Social and scientific motivations to move beyond groups in allele frequencies: The TOPMed experience

Sarah C. Nelson,[1,*] Stephanie M. Gogarten,[1] Stephanie M. Fullerton,[2] Carmen R. Isasi,[3] Braxton D. Mitchell,[4] Kari E. North,[5] Stephen S. Rich,[6] Matthew R.G. Taylor,[7] Sebastian Zöllner,[8,9] and Tamar Sofer[10,11,*]

Summary

For the genomics community, allele frequencies within defined groups (or "strata") are useful across multiple research and clinical contexts. Benefits include allowing researchers to identify populations for replication or "look up" studies, enabling researchers to compare population-specific frequencies to validate findings, and facilitating assessment of variant pathogenicity in clinical contexts. However, there are potential concerns with stratified allele frequencies. These include potential re-identification (determining whether or not an individual participated in a given research study based on allele frequencies and individual-level genetic data), harm from associating stigmatizing variants with specific groups, potential reification of race as a biological rather than a socio-political category, and whether presenting stratified frequencies—and the downstream applications that this presentation enables—is consistent with participants' informed consents. The NHLBI Trans-Omics for Precision Medicine (TOPMed) program considered the scientific and social implications of different approaches for adding stratified frequencies to the TOPMed BRAVO (Browse All Variants Online) variant server. We recommend a novel approach of presenting ancestry-specific allele frequencies using a statistical method based upon local genetic ancestry inference. Notably, this approach does not require grouping individuals by either predominant global ancestry or race/ethnicity and, therefore, mitigates re-identification and other concerns as the mixture distribution of ancestral allele frequencies varies across the genome. Here we describe our considerations and approach, which can assist other genomics research programs facing similar issues of how to define and present stratified frequencies in publicly available variant databases.

Introduction

Several databases publish allele frequencies, as well as other information on genetic variants and alleles, for use by the human genetics community. These databases include the National Center for Biotechnology Information's Single Nucleotide Polymorphism Database (dbSNP),[1] the Genome Aggregation Database (gnomAD),[2] the Functional Annotation of Variants - Online Resource (FAVOR),[3] and the TOPMed-specific variant server BRAVO (see web resources). TOPMed is one of the largest collections of whole-genome sequences to date, with >78.7% of variants discovered not previously reported in dbSNP[4] and ~158,000 whole-genome sequences available in dbGaP (see web resources). Indeed, TOPMed data in the BRAVO server may be the only public resource with a record of a given variant of interest. Yet, unlike other resources providing allele frequencies based on data that are not TOPMed-specific, BRAVO initially provided only TOPMed-wide allele frequencies, i.e., not "stratified" by genetic ancestry, race/ethnicity, study membership, or any other demographic or genetic features. Notably, TOPMed is a consortium aggregating whole-genome sequencing data from ~80 parent studies of diverse genetic ancestries, race/ethnicities and study designs. Therefore, to share stratified allele frequencies, TOPMed investigators had to decide how to stratify allele frequencies, taking into consideration the perspectives of the TOPMed parent cohorts.

To facilitate decision-making, the TOPMed ELSI (Ethical, Legal, and Social Issues) Committee discussed potential benefits and concerns with adding stratified frequencies to the BRAVO server. Here, we summarize concerns of TOPMed studies with respect to sharing of allele frequencies, describe the advantages and disadvantages of potential approaches for stratification, and explain the Committee's final recommendations, which were later approved by the TOPMed Executive Committee. While these discussions are guided by the TOPMed-specific experience, we contend this summary will be useful to other studies and consortia considering data sharing of summarized genetic information, such as allele frequencies.

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [2]Department of Bioethics and Humanities, University of Washington, Seattle, WA 98195, USA; [3]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA; [4]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [5]Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599, USA; [6]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA 22903, USA; [7]Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA; [8]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [9]Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109, USA; [10]Department of Medicine, Harvard Medical School, Brigham and Women's Hospital, Boston, MA 02115, USA; [11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
*Correspondence: sarahcn@uw.edu (S.C.N.), tsofer@bwh.harvard.edu (T.S.)
https://doi.org/10.1016/j.ajhg.2022.07.008.

Benefits of sharing allele frequencies

Allele frequencies are used by scientists and clinicians in multiple ways, including: (1) validation of a research result, e.g., when a trait-variant association is detected, verifying that the estimated allele frequency matches known frequencies; (2) identification of a population in which a variant is relatively common in order to design a replication study; (3) variant interpretation in a clinical context, i.e., as evidence for classifying variants as "pathogenic," "benign," or "of uncertain significance"; and (4) population genetics research, where allele frequencies can be used to study population history and evolutionary processes.

Stratified allele frequencies may be more helpful to the research community compared to population-wide frequencies. For example, stratified frequencies may allow for the identification of a population/group for a replication study following a genome-wide association study (GWAS), while population-wide frequencies generally will not. In addition, population genetic research may only be done with stratified frequencies. Similarly, given differences of allele frequencies across genetic ancestries, quality control procedures comparing allele frequencies computed in the new study to those in a database will be more precise when the frequencies in the database are inferred from a population with similar genetic ancestry to those in the new study. Finally, allele frequency may help inform clinical variant interpretation: current guidelines from the American College of Medical Genetics and Genomics (ACMG)[5] state: "If a variant is absent from (or below the expected carrier frequency if recessive) a large general population or a control cohort (>1,000 individuals) and the population is race-matched to the patient harboring the identified variant, then this observation can be considered a moderate piece of evidence for pathogenicity." Notably, allele frequency *greater* than expected given the disorder, or over 5%, can support a benign interpretation—a point on which the guidelines do not require a matched population. For example, identifying a variant as common in a specific population of similar ancestry, though rare in the U.S. overall, could lead to fewer unnecessary medical interventions for members of that population with the variant.

Concerns in sharing stratified allele frequencies

Despite the above benefits of stratified frequencies, there are also potential concerns, especially when the frequencies are made freely and publicly available via unrestricted-access variant servers. We group these concerns into four main categories below: re-identification, stigmatization, reification of race, and participant informed consent.

Re-identification

A risk of re-identification refers to the risk of determining whether an individual participated in a given study.[6] Previous research showed that, given genetic data (possibly chip-based) from an individual, one can use allele frequencies for a set of genetic variants measured in a specific study to determine whether the individual was a study participant. This information could be stigmatizing or otherwise compromise participant privacy, especially if study participation reveals sensitive phenotypic information about the individual. Studies from defined geographic areas and/or comprising families may have heightened concerns about re-identification.

Allele frequencies are a form of genomic summary result (GSR), for which NIH data sharing policies have evolved over the past two decades. Initially GSRs were made publicly available in NIH-designated repositories. GSRs were later moved out of public access following Homer et al.'s[7] findings about re-identification. After much deliberation, the NIH GSR sharing policy was revised again in 2018 such that unless a study was designated as "sensitive," GSRs could be made publicly available—i.e., unrestricted access (see web resources). Other publications have studied the risk of re-identification using GSRs but assuming that results from a genetic association study are available (i.e., including effect size estimates). For example, Lumley and Rice[7] showed that a phenotype prediction can be constructed for an individual based on their genetic data and effect-size estimates from GWAS. However, relevant to our question is the approach of using allele frequencies rather than effect estimates. Notably, in either case, re-identification risks are expected to decrease with the size and genetic heterogeneity of the research cohort,[6–9] suggesting mitigated risk if the sample size of the strata is large enough.

Stigmatization

There are potential harms from associating a group with stigmatizing phenotypes or other outcomes when pathogenic or other associated variants are present or common only in that group. Stigmatization is a type of group harm that may result in members of a group being "ostracized, humiliated, or discriminated against, resulting in the loss of social, economic, or political opportunities or even loss of face in the society."[10] Group-level harms are particularly salient in discussions of allele frequencies and other aggregate, summary-level data, as (1) features of those data are often annotated by, or otherwise attached to, socially identifiable groups (whether based on ancestry, demographics, geography, culture, or other affiliation) and (2) secondary use of data may be subject to little or no ethical oversight,[11] especially when available via publicly available variant servers. Notably, group harms are distinct from individual-level risks, such as loss of privacy or autonomy, and are often overlooked in ethical and regulatory frameworks that focus on individual-level protections.[12–15]

An illustration of potential group harms comes from the case of the Havasupai Tribe versus the Arizona Board of Regents.[16,17] While the tribe's understanding was that their samples were going to be used only for diabetes research, samples were later used for research on schizophrenia

and alcoholism, phenotypes considered stigmatizing to the tribe. While we are not aware of specific instances where allele frequencies generated by genetic research on Havasupai tribal members led to harm, the example illustrates how the *connection* of sensitive phenotypes to a socially identifiable group—whether from the pursuit of a specific research question or when a risk allele is found at higher frequency compared to other groups—can stigmatize the group and also impact future research efforts. Notably, the likelihood of group stigmatization will be influenced by the group identity; its historical experiences of racism, stigma, and discrimination; the phenotype in question; and other aspects of social context.[18]

Reification of race

Stratification may encourage or reinforce genetic or biological conceptions of race, which perpetuate a harmful history of scientific racism in the field of genetics.[19–22] Specifically, presenting genetic information (e.g., variant frequencies) using racial or ethnic groupings suggests those groupings are defined or distinguished by genetic differences. To the contrary, race is a social construct, and overwhelming genetic evidence exists to refute rather than support the idea of "biologically distinct subcategories" of humans.[21] This practice would also not meet the call to promote anti-racism in science.[20] As a practical matter, U.S.-based race categories, e.g., those defined by the 1997 Office of Management and Budget standards on race and ethnicity, do not apply to study participants outside of the US, which complicates stratifying allele frequencies in variant databases with an international collection of studies (such as available in TOPMed).

Stratification based on genetic ancestry rather than race or ethnicity is also potentially damaging. For example, prior studies have shown that genetic ancestry, when conceptualized at the continental level, can still be mapped onto common notions of race.[23] Lewis and colleagues have recently called for the abandonment of continental ancestry groupings, citing the problematic confounding between these categories and racial classifications.[24] Moreover, they note admixture analysis as compounding, rather than solving, this issue, as genomes of individuals are still described using a mixture of continental-level labels. Therefore, while statistical methods that do not require creation of discrete groups based on either genetic or social definitions may mitigate risks of conflating genetic ancestry with race, further empirical ELSI research is needed to explore the implications of using ancestry labels in allele frequency databases.

Consistency with informed consent

A study's decision to participate in the computation of stratified allele frequencies requires (1) interpreting whether potential uses of allele frequencies are consistent with participants' informed consent and (2) determining to what extent, from ethical and/or regulatory standpoints, individual-level consent constrains uses of summary-level data. Informed consent processes for large-scale genomic research are complicated,[25] including by potential downstream data uses often unconceived or unknowable at the time of participant recruitment. This leads to a common situation of requiring investigators and their ethics oversight boards to interpret legacy consent for new and potentially novel applications, such as a publicly available variant server. For example, for many TOPMed studies, participants were not explicitly asked at the time of recruitment about preferences for sharing summary results. However, this does not absolve investigators and institutions from considering whether such downstream uses are consistent, or at least not inconsistent, with participant wishes and understandings at the time of consent.

At the initiation of the BRAVO variant server, when only TOPMed-wide frequencies had been shared, TOPMed investigators were asked to determine whether their study could contribute to it and, if so, for which consent groups. In TOPMed, consents vary considerably between studies, and even within studies, especially for studies spanning multiple recruitment sites and institutions. In ongoing (versus legacy) studies, consents may also evolve over time. For broad consent such as general research use (GRU), the justification to contribute to BRAVO was fairly straightforward. Notably, at least one GRU study opted out due in part to concerns of downstream commercialization where the source population may not share the benefit. For the narrowest consent categories of disease-specific research, all TOPMed studies but one still agreed to contribute. In between broad GRU and disease-specific sits the consent category of health/medical/biomedical (HMB), which in the NIH standardized description "does not include the study of population origins or ancestry" (see web resources). Therefore, for studies with HMB consent, investigators had to consider whether presenting frequencies would constitute or enable "study of population origins or ancestry" which—admittedly—is not straightforward. Ultimately, most HMB studies elected to contribute to BRAVO. As stratified allele frequencies may enable additional downstream uses, compared to TOPMed-wide frequencies, TOPMed studies need to re-assess the consistency of involvement with their participants' informed consent.

The decision to share allele frequencies may be constrained by both ethical and regulatory standpoints. We bring both NIH policy and TOPMed precedent to bear on this question. First, as noted above, NIH policy on GSRs has evolved over time. Per the November 2018 policy update (see web resources), GSRs can be shared publicly and openly unless a study is designated as "sensitive." Uses of open access GSRs are typically constrained, if at all, only by user agreements (such as the BRAVO Terms of Use, which users attest to when creating a login; see web resources). For GSRs from sensitive studies, which can be accessed only by application, use is subject to the same limitations as the individual-level data. Thus, from a regulatory standpoint, individual consent limits the use of summary-level data for sensitive studies but not for non-sensitive studies.

Notably, prior to the NIH GSR policy update in 2018, the TOPMed ELSI Committee used a "publication analogy" to consider whether downstream uses of summary data should be constrained by individual-level consent.[26] Briefly, the publication analogy is the concept that once data are published in a journal article, they enter the public domain and may be used beyond the limits defined during the original participant consent process. The publication analogy could reasonably be applied to presentation of stratified allele frequencies in BRAVO, as it was previously to TOPMed-wide frequencies in BRAVO. Ultimately, the ELSI Committee's recommendation to TOPMed studies in considering expanding from TOPMed-wide to stratified frequencies was to rely on their investigators' expertise and experience, and the institutional history within the study, to make an informed decision on whether to contribute.

Considerations for implementation

There are multiple approaches to defining strata for computation of allele frequencies, each of which has advantages and disadvantages. In TOPMed, the specifics of proposed strata will likely influence whether study investigators agree to have their study data included. Here we describe potential strata definitions, illustrated in Figure 1 and summarized in Table 1, and note considerations for different options. In brief, allele frequencies can be computed in individuals grouped according to common characteristics, e.g., by social definitions of race/ethnicity or based on genetic ancestry patterns (grouping-based Approach 1 in Figure 1). Alternatively, allele frequencies can be computed without defining groups, but rather by first inferring the genetic ancestral background of each individual in the data and then using this inference to deconvolve frequencies (Approach 2 in Figure 1). The latter approach acknowledges that individual genomes are a mixture of genetic ancestries, and once the distribution of these ancestries within this mixture is known, one can infer frequencies of variants in each of these ancestries.

The TOPMed ELSI Committee recommended estimating ancestry-specific allele frequencies utilizing local genetic ancestry inference[27] and appropriate statistical methods. Investigators at the TOPMed Informatics Research Center previously performed local and global genetic ancestry inferences for TOPMed genomes. They condensed the 53 Human Genome Diversity Project (HGDP)[28] reference populations into seven "super populations"[29]—Europe, Middle East, Africa, Central and East Asia, South Asia, America, and Oceania—and assigned genetic ancestries to TOPMed samples. Specifically, each genome is divided into segments (local ancestry intervals, each encompassing a range of haplotypes). It is possible to compute ancestry-specific allele frequencies using statistical algorithms applied across all available TOPMed participants by incorporating local ancestry information.[30]

Computing ancestry-specific allele frequencies using local allele frequencies and statistical deconvolution (Approach 2 in Figure 1) in TOPMed alleviates the major limitations of the grouping-based approaches. First, the use of genetically inferred measures avoids harmonizing demographic categories across TOPMed. Demographic categories differ across studies for various reasons, including data collection methods and the differences in race/ethnicity categories in the countries in which the studies were undertaken, complicating harmonization. Furthermore, grouping approaches, whether based on social categories or genetic ancestry, risk excluding individuals who do not satisfy the conditions to be assigned to a specific group. The statistical deconvolution using local ancestry approach further limits the reification of race as a biological variable by not suggesting that race-based groupings correspond to genetic make-up—a limitation of other stratification schemes. Estimation of ancestry-specific allele frequencies also mitigates re-identification concerns because the mixture distribution of ancestral allele frequencies varies across the genome (i.e., by local ancestry). Therefore, applying this approach further does not require specifying a minimum number of TOPMed studies or of individuals in order to report a specific strata, unless a specific genetic ancestry is uniquely and completely represented by a specific TOPMed study. In contrast, when using grouping-based approaches, mitigating the risk of re-identification requires setting a minimum number of participants or studies per group/stratum and identifying that minimum threshold.

The computation of local ancestry data depends on the availability and selection of reference populations to define the possible ancestries. Currently, TOPMed will use the previously computed local ancestries corresponding to the seven super populations represented in the HGDP reference. Notably, each of these categories is composed of multiple, smaller sets of individuals from fine-scale categories. For example, the Europe super population includes individuals identified as French, Basque, Sardinian, Tuscan, and others. Each of these groups could in principle be used as a specific ancestry. Population genetics research highlights the usefulness of fine-scale population structure for human population history research and for medical genetics.[31–33] Recent studies using large genetic datasets, e.g. from the UK Biobank[34] and from the BioMe Biobank in New York City,[32] demonstrated that fine-scale population structure can be detected and may affect both epidemiological and medical research, further suggesting that more granular ancestries may be useful for allele frequency stratification.

However, there are limitations for using finer-scale ancestries for stratified allele frequencies. Critically, distinguishing between two closely related genetic ancestries (e.g., French and Basque) in local ancestry inference requires sufficiently large reference data, and using this information to compute accurate ancestry-specific allele frequency at a variant further requires a sufficient

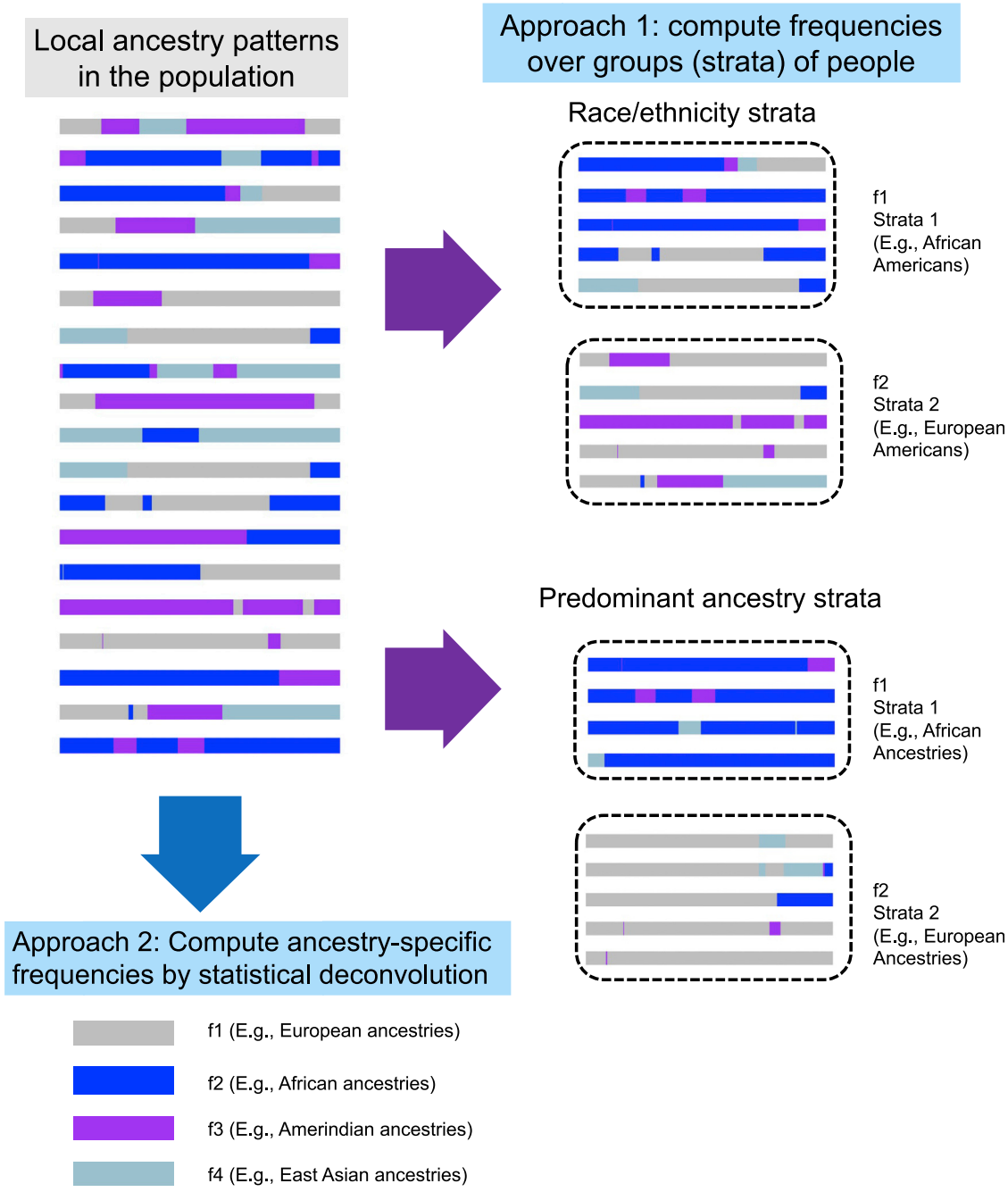

Figure 1. Visualization of approaches for computing stratified allele frequencies
This figure visualizes potential types of allele frequency stratification and their challenges by demonstrating admixture. Local ancestry patterns were simulated at random to generate admixed genomes, and each interval is colored by its sampled ancestry. Approach 1 for computing stratified allele frequencies relies on groupings of individuals based on self-reported race/ethnicity (top right, independent of inferred admixture patterns) or on identification and grouping of individuals whose genomes are mostly from a specific inferred ancestry (bottom right). Individuals may be excluded from grouping approaches due to missing race/ethnicity or high admixture. Approach 2 uses all individuals and relies on local ancestry inferences to compute ancestry-specific allele frequencies across the genomes. The use of the plural terms to describe continental ancestries (e.g., "European ancestries") emphasizes the fact that any selected ancestry is a reflection of a somewhat arbitrary reference population, encompassing a set of finer-scaled ancestries.

number of individuals having the relevant local ancestry at the variant.[30] Also, the decision as to which ancestries to include is not straightforward. Genetic ancestry is a continuum, representing the constant mixing of individuals across groups, rather than a collection of discrete categories.[35] Considering shorter genomic intervals for ancestry inference will lead to inference of more ancient admixture,[36] with potentially different ancestral populations.

We recognize that a local ancestry-based approach may still pose some of the same risks noted above, as it still relies on ancestry labels. A discussion published by Lewis

Table 1. The three types of stratified frequencies considered by the TOPMed ELSI Committee, including their advantages and disadvantages

	Ancestry-specific allele frequencies by statistical deconvolution	Ancestry-specific allele frequencies by grouping individuals according to ancestral patterns	Race/ethnicity group-specific allele frequencies
Grouping approach	No	Yes	Yes
All participants can contribute to analysis?	Yes	Only those who are "mostly" from one ancestry (i.e., predominant ancestry)	Only those with reported race/ethnicity
Risk of re-identification?	Very low; re-identification methods do not currently exist	Risk exists for groups with small sample size	Risk exists for groups with small sample size
Risk of reifying race as a biological variable?	Low	Medium (can be conflated with genetic ancestry)	High

Ancestry-specific allele frequencies by statistical deconvolution are computed based on inferred local ancestry patterns for individuals in the dataset. Ancestry-specific allele frequencies by grouping of individuals requires global ancestry inference and selection of a threshold of a minimum percentage of a specific genetic ancestry to categorize an individual into a genetic ancestry group. Race/ethnicity groups may be based on self-report or external ascription.

et al.[24] contends that categories, or labels, of genetic ancestries may lead to essentialization of the groups used to define genetic ancestry and the differences between them. It is possible that fine-scale ancestry will increase the risk of essentializing the biological notion of the groups from which these ancestries originated (e.g., an emphasis on Ashkenazi Jewish ancestry). While we do not propose to estimate frequencies by defined groups of people, tracking the estimated stratified frequencies to a set of "groups of origin" may still risk reification of the biological basis of social racial and ethnic categories.

An additional risk of our local ancestry-based approach stems from one of the benefits: the ability to include individuals and report on ancestry groups that would not be included or represented in a predominant ancestry approach. To illustrate this concern, we invoke the concept of admixture that, as noted above, is not without issue. Specifically, Hispanics/Latinos are admixed with three major categories of ancestries— European, African, and Amerindian—enabling the estimation of Amerindian allele frequencies, that, in the past, would only have been estimated from Native Americans/American Indians. Indigenous peoples are under-represented and understudied in genetic research for a number of reasons, including potential stigmatization, the undermining of tribal sovereignty, and lack of trustworthiness demonstrated by researchers.[37,38] Sharing Amerindian allele frequencies estimated from Hispanics/Latinos may be viewed, therefore, as a problematic "back door" access. Furthermore, these frequencies can potentially lead to group harm to American Indians. For example, as allele frequencies may enable population genetics research, they may be used to imply population history that is different from a population's own narrative, which may be used to undermine a population's legal claims to specific territories.[15] In TOPMed, however, we think that this risk is low, as Amerindian ancestry allele frequencies are estimated from Hispanics/Latinos (primarily) from diverse backgrounds (Mexican, Cuban, etc.) and are highly unlikely to be traceable to specific tribes.

There are additional limitations to the selected approach to estimating stratified allele frequencies. First, allele frequency estimates using statistical deconvolution may be less precise, compared to standard estimates, when a variant is only rarely available from a specific ancestry. Therefore, TOPMed may not be able to confidently share stratified variant frequencies for rare variants. TOPMed will continue to provide global (unstratified) TOPMed frequencies for rare variants. Second, this approach would not accommodate a study interested in sharing study-specific allele frequencies, e.g., for studies from a unique founder population. Such population-specific allele frequencies could be useful in improving clinical care in those populations. TOPMed may incorporate other mechanisms that would allow such sharing, e.g., by specific requests from such studies. Finally, implementing this approach requires the technical expertise and resources to calculate local ancestries.

Conclusions

Maximizing the sharing of genomic data from large-scale sequencing programs such as TOPMed is important for ensuring that the resulting resources are fully utilized by the scientific community. However, the details of how and where data are shared become critical for maximizing utility while also maintaining participant privacy and trust in the research enterprise.[39] NIH policy for sharing allele frequencies, a form of GSR, has fluctuated over the years from open, to restricted, and back to mostly open sharing—presumably in an attempt to balance potential risks with the benefits of open sharing. Notably, stratified allele frequencies are useful for quality control, design of replication studies, population genetics research, and clinical variant interpretation. While publicly available variant servers make frequency and other variant information easily accessible, how frequencies are calculated and presented has important scientific and social consequences.

The TOPMed program considered the risks and benefits of whether and how to add stratified frequencies to the BRAVO

variant server. Ultimately, we recommend providing ancestry-specific allele frequencies by statistical deconvolution, a method that does not require grouping individuals into demographic, genetic ancestry, or any other fixed category. This approach has the benefits of minimizing reification of race, maximizing use of available data, and mitigating risks of re-identification and reputational harms (see Table 1). An additional benefit is the potential to represent populations that are not generally available in a predominant ancestry framework (e.g., Amerindian ancestry). Overall, the risks associated with sharing allele frequencies outside the context of a specific GWAS (and therefore a specific phenotype) are small and become even smaller when combining populations from multiple contributing studies. Further, while we have drawn upon literature of potential harms and benefits to individuals and groups from genetic research more broadly, more empirical research is needed to assess both individual- and group-level implications of sharing stratified allele frequencies specifically.

Clinical variant interpretation is an evolving science; we recognize the clinical utility of our proposed approach to presenting stratified allele frequencies is not yet known. As a practical matter, clinicians may use patient demographics, including socio-political race/ethnicity categories, to select an appropriate population group against which to compare allele frequency.[5] The wider availability of ancestry-specific frequencies may change this practice. The effects of genetic ancestry on pathogenicity are still not studied (e.g., gene × gene interaction), and it is not clear whether ancestry-specific frequencies are indeed more useful compared to frequencies estimated on groups defined using social constructs such as race or ethnicity. However, ancestry-specific frequencies are an attempt at a more accurate representation of an individual's genome as a mosaic of local ancestry patterns and may help push evaluation of patient genomics in a non-categorical direction.

In addition to sharing stratified frequencies with the community by making them publicly available and supporting uses outlined above, stratified frequencies may also be used by TOPMed investigators to empower statistical genetic analyses. For example, investigators may apply admixture mapping methods while focusing on "ancestry-enriched" variants to increase power for discovery of genetic loci associated with complex traits in a diverse, admixed dataset.[40] Other future approaches may prioritize variants that tend to be more common in a given genetic ancestry to potentially improve polygenic risk prediction in individuals with a high genomic proportion of the same ancestry, to alleviate the problem of limited generalizability and portability of polygenic risk scores across populations, which may be related to genetic ancestry to some extent.[41,42]

Here we have recounted the TOPMed experience of working to add stratified allele frequencies to the BRAVO variant server in a way that is scientifically useful and socially responsible. These frequencies are not yet available on BRAVO, as some TOPMed studies are still deliberating. Our proposed approach mitigates concerns about various "grouping" strategies and more accurately reflects underlying genetic architecture. It also serves as an example for other genomics resources facing similar challenges. Importantly, we hope that this approach will improve genomic research in under-represented populations by enabling more accurate quality control for genetic data, identifying opportunities for replication of findings from GWASs, and improving clinical variant interpretation.

Acknowledgments

Molecular data for the Trans-Omics for Precision Medicine (TOPMed) program was supported by the National Heart, Lung, and Blood Institute (NHLBI). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering, were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Until October 2021, core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. K.E.N. is supported by R01HD057194, R01DK122503, R01HG010297, R01HL142302, R01HL143885, R01HG009974, and R01DK101855. T.S. is supported by R21AG070644, R21HL145425, and R03OD030598.

Declaration of interests

The authors declare no competing interests.

Web resources

NIH 2018 Update to NIH Management of Genomic Summary Results Access, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html

NIH Scientific Data Sharing, Restrictions: Data Use Limitations, https://sharing.nih.gov/genomic-data-sharing-policy/institutional-certifications/completing-an-institutional-certification-form#step-5

TOPMed ELSI Committee reports, https://topmed.nhlbi.nih.gov/elsi

TOPMed Whole Genome Sequencing Methods: Freeze 9, https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-9

TOPMed variant server, BRAVO (Browse All Variants Online) Home Page, https://bravo.sph.umich.edu/

TOPMed variant server, BRAVO (Browse All Variants Online) Terms of Use, https://bravo.sph.umich.edu/freeze8/hg38/terms

References

1. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
2. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A.,

Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141, 456 humans. Nature *581*, 434–443.
3. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet. *52*, 969–983.
4. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53, 831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.
5. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of medical genetics and genomics and the association for molecular pathology. Genet. Med. *17*, 405–424.
6. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., and Craig, D.W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. *4*, e1000167.
7. Lumley, T., and Rice, K. (2010). Potential for revealing individual-level information in genome-wide association studies. JAMA *303*, 659–660.
8. Visscher, P.M., and Hill, W.G. (2009). The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet. *5*, e1000628.
9. Bacanu, S.-A. (2017). Sharing extended summary data from contemporary genetics studies is unlikely to threaten subject privacy. PLoS One *12*, e0179504.
10. McGregor, J. (2010). Racial, ethnic, and tribal classifications in biomedical research with biological and group harm. Am. J. Bioeth. *10*, 23–24.
11. Fullerton, S.M., and Lee, S.S.-J. (2011). Secondary uses and the governance of de-identified data: lessons from the human genome diversity panel. BMC Med. Ethics *12*, 16.
12. Tsosie, K.S., Yracheta, J.M., and Dickenson, D. (2019). Overvaluing individual consent ignores risks to tribal participants. Nat. Rev. Genet. *20*, 497–498.
13. Tsosie, R. (2007). Cultural challenges to Biotechnology: native American genetic resources and the concept of cultural harm. J. Law Med. Ethics *35*, 396–411.
14. Hausman, D. (2008). Protecting groups from genetic research. Bioethics *22*, 157–165.
15. Greely, H.T. (2001). Informed consent and other ethical issues in human population genetics. Annu. Rev. Genet. *35*, 785–800.
16. Garrison, N.A. (2013). Genomic justice for native Americans: impact of the Havasupai case on genetic research. Sci. Technol. Human Values *38*, 201–223.
17. Sterling, R.L. (2011). Genetic research among the Havasupai: a cautionary tale. Virtual Mentor. *13*, 113–117.
18. de Vries, J., Jallow, M., Williams, T.N., Kwiatkowski, D., Parker, M., and Fitzpatrick, R. (2012). Investigating the potential for ethnic group harm in collaborative genomics research in Africa: is ethnic stigmatisation likely? Soc. Sci. Med. *75*, 1400–1407.
19. Race Ethnicity and Genetics Working Group (2005). The use of racial, ethnic, and ancestral categories in human genetics research. Am. J. Hum. Genet. *77*, 519–532.
20. Yudell, M., Roberts, D., DeSalle, R., Tishkoff, S.; and 70 signatories (2020). NIH must confront the use of race in science. Science *369*, 1313–1314.
21. ASHG (2018). ASHG denounces attempts to link genetics and racial supremacy. Am. J. Hum. Genet. *103*, 636.
22. Brothers, K.B., Bennett, R.L., and Cho, M.K. (2021). Taking an antiracist posture in scientific publications in human genetics and genomics. Genet. Med. *23*, 1004–1007.
23. Fujimura, J.H., and Rajagopalan, R. (2011). Different differences: the use of "genetic ancestry" versus race in biomedical human genetic research. Soc. Stud. Sci. *41*, 5–30.
24. Lewis, A.C.F., Molina, S.J., Appelbaum, P.S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S.M., Garrison, N.A., Ghosh, N., Hammonds, E.M., et al. (2022). Getting genetic ancestry right for science and society. Science *376*, 250–252.
25. McGuire, A.L., and Beskow, L.M. (2010). Informed consent in genomics and genetic research. Annu. Rev. Genomics Hum. Genet. *11*, 361–381.
26. NHLBI (2016). NHLBI TOPMed WGS ELSI Committee Report: Imputation Server. Available at. https://topmed.nhlbi.nih.gov/sites/default/files/ELSI%20Review%20of%20TOPMed%20Imputation%20Server.pdf.
27. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
28. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.
29. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012.
30. Granot-Hershkovitz, E., Sun, Q., Argos, M., Zhou, H., Lin, X., Browning, S.R., and Sofer, T. (2022). AFA: ancestry-specific allele frequency estimation in admixed populations: the hispanic community health study/study of latinos. Hum. Genet. Genomics Adv., 100096.
31. Novembre, J., and Peter, B.M. (2016). Recent advances in the study of fine-scale population structure in humans. Curr. Opin. Genet. Dev. *41*, 98–105.
32. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.
33. Belbin, G.M., Cullina, S., Wenric, S., Soper, E.R., Glicksberg, B.S., Torre, D., Moscati, A., Wojcik, G.L., Shemirani, R., Beckmann, N.D., et al. (2021). Toward a fine-scale population health monitoring system. Cell *184*, 2068–2083.e11.
34. Cook, J.P., Mahajan, A., and Morris, A.P. (2020). Fine-scale population structure in the UK Biobank: implications for genome-wide association studies. Hum. Mol. Genet. *29*, 2803–2811.
35. Mathieson, I., and Scally, A. (2020). What is ancestry? PLoS Genet. *16*, e1008624.
36. Gravel, S. (2012). Population genetics models of local ancestry. Genetics *191*, 607–619.
37. Chadwick, J.Q., Copeland, K.C., Branam, D.E., Erb-Alvarez, J.A., Khan, S.I., Peercy, M.T., Rogers, M.E., Saunkeah, B.R.,

Tryggestad, J.B., and Wharton, D.F. (2019). Genomic research and American Indian tribal communities in Oklahoma: learning from past research misconduct and building future trusting partnerships. Am. J. Epidemiol. *188*, 1206–1212.

38. Claw, K.G., Anderson, M.Z., Begay, R.L., Tsosie, K.S., Fox, K., Garrison, N.A.; and Summer internship for INdigenous peoples in Genomics (SING) Consortium (2018). A framework for enhancing ethical genomic research with Indigenous communities. Nat. Commun. *9*, 2957.
39. Trinidad, S.B., Fullerton, S.M., Ludman, E.J., Jarvik, G.P., Larson, E.B., and Burke, W. (2011). Research ethics. Research practice and participant preferences: the growing gulf. Science *331*, 287–288.
40. Shriner, D. (2013). Overview of admixture mapping. Curr. Protoc. Hum. Genet. *76*, 1.23.1–1.23.8.
41. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun. *10*, 3328.
42. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Current clinical use of polygenic scores will risk exacerbating health disparities. Nat. Genet. *51*, 584–591.