



TOPMed

Published on *NHLBI Trans-Omics for Precision Medicine* (<https://www.nhlbiwgs.org>)

Guidelines on the use and reporting of race, ethnicity, and ancestry in the NHLBI Trans-Omics for Precision Medicine (TOPMed) program

Updated 06/10/2020

V3.1

[View as PDF](#) [1]

Authors:

Alyna Khan, Caitlin McHugh, Matthew P. Conomos, Stephanie M. Gogarten, Sarah C. Nelson and the GAC Race and Ancestry Discussion Group

Contents

- [Motivation](#) [2]
- [Terminology](#) [3]
 - [Considerations for investigators](#) [4]
- [Harmonization of race and ethnicity in TOPMed](#) [5]
 - [Data collection methods often vary](#) [6]
 - [Harmonization of race and ethnicity by the Data Coordinating Center \(DCC\)](#) [7]
 - [Considerations for investigators](#) [8]
- [Genetic ancestry](#) [9]
 - [Considerations for investigators](#) [10]
- [Analysis](#) [11]
 - [Principal components adjustment](#) [12]
 - [Stratified analysis](#) [13]
 - [Using reported race/ethnicity as a covariate](#) [14]
 - [Differences in trait variance by race, ethnicity, and ancestry](#) [15]
 - [Considerations for investigators](#) [16]
- [Reporting](#) [17]

◦ [Considerations for investigators](#) [18]

• [Conclusion](#) [19]

• [Acknowledgements](#) [20]

• [References](#) [21]

[Back to top](#) [22]

Motivation

Defining race, ethnicity, and genetic ancestry and using these concepts in human genomic research programs such as TOPMed has wide-ranging implications for how the research is translated into clinical care, reported in the media, incorporated into public understanding, and implemented in public policy (Graves, 2011). In particular, if care is not taken in the way these concepts are deployed and communicated, harmful misconceptions of race/ethnicity and their relationship to genetics can arise (Harmon, 2018). In the extreme, findings from genetic research can be used to fuel racism and discrimination (Lee, 2008). More subtly, the way scientists deploy the concepts of race, ethnicity, and ancestry in the context of genetic studies can lead to the biological reification of social categories (Braun, 2006), particularly when ancestry is cast at the continental level (Fujimura, 2011). In addition to social harms, using poorly defined concepts or categories can also lead to results with poor scientific validity (Lee, 2001; Shields 2005).

TOPMed is a large consortium of genetic studies that encompass many different races, ethnicities, geographic locations, and ancestries (Taliun, 2019). This diversity of study populations is a major strength of TOPMed; such diversity enables the expansion of knowledge of genetic variation and an improved understanding of disease (Bently, 2017). Furthermore, this diversity highlights the importance of being transparent about genetic knowledge and its relevance to ancestry and to avoid the misappropriation of science to support racist beliefs (ASHG Executive Committee, 2018). Both diversity and transparency are necessary to fully realize the ability of the TOPMed program to contribute robustly and equitably to precision medicine.

Two main factors that motivate the development of these TOPMed guidelines are (1) to approach analytical and methodological decisions accurately and responsibly when using race/ethnicity and ancestry variables and (2) to communicate concepts of race, ethnicity, and ancestry appropriately and respectfully. We recognize that past efforts to set similar guidelines have been made with limited success in changing practice (Foster, 2009). Here, we aim for these guidelines to gain traction by providing concrete analytical and methodological points and by pairing recommendations with examples and observations from within the TOPMed data. Specifically, we can improve upon methodological transparency, reproducibility, and interpretability of analyses by articulating and justifying analytical decisions related to race/ethnicity and ancestry. These guidelines also aim to help investigators navigate some of the challenges in using socially and genetically defined groups in scientific discussions by presenting an overview of commonly used terminology, highlighting options for analysis, and discussing considerations in reporting findings.

[Back to top](#) [22]

Terminology

When presenting information on the race, ethnicity, or ancestry of participants in a study, it is essential to be clear about whether the labels used refer to reported information or to something inferred from genetics. People may use these terms in different ways, and even among dictionaries there is no clarity on their precise meaning. “Race” and “ethnicity” generally refer to social, not biological, categories, but they are often used interchangeably, or as the hybrid term “race/ethnicity.” “Ancestry” is generally used in genetic research to imply something about a person’s genetic origins; for example, whether the majority of their ancestors were from Africa, the Americas, Europe, or Asia (sometimes referred to as “continental ancestry”) (Royal, 2010). Ancestry can also be on a finer scale, such as having ancestors from specific countries or geographic regions. In this document, we will use the term “race/ethnicity” to refer to social categories, and “genetic ancestry” as describing genetic origins.

Both race/ethnicity and genetic ancestry can be relevant to consider in an analysis. Race is often tied to social factors influencing health; examples include racially-based housing discrimination influencing environment and increased stress levels in individuals experiencing racism. Genetic ancestry influences the relative frequencies of variants in different populations, as well as the patterns of linkage disequilibrium among variants. Because reported race/ethnicity and genetic ancestry may both appear in scientific discussions and communications, care must be taken to describe exactly what is being presented and why.

[Back to top](#) [22]

Considerations for investigators

- **Explicitly distinguish between variables that derive from non-genetic, reported information versus genetically inferred information.**
- **Avoid assuming that non-genetic, reported variables are by “self-report.”** Study- or cohort-specific documentation may help determine whether variables (e.g. race or ethnicity) were self-reported versus recorded by study personnel without soliciting self-report from the participant.
- **Avoid using terms that are historically linked to hierarchical, racial typologies** --- specifically, “Caucasian” (Moses, 2016; Krieger, 2005). “White” or “European” or “European-American” could be used instead.
- Follow the [APA’s guidelines on bias-free language regarding racial and ethnic identity](#) [23].

[Back to top](#) [22]

Harmonization of race and ethnicity in TOPMed

[Back to top](#) [22]

Data collection methods often vary

Race and/or ethnicity are often collected by a study and included with other phenotypes describing study participants (such as sex, age, and height). A common method of collecting this information is for study participants to fill out a form indicating their race and/or ethnicity (typically choosing from among a set of options provided by the study team), which leads to “self-reported” values. Other collection methods are also possible, including designation by a third party (health care provider or study data collector) who

typically infers the participant's ascriptive race, or through study documents that describe the recruitment population but do not ask whether the self-reported race/ethnicity of specific individuals differs from the target population. Whether self-reported or ascriptively assigned, the race and/or ethnicity of a specific participant is always a function of the specific options provided in study instruments, which will often vary by location or the research interests of investigators. Some studies may ask only one question about race or ethnicity, while others may ask separate questions. Some studies may give participants a wide array of possible choices for race and allow them to select more than one, while other studies may ask them to select the best match from a very limited set of choices.

[Back to top](#) [22]

Harmonization of race and ethnicity by the Data Coordinating Center (DCC)

The diversity in data collection methods presents a challenge for investigators attempting to combine data from multiple studies. Race/ethnicity has been redefined throughout history as is reflected in the changes to race and ethnicity classification for federal data over the years (Brown, 2020; NOT-OD-15-089). Unlike phenotypes measured with different units for which translation is required prior to data aggregation, there is often no straightforward method to convert one set of race/ethnicity categories into another. This is particularly the case when study cohorts include individuals sampled from distinct national contexts where socio-cultural understandings of racial and/or ethnic identity differ. Previous efforts by the TOPMed DCC to harmonize race and ethnicity information included the capture of three variables: `race_1`, `ethnicity_1` and `hispanic_subgroup_1`. As TOPMed studies expanded to include participants from outside the U.S., using a U.S. administrative definition of race and ethnicity became a challenge. The DCC's new attempt maps each US-based study's reported information onto the categories currently used by the U.S. Census (U.S. Census Bureau, 2017) in a `race_us_1` variable, including the category "Multiple" when participants report more than a single race. The variable was renamed to emphasize that values represent U.S. Census categories, and the definition of the harmonized variable is appropriate only for studies with participants living in the U.S. In the U.S. Census, "ethnicity" has the narrow definition of "Hispanic/Latino" or "Not Hispanic/Latino." In this new framework, the variable `hispanic_or_latino_1` specifically refers to whether a participant is "Hispanic/Latino", "not Hispanic/Latino", or "both." While the assignment of "Multiple" or "both" is not a helpful one for analysis, we include it as a category to preserve data about multiple responses wherever possible. However, we recognize the limitations of this approach, as much more detailed information collected by some studies is not represented in the harmonized phenotypes.

[Back to top](#) [22]

Considerations for investigators

1. **When using DCC-harmonized race and ethnicity variables**, keep in mind that these harmonized variables are based on reported data (often self-reported but not always) rather than genetic inference and often represent a simplification of more complex sources of information that may not translate well between different studies and jurisdictions (e.g. different countries).
2. **When attempting to use race categories for non-U.S. populations, consult with investigators from the TOPMed study/studies on how best to refer to population groups.**
3. **Avoid assuming U.S. race categories for non-U.S. participants.**

[Back to top](#) [22]

Genetic ancestry

Genetic ancestry can be inferred indirectly by examining genetic similarity, either among participants with reported values, or to reference samples of known ancestry (Mathieson, 2020). Principal Component Analysis (PCA) describes the variation in the genetic data as a continuous, multidimensional distribution, with participants whose ancestors came from the same geographical area usually clustering together in PC space. Reported race and ethnicity are often used to interpret which clusters correspond to which geographical areas, (e.g. a cluster of self-identified Blacks would indicate African ancestry). Even in the absence of reported race in a set of participants, reference populations from specific geographic regions may be included in the analysis and used to interpret the meaning of the observed clusters (e.g., study participants clustering with Yoruba in Ibadan, Nigeria samples from the 1000 Genomes Project would also indicate African ancestry) (The 1000 Genomes Project Consortium, 2015). Software such as ADMIXTURE (Alexander, 2009) allows inference of what fraction of a person's genome comes from various ancestral populations in different geographic regions, again using reference samples from these regions. Because race is highly correlated with genetic ancestry, and indeed often used as an interpretive tool in ancestry analysis, separating the two concepts is often difficult. In sample sets where participants' ancestors came from geographically isolated regions, it can be tempting to interpret the results of PCA as grouping participants into discrete racial categories. In sample sets with substantial admixture, although clustering is still evident, the clusters are not discrete.

While self-reported race is often used as a proxy for genetic ancestry (e.g., reporting the frequency of a variant in subjects who checked "Asian" on a study intake form), this can be problematic. Although race and ethnicity are often highly correlated with genetic ancestry, they are not the same. For example, a light-skinned person of majority African ancestry may identify as White, while another person with more European ancestry may identify as Black, influenced by many factors, including their family and/or culture. The problem is especially acute in admixed populations, defined as groups of people whose genetic ancestry spans multiple continents. Such groups are often treated as homogeneous when they are, in fact, extremely heterogeneous. For example, people who identify as Hispanic/Latino have a wide variety of cultural backgrounds and genetic ancestries, with different proportions of admixture from Africa, the Americas, and Europe (Conomos, 2016).

[Back to top](#) [22]

Considerations for investigators

1. **Avoid reinforcing the idea that race is the same as genetic ancestry.** When presenting genetic concepts such as principal components or allele frequencies, use labels like "European ancestry" and "African ancestry" rather than "White" and "Black." If reported race values were used as a proxy for ancestry, note this in the methods.
2. **Avoid using reported race as a proxy for genetic ancestry.** Race and ethnicity are often highly correlated with genetic ancestry, but they are not the same. If using reported race as a covariate, whether as a proxy of genetic or non-genetic factors, justify the reasoning and mention this in the methods section.

[Back to top](#) [22]

Analysis

When considering how to use race/ethnicity and/or genetic ancestry information in a genetic analysis, an analyst must first assess the goals of the analysis and the intended purpose of inclusion of those variables. In a genetic association test, the goal is to identify variants that are associated with a particular trait or disease. Genetic ancestry may be a confounding factor in the analysis if it is associated with the trait or disease of interest, as allele frequencies at many variants differ between ancestral populations.

[Back to top](#) [22]

Principal components adjustment

To adjust for confounding due to genetic ancestry, it is often recommended to include PCs from a PCA of sample genotype data as covariates in the association test. Study participants with similar genetic ancestry tend to cluster together in the top several PCs (Novembre, 2008). The number of PCs to be included as covariates for this purpose can be determined by using information about reported race/ethnicity to interpret patterns in the data, and selecting only PCs that separate known populations. If no such information is available, projecting reference samples of known ancestry onto the PCs can aid in interpretation. An alternative but conceptually similar approach to using PCs is to include as covariates ancestry proportions for each subject, estimated using reference samples of known ancestry and software such as ADMIXTURE.

[Back to top](#) [22]

Stratified analysis

Another approach to address confounding due to genetic ancestry is to conduct a stratified analysis, where different racial/ethnic or ancestry groups are analyzed separately, followed by meta-analysis. While meta-analysis is a useful tool for combining results from multiple studies conducted separately, it is often unnecessary in TOPMed given the cross-study harmonization of phenotype and genotype data and the development of computational tools that can efficiently analyze samples with well over 100K individuals. Still, meta-analysis can be an effective strategy to account for confounding, especially if investigators expect heterogeneity in other model parameters, such as covariate effects. However, we encourage investigators who take this approach to focus on the meta-analysis results and exercise caution when interpreting the stratum-specific results.

To demonstrate the need for caution, we highlight a common motivation for interpreting the stratum-specific results: to determine whether participants of a particular ancestry are “driving” the association signal observed in either the pooled- or meta-analysis. In answering this question, the precise definition of the strata must be taken into consideration, along with any attendant limitations. For example, participants within strata based on reported race/ethnicity will typically not have homogeneous genetic ancestry. Therefore, extrapolating results from such strata to infer differences among ancestries is problematic, particularly in studies with recently admixed individuals, such as Hispanic/Latinos or African Americans, where ancestry proportions can vary greatly among people who identify as the same race/ethnicity. Reported race/ethnicity is often missing for a subset of participants within a study who would then be omitted from an analysis stratified on these variables, leading to decreased power in stratified analyses compared with pooled analyses. Alternatively, participants could be stratified based on inferred genetic ancestry; e.g., using PC values or estimates of admixture proportions. While this addresses the missing data problem, it also requires some definition of strata, which remains problematic in studies with admixed

individuals. Clustering or machine learning algorithms may be used to infer strata, but these approaches often rely on reported race/ethnicity information to define a set of stratum labels and train the model. One such method, HARE (Fang, 2019), uses a support vector machine (SVM) to estimate probabilities of stratum membership for each participant based on the similarity of their ancestry PC values to those of participants with provided race/ethnicity values. Even with such an approach, participants within inferred strata will still typically have non-homogeneous genetic ancestry, and generalization of results will be problematic.

In contrast, pooled analysis adjusted for ancestry PCs or ancestry proportions does not require arbitrary clustering decisions and also allows inclusion of all participants in the analysis, including those with either missing or underrepresented race/ethnicity (Manolio, 2019). Furthermore, in our experience analyzing TOPMed data, we typically do not find additional signals from stratified analyses that are not also captured by pooled analysis. Attempts to further characterize ancestry-specific contributions to association signals may be better framed as determining on which ancestral haplotype(s) a particular variant is found. This question can be addressed more directly by performing a local ancestry analysis and using methods such as admixture mapping (Shriner, 2014) to locate the variant of interest on a particular haplotype, although this approach is not without limitations.

[Back to top](#) [22]

Using reported race/ethnicity as a covariate

Reported race/ethnicity may also be a confounding factor in the analysis if it is associated with both the frequency of tested variants and the trait or disease of interest. In most cases, however, association between reported race/ethnicity and allele frequencies is due to the correlation between race/ethnicity and genetic ancestry, so the inclusion of PCs as covariates is usually sufficient to account for confounding, and no additional adjustment for race/ethnicity may be necessary. On the other hand, for some analyses, inclusion of race/ethnicity as an additional covariate may further explain variation in the trait or disease of interest that is dependent on aspects of social identity (e.g. systemic or individual racial discrimination) rather than genetic ancestry. For example, African Americans with a high proportion of European ancestry may suffer the same lack of access to adequate health care as African Americans with little to no European ancestry. As another example, diet is correlated with many health outcomes and is often culturally driven, rather than driven by genetic ancestry. In such instances, including race/ethnicity as a covariate may improve statistical power to detect association.

[Back to top](#) [22]

Differences in trait variance by race, ethnicity, and ancestry

In addition to shifts in trait means by ancestry, the variance of a trait may also be different between different racial/ethnic groups and/or as a function of genetic ancestry. Genetic association tests may yield test statistics that are artificially high, known as inflation, if the differences in the trait variance are not properly accounted for in the statistical model. By allowing for heterogeneous residual variance (heteroskedasticity) across different groups, one is able to better control the inflation and false positive rate (Conomos, 2016). This is an especially important consideration for consortia such as TOPMed that comprise multiple component studies, with participants recruited from different geographic locations. While grouping by study is somewhat effective in reducing inflation, we have found that further splitting studies into racial/ethnic or genetic ancestry subgroups improves the statistical results. Unfortunately, splitting participants into discrete subgroups based on reported race/ethnicity or inferred genetic ancestry raises the same issues regarding missing data and non-homogeneity described in the stratified analysis

subsection above. Despite this, we have used reported race to define residual variance groups for many TOPMed analyses, utilizing HARE to impute missing values from a participant's ancestry PC values. This approach solves the missing data problem, but it does not solve the non-homogeneity problem introduced by requiring discrete groups. Nonetheless, because the grouping is only used to adjust the variance in the analysis, and we are not focused on inference regarding the variance component estimates, we contend that having some participants assigned to groups that may not be a particularly good fit will still provide better model performance than assuming homogeneity and not adjusting the variance at all. However, developing a model that allows for residual variance heteroskedasticity without requiring discrete grouping of subjects is of interest and an active area of research (Musharoff et al. 2018).

[Back to top](#) [22]

Considerations for investigators

1. **Articulate and justify why variables were used in a given analysis; in particular, describe analytical decisions to use non-genetic versus genetically inferred variables.** Analytical decisions are nuanced and often reflect a weighing of various pros and cons to different approaches.
2. **When using PCA to adjust for confounding due to genetic ancestry, use PCs that separate known populations.** The number of PCs for adjustment should be determined after examining the results, not *a priori*, e.g. 10 PCs.
3. **Pooled analysis including all samples and adjusting for PCs is recommended.** In considering additional stratified analysis, keep in mind that individuals within categorical racial/ethnic groups do not have homogeneous genetic ancestry. If using stratified analysis, describe why this approach was taken and what the limitations may be.
4. **If using race/ethnicity as a covariate, keep in mind that this variable may explain variation that is dependent on aspects of social identity, not genetics. Justify why this variable was used and how it contributes to the analysis.** In most cases, association between reported race/ethnicity and allele frequencies is due to the correlation between race/ethnicity and genetic ancestry, so the inclusion of PCs as covariates is usually sufficient to account for confounding.

[Back to top](#) [22]

Reporting

Race, ethnicity, and ancestry variables will inevitably be included in reports of TOPMed data and analyses including in presentations internal to the program, in abstracts and presentations for external conferences, and in peer-reviewed publications. Reporting on these variables and constructs is typically necessary to describe one's approach and methods as well as to provide interpretation of results. Past studies have found inadequate descriptions of race, ethnicity, and ancestry variables in the scientific literature (Ali-Khan, 2011; Fullerton, 2010), which can lead to both scientific and social harms (see [Motivation](#) [2]). In recognizing this challenge, some journals and funders have developed guidelines to help address these issues ("Preparing for Submission," ICMJE). Below we offer some recommendations on the reporting of race, ethnicity, and ancestry variables in TOPMed. Notably, these recommendations are meant to augment rather than supplant existing reporting recommendations from journals and funders.

[Back to top](#) [22]

Considerations for investigators

- **Consult with investigators from the TOPMed study/studies used in the analysis about any preferences or study-specific reporting guidelines.** For example, the HCHS/SOL cohort [recommends the term “Hispanic/Latino”](#) [24] rather than either term in isolation. However, the non-gendered term “Latinx” seems to be increasingly recommended for use. The Samoan Adiposity Study requests the term “Samoan” be used instead of “Pacific Islander” unless there are other Pacific Islander groups included. They also note that grouping SAS participants under “Other” perpetuates a bias towards majority groups. The BAGS study notes that the term “African Caribbean” is the most appropriate description of the ancestry of their participants. When considering group identity in a nested structure (e.g. calling a group a “subgroup”), be cognizant that this term may imply a hierarchy for some populations and that members of an assigned “sub” group may not consider themselves to be a part of the “larger group” being identified. Given the number and complexity of TOPMed studies, and the potential for conflicting study-specific recommendations in cross-study analyses, we strongly encourage authors to discuss these issues with TOPMed Working Group members and study representatives. Study investigators are also encouraged to send study-specific considerations and recommendations to the DCC (gccbiost@uw.edu [25]) to enable the creation of a centralized resource for researchers working with TOPMed data.
- **When invoking health disparities as a justification for genomic research, acknowledge the broader social context of health disparities.** Health disparities are differences in health “linked with economic, social, or environmental disadvantage” (“Disparities”, *Healthy People 2020*). While health disparities often disproportionately affect minority racial and ethnic groups, the underlying reasons are typically due to social and structural determinants of health rather than genetic factors (Sankar, 2005; Williams, 2010; Meagher, 2017). Genetic research may be part of the solution to address health disparities, but should be integrated into “social models of disease and interdisciplinary research methods” (West, 2017).

[Back to top](#) [22]

Conclusion

The diversity of the TOPMed program in terms of contributing studies, populations, genetic ancestries, and areas of phenotypic focus presents both analytical challenges and opportunities. Deliberate and considered decision-making around the use of race, ethnicity, and ancestry variables is essential for producing valid science and working to avoid harmful misappropriation of research findings in support of racism, discrimination or stigmatization. Methodological advancements are necessary to continue improving how researchers use and define these concepts; at the same time, awareness and sensitivity among researchers are needed to encourage good data stewardship, foster collaboration, and work towards expanding the diversity and representation needed to further translational genomic research (Sirugo, 2019).

[Back to top](#) [22]

Acknowledgements

The authors are grateful to the TOPMed ELSI Committee and TOPMed Analysis Committee for their feedback on, and support of, this project.

References

- Alexander, D. H., et al. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research*, vol. 19, no. 9, 2009, pp. 1655–1664., doi:10.1101/gr.094052.109.
- Ali-Khan, Sarah E., et al. "The Use of Race, Ethnicity and Ancestry in Human Genetic Research." *The HUGO Journal*, vol. 5, no. 1-4, July 2011, pp. 47–63., doi:10.1007/s11568-011-9154-5.
- "ASHG Denounces Attempts to Link Genetics and Racial Supremacy." *The American Journal of Human Genetics*, vol. 103, no. 5, 2018, p. 636., doi:10.1016/j.ajhg.2018.10.011.
- Bamshad, Michael, et al. "Deconstructing the Relationship between Genetics and Race." *Nature Reviews Genetics*, vol. 5, no. 8, 2004, pp. 598–609., doi:10.1038/nrg1401.
- Bentley, A.R., Callier, S. & Rotimi, C.N. Diversity and inclusion in genomic research: why the uneven progress?. *J Community Genet* 8, 255–266 (2017). <https://doi.org/10.1007/s12687-017-0316-6> [26]
- Braun, Lundy. "Reifying Human Difference: The Debate on Genetics, Race, and Health." *International Journal of Health Services*, vol. 36, no. 3, 2006, pp. 557–573., doi:10.2190/8jaf-d8ed-8wpd-j9wh.
- Brown, Anna. "The Changing Categories the U.S. Census Has Used to Measure Race." *Pew Research Center*, Pew Research Center, 25 Feb. 2020, www.pewresearch.org/fact-tank/2020/02/25/the-changing-categories-the-u-s... [27].
- Conomos, Matthew P et al. "Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos." *American journal of human genetics* vol. 98,1 (2016): 165-84. doi:10.1016/j.ajhg.2015.12.001
- "Disparities." *Disparities | Healthy People 2020*, <https://www.healthypeople.gov/2020/about/foundation-health-measures/Disp...> [28].
- Fang Huaying et al. "Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies." *American Journal of Human Genetics*, vol. 105, no. 4, Oct. 2019, pp. 763–772.
- Foster, Morris W. "Looking for Race in All the Wrong Places: Analyzing the Lack of Productivity in the Ongoing Debate about Race and Genetics." *Human Genetics*, vol. 126, no. 3, 2009, pp. 355–362., doi:10.1007/s00439-009-0674-1.
- Fujimura, Joan H., and Ramya Rajagopalan. "Different Differences: The Use of 'Genetic Ancestry' versus Race in Biomedical Human Genetic Research." *Social Studies of Science*, vol. 41, no. 1, July 2010, pp. 5–30., doi:10.1177/0306312710379170.
- Fullerton, Stephanie M., et al. "Population Description and Its Role in the Interpretation of Genetic Association." *Human Genetics*, vol. 127, no. 5, 2010, pp. 563–572., doi:10.1007/s00439-010-0800-0.

- Graves, JL. "Evolutionary Versus Racial Medicine" *Race and the Genetic Revolution: Science, Myth and Culture*, edited by Sheldon Krimsky and Kathleen Sloan, Columbia University Press, 2011, pp. 142 - 170.
- Harmon, Amy. "Why White Supremacists Are Chugging Milk (and Why Geneticists Are Alarmed)." *The New York Times*, The New York Times, 17 Oct. 2018, <https://www.nytimes.com/2018/10/17/us/white-supremacists-science-dna.html> [29].
- Krieger, Nancy. "Stormy Weather: Race, Gene Expression, and the Science of Health Disparities." *American Journal of Public Health*, vol. 95, no. 12, 2005, pp. 2155–2160., doi:10.2105/ajph.2005.067108.
- Lee, Sandra Soo-Jin, et al. "The Meanings of 'Race' in the New Genomics: Implications for Health Disparities Research." *Yale Law School Legal Scholarship Repository*, digitalcommons.law.yale.edu/yjhple/vol1/iss1/3/.
- Lee, Sandra, et al. "The Ethics of Characterizing Difference: Guiding Principles on Using Racial Categories in Human Genetics." *Genome Biology*, vol. 9, no. 7, 2008, p. 404., doi:10.1186/gb-2008-9-7-404.
- Manolio, Teri A. "Using the Data We Have: Improving Diversity in Genomic Research." *The American Journal of Human Genetics*, vol. 105, no. 2, 2019, pp. 233–236., doi:10.1016/j.ajhg.2019.07.008.
- Mathieson I, Scally A (2020) What is ancestry? PLoS Genet 16(3): e1008624. <https://doi.org/10.1371/journal.pgen.1008624> [30]
- Meagher, Karen M., et al. "Precisely Where Are We Going? Charting the New Terrain of Precision Prevention." *Annual Review of Genomics and Human Genetics*, vol. 18, no. 1, 2017, pp. 369–387., doi:10.1146/annurev-genom-091416-035222.
- Moses, Yolanda, et al. "Why Do We Keep Using the Word 'Caucasian'?" *SAPIENS*, Johann Friedrich Blumenbach/Wikimedia Commons, 7 Dec. 2016, <https://www.sapiens.org/column/race/caucasian-terminology-origin/> [31].
- Musharoff, Shaila, et al. "Existence and Implications of Population Variance Structure." *BioRxiv*, Cold Spring Harbor Laboratory, 1 Jan. 2018, www.biorxiv.org/content/10.1101/439661v1 [32].
- Novembre, John et al. "Genes mirror geography within Europe." *Nature* vol. 456,7218 (2008): 98-101. doi:10.1038/nature07331
- "Preparing for Submission." *ICMJE*, section d-i, <http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html> [33]
- Royal, Charmaine D., et al. "Inferring Genetic Ancestry: Opportunities, Challenges, and Implications." *The American Journal of Human Genetics*, vol. 86, no. 5, 2010, pp. 661–673., doi:10.1016/j.ajhg.2010.03.011.
- Sankar, Pamela. "Genetic Research and Health Disparities." *Jama*, vol. 291, no. 24, 2004, p. 2985., doi:10.1001/jama.291.24.2985.
- Shields AE, Fortun M, Hammonds EM, King PA, Lerman C, Rapp R, Sullivan PF; "The Use of Race Variables in Genetic Studies of Complex Traits and the Goal of Reducing Health Disparities: A

Transdisciplinary Perspective." *American Psychologist*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/15641924/ [34].

Shriner, Daniel. "Overview of admixture mapping." *Current protocols in human genetics* vol. Chapter 1 (2013): Unit 1.23. doi:10.1002/0471142905.hg0123s76

Sirugo, Giorgio, et al. "The Missing Diversity in Human Genetic Studies." *Cell*, vol. 177, no. 4, 2019, p. 1080., doi:10.1016/j.cell.2019.04.032.

Taliun et al. "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program" (2019). Available at: <https://www.biorxiv.org/content/10.1101/563866v1> [35] (preprint)

The 1000 Genomes Project Consortium. "A Global Reference for Human Genetic Variation." *Nature*, vol. 526, no. 7571, 2015, pp. 68–74., doi:10.1038/nature15393.

West, Kathleen Mcglone, et al. "Genomics, Health Disparities, and Missed Opportunities for the Nation's Research Agenda." *Jama*, vol. 317, no. 18, Sept. 2017, p. 1831., doi:10.1001/jama.2017.3096.

Williams DR, Mohammed SA, Leavell J, Collins C. Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. *Ann N Y Acad Sci*. 2010;1186:69–101. doi:10.1111/j.1749-6632.2009.05339.x

[Back to top](#) [22]

Source URL (modified on 06/10/2020 -

12:46pm):<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed>

Links

[\[1\] https://www.nhlbiwgs.org/printpdf/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed](https://www.nhlbiwgs.org/printpdf/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed) [2]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#motivation> [3]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#terminology> [4]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#considerations-a-id-co-nsiderations1-a-for-investigators> [5]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#harmonization-of-race-and-ethnicity-in-topmed> [6]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#data-collection-methods-often-vary> [7]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#harmonization-of-race-and-ethnicity-by-the-data-coordinating-center-dcc-> [8]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#considerations-a-id-co-nsiderations2-a-for-investigators> [9]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#genetic-ancestry> [10]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#considerations-a-id-co-nsiderations3-a-for-investigators> [11]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#analysis> [12]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#principal-components-adjustment> [13]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#stratified-analysis> [14]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#using-reported-race-ethnicity-as-a-covariate> [15]

<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#differences-in-trait-var>

[iance-by-race-ethnicity-and-ancestry](#) [16]
<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#considerations-a-id-co>
[nsiderations4-a-for-investigators](#) [17]
<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#reporting> [18]
<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#considerations-a-id-co>
[nsiderations5-a-for-i](#) [19]
<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#conclusion> [20]
<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#acknowledgements>
[21] <https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#references> [22]
<https://www.nhlbiwgs.org/guidelines-use-and-reporting-race-ethnicity-and-ancestry-topmed#top> [23]
<https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/racial-ethnic-minorities> [24]
https://sites.csc.unc.edu/hchs/sites/default/files/public_docs/Special%20Notes%20for%20Primary%20Authors%20of%20HCHS_revisedNov2018.pdf [25] <mailto:gcabiost@uw.edu> [26] <https://doi.org/10.1007/s12687-017-0316-6>
[27] <http://www.pewresearch.org/fact-tank/2020/02/25/the-changing-categories-the-u-s-has-used-to-measure-race/>
[28] <https://www.healthypeople.gov/2020/about/foundation-health-measures/Disparities> [29]
<https://www.nytimes.com/2018/10/17/us/white-supremacists-science-dna.html> [30]
<https://doi.org/10.1371/journal.pgen.1008624> [31]
<https://www.sapiens.org/column/race/caucasian-terminology-origin/> [32]
<http://www.biorxiv.org/content/10.1101/439661v1> [33]
<http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html> [34]
<https://pubmed.ncbi.nlm.nih.gov/15641924/> [35] <https://www.biorxiv.org/content/10.1101/563866v1>