HPRC-ELSI Working Group Statement on the Use of 'ASW' and 'MXL' Cell Lines from the 1000 Genomes Project Collection at Coriell Institute

[Finalized and Submitted to HPRC Coordinating Center: December 21, 2023]

Background: A team of ethical, legal, and social implications (ELSI) and genomics scholars comprise the ELSI Working Group (WG) of the NIH-NHGRI funded Human Pangenome Reference Consortium (HPRC). This group advises the HPRC on issues related to ELSI that arise in real time throughout the course of the project. One of these issues includes population sampling, which began in Phase 1 with sampling existing cell lines at the Coriell Institute that were generated through the 1000 Genomes Project.

Many 1000 Genome Project (1KG) samples are labeled according to where they were collected, combined with ancestry that donors described. These community- engaged labels are called 'sub-populations' on official websites describing the data, and in research publications; continental 'super-population' labels into which the more granular descriptions are combined (often inconsistently) are also associated with the samples. Scientific research shows that these labels offer incomplete representation of genomic diversity among populations they are meant to include, leading to false positive and false negative findings in genetic association studies that assume genetic population structure at the continental level. For example, a reference panel recently developed to account for European subcontinental diversity showed that 1KG covered only 26.8% of European genomic diversity, by comparing European to European American samples.[1]

Based on these limitations and many other methodology-based concerns about the use of continental-level population descriptors, their use was explicitly discouraged in a 2023 report with recommendations by the National Academies of Science, Engineering, and Medicine (NASEM) Committee on the use of population descriptors such as race, ethnicity, and ancestry in genetics and genomics research.[2] At the time of publication, genetics and genomics investigators still regularly use the 1KG 'super-population' continental labels in analyses, perpetuating harmful misconceptions about inflated genetic heterogeneity between continents and a false sense of homogeneity within continental groupings.

There is an effort funded by NIH-NHGRI with one of the HPRC ELSI WG members (C. Adebamowo, PI)–to engage with some of the 1KG communities sampled in Africa, with the goal of learning about their current perspectives on the creation of inducible pluripotent stem cell (iPSC) lines using their donated samples to the 1KG Project. Since the community advisory boards established during 1KG are largely not reachable, and most are no longer active, an alternative approach to attain answers to these questions is to engage with members of the communities who participated.

There is active, ongoing discussion within HRPC about prospective recruitment beyond the use of 1KG cell lines; specifically related to the type(s) of information that should accompany samples and data generated from samples as metadata on provenance; e.g., in the form of population descriptors. We will continue to address this important theme in year 5 of Phase 1 and into Phase 2 of HPRC, if funded. It is a major ongoing task for the Population Sampling and Representation (PSR) and ELSI WGs.

---

[1] Gouveia, M.H., Bentley, A.R., Leal, T.P. *et al.* Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. *Nat Commun* **14**, 6802 (2023). https://doi.org/10.1038/s41467-023-42491-0.

[2] National Academies of Sciences, Engineering, and Medicine. 2023. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26902.

Reason(s) for the Inquiry: In this case, the issue of Indigenous Data Sovereignty was raised by the ELSI WG in response to an HPRC investigator inquiry about the use of 1KG Project cell lines labeled 'Mexican Ancestry from Los Angeles (MXL)' and/or 'African ancestry in the American Southwest (ASW)' to infer genetic information about 'Indigenous American' or 'Native American' groups, as a proxy for those unsampled.

Because many Tribes in the US have laws or restrictions about research involving members of their Tribes and because there are hesitations about using genomic data from individuals to infer information about others, concerns about appropriate uses of 'ASW' and 'MXL' cell lines were raised.

As a result, further production of HPRC assemblies using these cell lines was paused. Chairs of the HPRC Technology & Production, and the PSR WGs requested input from the ELSI WG on how to proceed. HPRC leadership and the Coordinating Center then requested a written statement on the ELSI WG's process of deliberation, what considerations were discussed, and our recommendations for the Consortium on how to proceed regarding the use of these samples.

Facts about the 1KG Cell Lines in Question:

- 'MXL'-labeled samples from the 1000 Genomes Project (1KG) collection is a cohort of 104 samples collected in Los Angeles from those of Mexican ancestry (at least 3 of 4 grandparents born in Mexico)
- ASW is a group of 107 samples from people with African-American ancestry (all four grandparents African American resided in Oklahoma). They trace their lineage to Oklahoma Black townships established in the 1890s. There may be some ancestry from American Indians who were forced to Oklahoma on the Trail of Tears, six decades before the Black townships were established. The ASW cohort did not identify as Indigenous, and did not want to be labeled "African or African American," but rather, chose the label: "African ancestry collected in the American Southwest."
- These samples were *not* selected for inclusion in HPRC using these (or any other) population descriptors, but by using genotype data that is publicly available, with an algorithm that predicted they would add more common genetic variants (at 1% frequency or higher among all samples the 1KG collection) to the human pangenome reference resource.
- All 1000 Genomes Project cell lines are held at Coriell Institute and managed through a shared governance model with the National Human Genome Research Institute (NIH-NHGRI); these biomaterials are managed with certain commercial restrictions, separately from the sequence data, which are open access, and consented for broad future use(s) without restrictions.

Data from the 1KG Sample Collection:

- Data generated from these cell lines during 1KG's Phase 3 Next-Gen sequencing effort are publicly available on the NIH-NHGRI AnVIL platform, the U.S. National Center for Bioinformatics (NCBI)[3], the International Genome Sample Resource (IGSR)[4], and Amazon Web Services (AWS) S3.[5]
- These data are considered fully open access, and the consent protocol specified they could be used for broad future research purposes, yet to be defined, 'forever' and without the opportunity to be recontacted for further inquiry.
- Adding cell lines from this collection to the production pipeline for high-quality, telomere-to-telomere (T2T) pangenome assemblies to be included in the HPRC collection would add more accurate and

[3] ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp
[4] https://www.internationalgenome.org/data/
[5] https://registry.opendata.aws/1000-genomes/

detailed information about the genomes of the individuals sampled. Data would be considered fully open under the current model, available for public access, download, and use without restrictions.

Process of Deliberation: ELSI WG members discussed the widespread issue of using 1KG samples and other datasets collected outside the US in South and Central America as a proxy for 'Indigenous' or 'Native American' genetic ancestry. This topic has come up on HPRC Steering Committee (SC) calls, as well as PSR WG and ELSI WG calls, in the context of educating HPRC leadership about the history and current status of human genomics practices that disregard Indigenous data sovereignty.

- When the issue was first raised, the ELSI team reached out to the original 1KG investigator Julio Licinio, who organized community meetings for the 'MXL' participants. He stressed that Indigeneity was not a factor in recruiting or labeling this sample population, specifically cautioned against using the cohort as a proxy for indigeneity, and noted the community members elected to be described in the specific way they are labeled as 'MXL'.
- Concerns were raised about these samples being used as proxies to infer and label Indigenous American ancestry, at multiple points throughout this project; including the PSR WG, Steering Committee, and Technology & Production WG calls. When 171 cell lines from the 1KG collection at Coriell were selected by the 'MaxVar' algorithm and sent to the PSR WG for approval to send to production via email, the ELSI team flagged the 'MXL' and 'ASW' cell lines for review by the WG, to ensure that all ELSI team members who are knowledgeable about Indigenous Data Sovereignty had a chance to weigh in.
- Some members of the ELSI team felt that the issues had been resolved when outreach to the original 1KG organizers revealed that the samples were neither selected nor interpreted to represent Indigeneity, but other members of the team felt there was more to discuss. They did not feel comfortable moving forward without sufficient time to consider the nuances and complexity of this issue; especially in the context of a field (human genetics) that regularly disregards the 1KG community-driven cohort (so-called 'sub-population') labels, in favor of continental-level ('super-population') categories, or inferring and/or assigning labels based on dimension-reduction and clustering techniques using genotypes.
- Some ASW and MXL samples were included in the 2021/22 tranche for HPRC, but only after the decision to use the MaxVar algorithm, rather than population labels, to decide inclusion in HPRC. That is, the population label was not the basis for inclusion.
- A subset of the ELSI team with background expertise and/or involvement with academic professional movements related to Indigeneity and data sovereignty was invited to weigh in on the subject over email, in smaller group discussions, and in 2-3 consecutive ELSI WG meetings.
- A formal response was drafted in Google Docs, and the ELSI team had two weeks to provide feedback on the draft response.
- The full ELSI WG discussed the draft response, and additional comments were added to the draft which clarified and elaborated on the considerations.
- A final draft response was circulated to the ELSI team, and they were given 48 hours to provide feedback before the statement was finalized and submitted to the Coordinating Center at WashU.

Considerations and Perspectives:

- Many people of Mexican ancestry, heritage or descent do have some Indigenous genetic ancestry, estimated as high as 60-80% in some parts of Mexico[6], with the caveat that methods used to estimate

[6] https://onlinelibrary.wiley.com/doi/10.1002/ajhb.23032

ancestry proportions are based on the same approaches cautioned against in the NASEM report (i.e., using existing reference data to infer ancestry labels and assign them to new samples/data).

- The 'MXL' cohort developed its own name after community consultation with study investigators, and the organizer cautioned against attributing Indigeneity.
- One member of the working group expressed discomfort that the 'MXL' samples were likely to be used as a way of gathering and producing information about genetic architecture and variation on Indigenous Mexican ancestral haplotypes without directly engaging with Mexican Indigenous groups.
- The 'ASW' sample population is sometimes classified using the 1KG 'super-population' label for Indigenous American (AMR) and sometimes using the continental label for African ancestry (AFR). Both are incorrect for some members of the cohort, because continental labels are inappropriate for the purpose of characterizing ancestry, heritage, or genetic background.
- The International Genome Sample Resource (IGSR) attributes the 'ASW' sample population to continental label 'AFR' although the donors were all based in the USA and their grandparents were US residents, so at least 3 generations based in North America. This illustrates the ambiguity of continental "origin" and "ancestry" labeling, elucidating the need for data resource management teams to be aware of, and clarify these labeling issues.
- In some reports in the literature and on the IGSR site, 'MXL' samples are attributed the 'AMR' label, and are sometimes assumed or inferred through statistical population reference-based analyses and dimensionality reduction techniques to be Indigenous American.
- The MXL cohort may have admixture with North and Central American Indigenous populations, but it was not collected in a way to ensure that, and should not be used as a proxy for Indigeneity.
- The ELSI team agreed to approve moving forward with the 'MXL' and 'ASW' samples selected for inclusion in HPRC, although one working group member still cautioned about using 'MXL' as a way to include samples without engaging the affected communities; and the WG determined that this approval should have a few caveats. Primarily, there should be some way to alert data users of the ways in which they are permitted and restricted from labeling samples–and, guiding principles leading to best practices should be developed for HPRC.

Recommendation(s): In general, HPRC should not perpetuate the oversimplification of assigning 1KG sample populations to continental-level designations. ELSI and PSR WGs will continue to collaborate to address the use of population descriptors, labeling and metadata included with samples and data, and methods for linking samples and data to terms of use.

HPRC should take steps to ensure that investigators won't use labeled genetic data from these samples as a proxy for Indigenous (Native North, Central, or South American) ancestry or otherwise deviate from the 1KG labels that were established through community consultation. In particular, categorizing them according to "continental" or other "superpopulation" labels is inaccurate for some members of the cohorts, apt to be misleading, and deviates from the recent NASEM report recommendations.

The 1KG cohorts have specific recommendations for use of population labels (e.g., for 'MXL' samples[7]) as well as rules for their general use[8] and, at times, some instances of 1KG 'super-population' (continental) labels used in the HPRC have not been compliant with those terms.

[7] https://catalog.coriell.org/1/NHGRI/Collections/HapMap-Collections/Mexican-Ancestry-in-Los-Angeles-CA-USA-MXL

[8] https://catalog.coriell.org/0/Sections/Support/NHGRI/NHGRI_Pop_Ref.aspx?PgId=688

HPRC investigators, and as a consortium, should make every effort to comply with the original 1KG project recommendations for specific cohort sample descriptions and the resource in general, as well as the NASEM report's recommendations on continental labels, and apply these standards to samples and data generated by HPRC. The following rationale is provided in the NASEM report summary, cautioning against the use of continental labels to describe discrete categories of humans:

> *Avoiding Typological Thinking*. Erroneous categorical assumptions can be scientifically and ethically detrimental, particularly when applied to studies of human history, identity, variation, and traits or diseases. There is a pervasive misconception that humans can be grouped into discrete, innate biological categories. The committee cautions against the use of typological categories, such as the racial and ethnic categories established by the U.S. Office of Management and Budget in Statistical Directive 15, for most purposes in human genomics research. While the use of these categories may be required of researchers under certain circumstances (for example, in describing participants in studies receiving federal funding), the fundamentally sociopolitical origins of these categories make them a poor fit for capturing human biological diversity and as analytical tools in human genomics research. Furthermore, use of these categories reinforces misconceptions about differences caused by social inequities. Current practices in human genetics, including the use of descriptors such as continental ancestry, also reinforce these views. (Emphasis added; pg.7).

The following recommendations are more specific and address the ways in which HPRC ELSI team members suggest moving forward in parallel with formal approval for including 'MXL' and 'ASW' cell lines from 1KG into HPRC.

1) HPRC should develop "best practices" for use of HPRC data (including but not limited to the following recommended restrictions and permissions)
   a) Do not use 'super-population' labels to describe, group, or analyze data; select samples for genomic diversity; or report results. Instead, point to population descriptors determined by communities.
   b) Do not use any semantic labels or genetic information in the samples as a proxy measure for Indigenous American ancestry, or to assign 'Native American'-associated labels to groups of samples via cluster analyses, unless derived from samples that entailed engagement with the communities being characterized.
   c) Do not re-identify anyone in any of the HPRC collections, for any reason

2) The Coriell Institute entit(ies) that approve requests for cell lines should clarify permissions and restrictions associated with the use of HPRC data; and ensure compliance with these recommendations in meta-data catalogs.

3) Data downloads from the HPRC website and other open data access points from third-party sites would ideally include a pop-up window, or a link to the HPRC website, with guidance for investigators and an agreement associated with cautioning language about population descriptors; e.g., potentially inaccurate 'super-population' labels; attributing racial, ethnic, geographic, or genetic ancestry (including Indigenous ancestry); and prohibiting the re-identification of individuals from the data.

4) HPRC-ELSI leadership should collaborate with other HPRC investigators and scholars to co-develop guiding principles and best practices for the consortium, which would ideally specify the appropriate use(s) of the data, and include restrictions, recommendations, and an ethical framework for broad future use of legacy sample collections.
    a) These guiding principles and related best practices should be (at minimum) posted on the HPRC website, and published in a high-impact, peer-reviewed journal to ensure visibility of the effort and guidance.
    b) These should address population descriptors, ideally in a way that links samples and data to appropriate use(s) and specific considerations.

5) With these caveats, ‘MXL’ and ‘ASW’ 1KG cell lines have ELSI team support for inclusion in HPRC.