestry components represented? African, American, East Asian, European, Oceanian, South Asian, and West Asian. Particularly confusing for Asian ancestry panels.

Response #2.9: As suggested, we re-did Figure 3 (now Figure 2) to show all the ancestry components: African, American, East Asian, European, Oceanian, South Asian, and West Asian. [223-228]

Comment #2.10: I think the relationship between the data presented here and the ancestry cluster identification information described in https://support.researchallofus.org/hc/en-us/article_attachments/19370367115796 in the All of Us research portal needs to be described. For example, the documentation describes use of an “other” ancestry cluster, which is not described here.

Response #2.10: As suggested, we compare the continuous ancestry percentages that we infer in our manuscript to the categorical ancestry predictions provided on All of Us Research Workbench. We provide a description of this new analysis in the Results section Supervised: genetic ancestry inference and the results are shown in Supplementary Figure 4. The categorical ancestry group predictions provided on the Researcher Workbench were inferred independently of our own approach, as discussed in the manuscript, but there is a high correspondence between group membership and percent continental ancestry.

Comment #2.11: In Figure 5- do you mean “Native American (teal)” or “American (teal)”? The second would be more consistent with terminology in the rest of the paper.

Response #2.11: Thanks for catching that mistake. This was removed in the revised manuscript.

Comment #2.12: I'm not sure I agree with the second part of this statement- I worry this could be taken as inferring that, on the whole, it is ok to assume that SIRE groups generally have the same pattern of genetic ancestry and that SIRE is an ok proxy for genetic ancestry at the individual level, which I think is almost certainly not what the authors mean. "tend to show group-specific patterns of genetic ancestry, which may be relevant to the diagnosis and treatment of disease". Would clarify/reword.

Response #2.12: Following the recommendation of the All of Us consortium, we have removed all analysis and comparison of race and ethnicity in the revised manuscript and focus exclusively on population structure and genetic ancestry (see new title). This entailed removing Figure 1 and stratifying groups by genetic similarity, rather than race/ethnicity, in revised Figure 2.

Comment #2.13: The authors should cite/describe differences between their methods and those presented in <https://jamanetwork.com/journals/jamacardiology/fullarticle/2808413>.

Response #2.13: We cite this relevant paper as suggested. The paper is a very short report that provides only a visual comparison of PCA data stratified by SIRE, rather than any quantitative analysis of genetic diversity, clustering, or ancestry. The visual evaluation of PCA data did not allow for any conclusion about how the data are clustered.

Reviewer #3 (Remarks to the Author):

Comment #3.1: The paper should be a good additional to the literature in order to understand the diversity of the All of Us dataset. I did not identify flaws in the analysis, interpretation and conclusion, but suggested the authors consider speaking to how the use of race/ancestry is affected by the recent NASEM Population Descriptors report. Specific suggestions below.

Response #3.1: We appreciate the reviewer's positive response to our manuscript. We have added a new Methods section "Note on genetic ancestry inference" to elaborate on how our use of race/ancestry is affected by the recent NASEM Population Descriptors report. [146-157]

Lines

Comment #3.2: 49-50: Suggest explaining why Eurocentric bias exacerbate health disparities.

Response #3.2: Done [43-44].

Comment #3.3: 60: Suggest adding "self-reported".

Response #3.3: Please note that following the recommendation of the All of Us consortium, we have removed all analysis and comparison of race and ethnicity in the revised manuscript and focus exclusively on population structure and genetic ancestry (see new title). This entailed removing Figure 1 and stratifying groups by genetic similarity, rather than race/ethnicity, in revised Figure 2.

Comment #3.4: 61: Suggest adding characteristic "derived from" the genome.

Response #3.4: This text got replaced with a new Methods section “Note on genetic ancestry inference”. [146-157]

Comment #3.5: 71-72: Suggest speaking to how genetic ancestry is associated with phenotype and thus perceived race/ethnicity which also affects health outcomes.

Response #3.5: This section was removed following the suggestion of the All of Us consortium (see response to comment #3.3 above).

Comment #3.6: 107: Suggest defining PMI: Skip.

Response #3.6: This section was removed following the suggestion of the All of Us consortium (see response to comment #3.3 above).

Comment #3.7: 108: Suggest explaining why data is currently unavailable for AI/AN

Response #3.7: The All of Us program recognizes that Tribes in the United States are sovereign nations, and some Tribes have concerns about taking part in research studies. All of Us has been engaged in a Tribal consultation process to learn about their views and to ensure that research using the program’s biospecimens and data from Tribal members is done in a way that is respectful of applicable Tribal customs, culture, and laws. A primary concern of Tribal leaders has been minimizing the risk of harm to Tribal communities once data from self-identified AI/AN participants is shared with researchers. In response to these concerns, the program paused any work proposing to share data volunteered by participants self-identifying as AI/AN in order to consult on the best path forward. See <https://allofus.nih.gov/about/diversity-and-inclusion/tribal-engagement>

Comment #3.8: 161-162: With zip code data not available for states listed, how much of the cohort is affected?

Response #3.8: We no longer use zip code data, and instead show ancestry/admixture results on a state-by-state level (see responses to Comments #1.6-#1.11 from reviewer #1 and new Figure 4).

Comment #3.9: 190-191: Sentence confusing -- "and participants"?

Response #3.9: Typo fixed.

Comment #3.10: 216: "there are" is repeated.

Response #3.10: Typo fixed.

Comment #3.11: 262: Again suggest adding "derived from".

Response #3.11: Fixed. [243-244]

Comment #3.12: 263: Suggest mentioning NASEM report on Population Descriptors and how it affects the use of race/ethnicity.

Response #3.12: As suggested, we have added a new Methods section “Note on genetic ancestry inference” to explain how we defined genetic ancestry following suggestions of the NASEM Population Descriptors report. Please note that following the recommendation of the All of Us consortium, we have removed all analysis and comparison of race and ethnicity in the revised manuscript and focus exclusively on population structure and genetic ancestry (see new title). [146-157]

Reviewer #1 (Remarks to the Author):

Overall, I am satisfied with the revisions provided by the authors. Additionally, I appreciate the authors' extensive re-analysis and revision of Figure 4.

A few points that would benefit the clarity of the manuscript:

Comment #1.1: Could a colorbar be added to the heatmap plots, like those in Sup Figure 2C?

Response #1.1: We added color bars to **Supplementary Figures 2 and 4** as suggested.

Comment #1.2: For the subcontinental PCA/UMAP analyses, was dimensionality reduction performed on the entire cohort and then sub-populations are just highlighted? Or was dim reduction performed ONLY on participants within a given inferred continental ancestry group? Either is fine, but it would be helpful to clarify the exact steps.

Response #1.2: Dimensionality reduction was performed only on participants within a given continental ancestry group, together with their respective reference populations, for the subcontinental PCA/UMAP analysis. We have added this information to the **Supplementary Figure 6 and 7 legends**.

Comment #1.3: On the figure title for Sup Figure 5, what does "coherent" mean in this context? Perhaps a different word would be more clear.

Response #1.3: We changed the **Supplementary Figure 5 legend** to read "Continental ancestry reference clusters..." for clarity as suggested.

Reviewer #4 (Remarks to the Author):

Comment #4.1: Overall, the authors do a good job at addressing the comments of the previous reviewer (Reviewer 2) and have performed multiple re-analyses that have improved the paper. The narrower focus on population structure and ancestry versus adding ethnicity into the mix is probably a good choice, given the ease of which this complex field can be misappropriated. The paper is short - which makes it a nice and easy read.

There are, however, a few comments that I believe could be addressed to balance this brevity with a bit more clarity to improve interpretability of the current results and especially to clarify methodological and conceptual challenges posed by these kinds of analyses.

Response #4.1: We appreciate the reviewers' thoughtful and nuanced suggestions related to the complexities of ancestry inference. We have incorporated these important caveats and additional detail into the revised manuscript as described for each point below.

Comment #4.2: Lines 50-61. The statement on the "Conceptual, mechanistic, and operational" definitions of genetic ancestry is a positive addition. A few points, however.

"Conceptually, genetic ancestry reflects the geographic origins of an individual's ancestors" – I'd argue that this is perhaps not a useful way of describing how "genetic ancestry" is being conceptualised here or that this is broadly agreed upon in the field. Even within the paper, ethno-linguistic labels used in tandem with geography for proxy ancestral populations like "Bantu" and "Han". Human populations have moved a lot over time, especially recently, so you will likely have ancestors that lived in various locations depending on what slice of time you are looking at and at what resolution.

Broadly, in this method, you are making an indirect inference about the sharing of more recent genetic ancestors between an individual and some ethno-linguistic or geographic group relative to some others. Whilst this can certainly be useful for some downstream applications, is not necessarily informative about the geographic origins of a person's ancestors as a whole. A few tweaks for clarity would probably minimise this potential for misunderstanding.

Response #4.2: As per the reviewer's suggestion, we have deleted the phrase "Conceptually, genetic ancestry reflects the geographic origins of an individual's ancestors" from the Introduction section and instead focus on (1) the mechanistic definition of genetic ancestry and (2) how genetic ancestry is typically inferred. We agree that this change could minimize the potential for misunderstanding. **See page 3, lines 54-55.**

Comment #4.3: When using the term "surrogates for ancestral populations" I believe it is worth noting (e.g. in Note on ancestry inference or in the introduction) that discrete ancestral populations don't (and never have) existed. Population boundaries both through time and in the present-day are blurry and genetic ancestry itself doesn't neatly map onto these PCA clusters or labelled reference populations.

Whilst the fact that "ancestral source populations" model a more complex reality may be obvious to

more informed readers, the All Of Us project has a broad user-base, so ensuring this point is drilled home is appropriate I think.

Response #4.3: We agree that this is an important caveat, which we have added to the “Note on genetic ancestry inference” as suggested, borrowing liberally from the reviewer’s wording, which we found to be quite useful (thanks!). We state that “Although reference populations are taken as surrogates for ancestral populations, it should be stressed that human populations are an idealized concept and discrete ancestral populations did not exist just as modern populations are not discrete. Rather, population boundaries past and present are fuzzy and genetic ancestry does not map neatly onto clusters defined by PCA or labelled reference populations”. **See pages 5-6, lines 155-161.**

Comment #4.4: Following from the above, convex “ancestry components” as output by the Rye method may require some more careful interpretation.

For example: “All groups show evidence of admixture with multiple ancestry components present in different proportions.” is perhaps a reasonable interpretation in some cases (e.g. this is widely described for African Americans). However, could some of these results not also simply be explained by incomplete sampling of reference populations + spatial genetic structure?

Imagine performing this inference on a member of a population roughly equally dissimilar (central) to multiple reference pops simply through genetic drift / isolation by distance. When you force discretisation, I’d wager you will always see evidence of such “admixture”, which could then just disappear if the reference populations changed or another was added e.g. should we really just interpret cluster 5 as simply the product of recent admixture between Europeans, South Asian and West Asian populations? Is there any explanation for this that maps to historic events? It seems to me that alternatives (such a spatial population structure) could also explain these results.

These issues surrounding potential for misinterpretation should, in my opinion, be discussed more explicitly or even tested for / demonstrated in Supplementary Materials (if possible!) e.g. by simply removing one or more reference populations from a given analysis. Given the method is described as “accurate” elsewhere (line 133; I’m not sure how this can be discerned necessarily unless through simulations for example), I’d recommend instead taking the chance to emphasise what the outputs of the methods are actually telling us, and where modelling things using incomplete / discrete source populations may break down.

Response #4.4: The reviewer raises the important point that the ancestry inference results we report are dependent on the reference populations used for the analysis. As suggested, we performed a sensitivity analysis for cluster 5 by sequentially adding and removing reference populations. It is very reasonable to expect that incomplete sampling of reference populations, coupled with spatial population structure, may better explain the observed three-way European, South Asian, West Asian admixture seen for this cluster, particularly if the participants in cluster 5 have ancestral origins in the Arabian Peninsula, Iraq or Iran, south and east of the West Asia reference populations we have from the Human Genome Diversity Project (HGDP). In this case, the South Asian ancestry component of cluster 5 could be attributable to missing reference populations. The closest HGDP reference population we have

is for Central Asia (Kalash). However, adding the Central Asian reference population did not noticeably change the ancestry inference for cluster 5 – the Southern Asian ancestry component remains. Next, when the South Asian reference population was removed, the Central Asian component takes its place. Finally, when the West Asian reference population was removed, the South Asian ancestry component takes its place, with little to no contribution from the Central Asian population. These results underscore the extent to which the reference populations influence the results of the ancestry analysis, but do not point to a missing Central Asian ancestry component in our data. Fortunately, however, the challenge of incomplete reference populations only applies to a small percentage of All of Us participants (~3%), the majority of which show ancestral origins from global regions that are well-covered by the reference populations used here.

The results of the new continental ancestry sensitivity analysis are shown in **Supplementary Figure 8** and reported in the Results section (**page 7, lines 228-244**). We also added a description of this important caveat the Discussion section (**pages 8-9, lines 292-303**).

Comment #4.5: As an aside to the previous comment, apologies as I didn't see the initial draft, but almost all individuals in e.g. with any apparent "African" ancestries in Figure 2 also have some evidence of "European" ancestries, and European reference pops are now not included (it appears they may have been before?). Given this, is the ubiquitous "East Bantu" component in Figure 3a just leaking in as the "closest proxy" for European ancestries here?

Given the apparent lack of any evidence for Eastern Bantu related ancestry components in the African American population in the USA in previous research (e.g. see Ongaro et al. 2019, The Genomic Impact of European Colonisation of the Americas), this is an unexpected result that may demonstrate another example of the challenges of reference population curation and potential issues of interpretability described above.

Response #4.5: Per the previous comment, we agree completely that the ancestry inference results are dependent on the reference populations used for the analysis. Furthermore, as the reviewer guessed, European reference populations were previously included in the African subcontinental analysis but later removed based on a previous comment from reviewer #2. We performed a sensitivity analysis for this group, similar to that described in the response to comment 2.4, to assess whether the East Bantu component can be attributed to missing European or North African reference samples. This does not seem to be the case. When both European and North African reference populations are included in the analysis, the average European ancestry percentage for these samples is 4.3% and the average "East Bantu" ancestry is 7.7%. Having said that, it is likely the case that the "East Bantu" component is the closest proxy for some other reference population that is not included in the study.

The results of the new subcontinental ancestry sensitivity analysis are shown in **Supplementary Figure 9** and reported in the Results section (**pages 7-8, lines 246-255**). We also added a description of this important caveat the Discussion section (**pages 8-9, lines 292-303**). We also added the suggested reference to put this analysis in the context of previous work (**reference #40**).

REVIEWERS' COMMENTS

Reviewer #4 (Remarks to the Author):

The authors have done a great job at responding to my questions and adding new analyses quickly and thoughtfully.

I believe the paper should now be accepted for publication.

Author responses: We are delighted to read that the last reviewer was satisfied with our responses and new analyses and recommends that the paper should now be accepted for publication.