Title

Experience using conventional compared to ancestry-based population descriptors in clinical genomics laboratories.

Permalink
https://escholarship.org/uc/item/0xq3s0tj

Journal
American Journal of Human Genetics, 112(3)

Authors
Hatchell, Kathryn
Poll, Sarah
Russell, Emily
et al.

Publication Date
2025-03-06

DOI
10.1016/j.ajhg.2025.01.008

Peer reviewed

Experience using conventional compared to ancestry-based population descriptors in clinical genomics laboratories

Kathryn E. Hatchell,[1,*] Sarah R. Poll,[1] Emily M. Russell,[1] Trevor J. Williams,[1] Rachel E. Ellsworth,[1] Flavia M. Facio,[1] Sienna Aguilar,[1] Edward D. Esplin,[1] Alice B. Popejoy,[2,3] Robert L. Nussbaum,[4] and Swaroop Aradhya[5,6]

Summary

Various scientific and professional groups, including the American Medical Association (AMA), American Society of Human Genetics (ASHG), American College of Medical Genetics (ACMG), and the National Academies of Sciences, Engineering, and Medicine (NASEM), have appropriately clarified that certain population descriptors, such as race and ethnicity, are social and cultural constructs with no basis in genetics. Nevertheless, these conventional population descriptors are routinely collected during the course of clinical genetic testing and may be used to interpret test results. Experts who have examined the use of population descriptors, both conventional and ancestry based, in human genetics and genomics have offered guidance on using these descriptors in research but not in clinical laboratory settings. This perspective piece is based on a decade of experience in a clinical genomics laboratory and provides insight into the relevance of conventional and ancestry-based population descriptors for clinical genetic testing, reporting, and clinical research on aggregated data. As clinicians, laboratory geneticists, genetic counselors, and researchers, we describe real-world experiences collecting conventional population descriptors in the course of clinical genetic testing and expose challenges in ensuring clarity and consistency in the use of population descriptors. Current practices in clinical genomics laboratories that are influenced by population descriptors are identified and discussed through case examples. In relation to this, we describe specific types of clinical research projects in which population descriptors were used and helped derive useful insights related to practicing and improving genomic medicine.

Introduction

Different group-based descriptors have been devised to classify humans into population subgroups, which are dynamic and context dependent, with major changes over time. Race and ethnicity (“conventional population descriptors”) categories in the United States are often self-identified or assigned to people based on socially determined factors, such as physical appearance, language, cultural practices, or family history, and have been in use for demographic or sociological purposes for hundreds of years (https://www.census.gov/newsroom/blogs/random-samplings/2015/11/measuring-race-and-ethnicity-across-the-decades-1790-2010.html).

More recently, population genetics approaches have been used to subgroup and stratify populations for group-based analyses. While most genetic variants are commonly shared among all human populations and are not specific to socio-cultural or broad geographic groups, some variants do have ancestry-derived frequencies. The frequencies of these polymorphic alleles are more similar among individuals whose predecessors lived in a particular geographic region with limited reproductive contact outside the region.[1] Population descriptors named after these regions can vary in size and resolution and are typically based on the geographic location or how local individuals self-identify.[2,3] An individual may be assigned a particular genetic ancestry (“ancestry-based population descriptor”) based on the similarity of their genotypes at these polymorphic loci to the genotypes characteristic of various population-based reference data from different geographic areas.

Debate continues on the use of both conventional and ancestry-based population descriptors in health research.[4] For example, as social constructs, conventional population descriptors remain useful in evaluating disparities in the etiology of disease, healthcare delivery, and health outcomes[5]; eliminating the use of race/ethnicity may exacerbate inequalities in health outcomes.[6] These conventional population descriptors have limitations, including imprecise labels that change over time, reinforcing harmful stereotypes and not representing underlying genetic diversity within categories.[7,8] The use of genetic ancestry-based population descriptors also has advantages and disadvantages. For example, while ancestry-based population descriptors, assigned using genetic markers, are rooted in biology, the accuracy of ancestry estimates is limited by available reference sequences. Conclusions based on

[1]Labcorp Genetics, Inc. (formerly Invitae Corp.), San Francisco, CA, USA; [2]Department of Public Health Sciences (Epidemiology Division), University of California Davis School of Medicine, Davis, CA, USA; [3]UCDavis Health Comprehensive Cancer Center, University of California Davis Medical Center, Sacramento, CA, USA; [4]Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA; [5]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
[6]Present address: Global Medical and Clinical Affairs, Illumina, 5200 Illumina Way, San Diego, CA 92122, USA
*Correspondence: khatchell@illumina.com

https://doi.org/10.1016/j.ajhg.2025.01.008.

ancestry-based population descriptors alone should be treated with caution, as they are often confounded by social determinants of health and could mask contributions to health differences from other sources, such as structural racism.[9]

In 2023, the National Academies of Sciences, Engineering, and Medicine (NASEM) issued a report entitled "Using Population Descriptors in Genetics and Genomics Research,"[10] which evaluated existing methodologies and explored the benefits, pitfalls, and challenges of using population descriptors associated with concepts of descent in human genetics and genomics research. In parallel with or in response to the NASEM report, several prominent scientific professional organizations and journals in human genetics and medicine provided their own guidance, describing the appropriate use of conventional population descriptors, as well as considerations for estimating genetic ancestry from existing population-labeled reference datasets.[5,11–13] Together, these efforts strive to reduce inaccurate assumptions about the relationship between genetic ancestry constructs and subject- or clinician-reported race and ethnicity categories. Integrating this guidance into standard practice may reduce inadvertent harm and improve equity for populations that have historically been excluded by the scientific community and continue to be underrepresented in genomics research and reference datasets.[13]

Implementation of this new guidance is likely to vary among the array of research endeavors and data collected across genomic health programs worldwide. Its relevance and readiness for implementation in clinical genomics laboratories is even less clear. Clinical genomics laboratories often serve as a rich source of legacy datasets for downstream research studies that shed light on the clinical utility of genetic testing in individuals with different backgrounds. Unlike academic research protocols that tend to target specific populations while engaging appropriate community members in study design, clinical genomics laboratories typically serve all comers rather than targeted populations and therefore may represent unselected populations. Furthermore, demographic data available to clinical genomics laboratories, including conventional population descriptors, are frequently provided by the ordering clinician when filling out the test requisition form (TRF). Not all clinical genomics laboratories have the capability to generate ancestry-based estimates, and how these estimates would be used for clinical genetic testing and interpretation is yet undefined. Thus, the incorporation of many of the recommendations from the NASEM report may not be relevant or feasible to implement in clinical genomics laboratories. Strategies are needed to utilize conventional population descriptors from legacy data, such as those generated from over 5 million subjects who have received genetic testing through our laboratory, while developing effective protocols to generate ancestry-based population descriptors from past and future subject samples. This perspective piece offers insights and examples from over a decade of experience using both conventional and ancestry-based population descriptors in clinical laboratory genetic testing and associated research.

Collection of conventional population descriptors in clinical genomics laboratories

In current practice, subject data that are collected, requested, and reported to clinical laboratories for genetic testing are often limited by a predetermined list of discrete categories representing population descriptors such as race, ethnicity, and/or ancestry. The specific population descriptors and associated categories or group labels from which clinical providers and/or subjects can choose when ordering a genetic test are typically provided by the laboratory and may or may not include an open-ended write-in option.[14] For example, the current Invitae (now part of Labcorp) TRF includes nine discrete population descriptor categories and an open-ended "other" option. The lack of standards for the format and content of these questions in clinical genetic testing has led to vast differences between laboratories in terms of the data that are collected and reported, and there are no consensus guidelines to help laboratories determine what should be done moving forward.[15]

Over the last decade, our laboratory has performed different types of genetic tests ranging from single-gene tests to multigene panels to whole-genome sequencing for >5 million individuals. Test orders were primarily from the United States (82.3%), and approximately 17.7% were from >120 other countries. In test orders from the United States ($n = 4,497,419$), data for up to ten subject self-identified or clinician-indicated population descriptors plus an open-ended "other" option were provided on the TRF, including six major United States census categories for race and ethnicity ("American Indian or Alaska Native," 'Asian," 'Black or African American," "Hispanic or Latino," "Native Hawaiian or Pacific Islander," and "White"), as well as categories related to ancestry from specific founder populations: "Ashkenazi Jewish," "French Canadian," "Mediterranean," and "Sephardic Jewish" (Figure 1).

Complexities of conventional population descriptors collected in clinical genomics laboratories

Race and ethnicity categories reported on clinical forms may be inconsistent over time and often differ depending on the person providing the information (e.g., clinical provider or office staff versus subject). Within a subset of 486,584 individuals who underwent multiple tests at our laboratory between 2014 and 2024, we observed that 15,845 (3.3%) were assigned different racial and ethnic categories over time (Figure 2). Interestingly, these changes were not limited to additions or deletions of race and ethnicity categories, and some changed entirely from one descriptor to another (e.g., Ashkenazi Jewish only to White only).

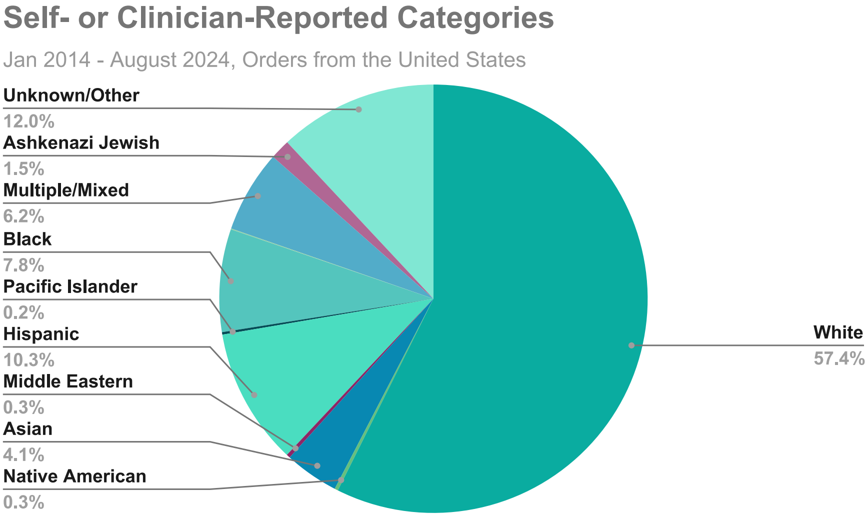

Figure 1. Pie chart showing the relative proportion of self-identified or clinician-indicated categories from 4,497,419 orders submitted from January 2014 to August 2024 within the United States
The majority ($n = 2{,}579{,}943$; 57.4%) of genetic testing orders came from individuals identified as non-Hispanic White (which is consistent with population statistics reported in the 2020 US census[16]). The remaining 42.6% of this sample consisted of 30.7% who identified with at least one category other than non-Hispanic White (6.2% of whom identified with multiple categories). Within the unknown/other group (12.0%), 4.6% left it blank, while 2.1% wrote in a custom option and the remaining selected "unknown," an option only available online. The multiple/mixed group includes individuals who selected two check boxes or wrote in multiple distinct ethnicities in the free-text field. Institutional review board (IRB) approval for these heretofore unpublished results was granted by the WCG IRB (study number 1167406), and all subjects provided informed consent.

Heterogeneity in the population descriptors reported for genetic testing subjects is lost during research or other activities when restricted to discrete categorical data types, with a limited number of groups lumped into large "bins" without providing the chance to maintain multiple categories or an open-ended response option for subjects or providers to write in a custom description. Among orders with a custom response, 29.6% ($n = 41{,}553$) contained free text seen fewer than five times in the laboratory's tested population, emphasizing the vast diversity of self- or clinician-indicated racial and ethnic identities.

To analyze open-ended responses provided in the "other" free-text box, data cleaning was imperative and nontrivial. There were 59 unique open-ended responses with >1,000 occurrences among all recorded orders, which required manual review. Regular expressions[17] were used to identify patterned strings of text while collapsing similar entries with alternate spellings and punctuation (Table S1). While 2.1% of all responses could not be aligned to a structured category, 3.1% of all orders (n = 140,224) contained free-text data, approximately one-third of which were grouped into pre-existing categories. Integrating free-text entries into larger categories allows more granular descriptions of subject demographic groups (and non-categorical responses) to be included in statistical analyses of large datasets; however, collapsing free-text data into larger pre-existing categories should only be attempted alongside careful preservation of the original input data, as more specific or detailed analyses of subgroups may be desirable in the future.

Another caveat to using conventional population descriptors in clinical genomics laboratories is that descriptors may have been reported by clinicians rather than their subjects. In one study of subjects from our lab who self-identified with demographic categories through an online portal (n = 4,618), comparisons between categories selected by subjects and clinician-indicated categories reported on clinical lab TRFs revealed complex trends reflecting differences in the total number of categories reported for subjects, data collection approaches, and the presence or absence of an open-ended response field.[18] Specifically, subjects who self-identified with only one category in their private subject portal had high levels of agreement with the TRF completed by a healthcare provider (n = 3,686; ~80%). In contrast, subjects who identified with 2–3 categories in their portal were often assigned only one category on the clinician-provided form. Importantly, there was an open-ended field provided on the TRF (which was not an option in the patient portal) that provided clinicians an opportunity to add custom descriptions, presumably through a conversation with their subjects. Providing only multiple-choice options (including "other" but without an open-ended response option) led to higher rates of missed identities for some subjects through the portal relative to TRFs that included custom entries. It is possible that subjects may have been reluctant or did not have the opportunity to share their full self-described identities with providers. It would be informative for future studies to offer subjects an open-ended [custom descriptor] field in addition to multiple-choice categories that match those on clinical laboratory forms.[18] These complex issues in the reporting of race and ethnicity highlight that relying solely on self-identified or clinician-reported population descriptors may reduce the power of certain study designs, including less precise matching for case-control studies, greater missingness of data, and reduced statistical power due to non-systematic misclassification bias.

Determination of ancestry-based population descriptors in the clinical genomics laboratory

A number of approaches have been developed to infer individual genetic ancestry from data generated during

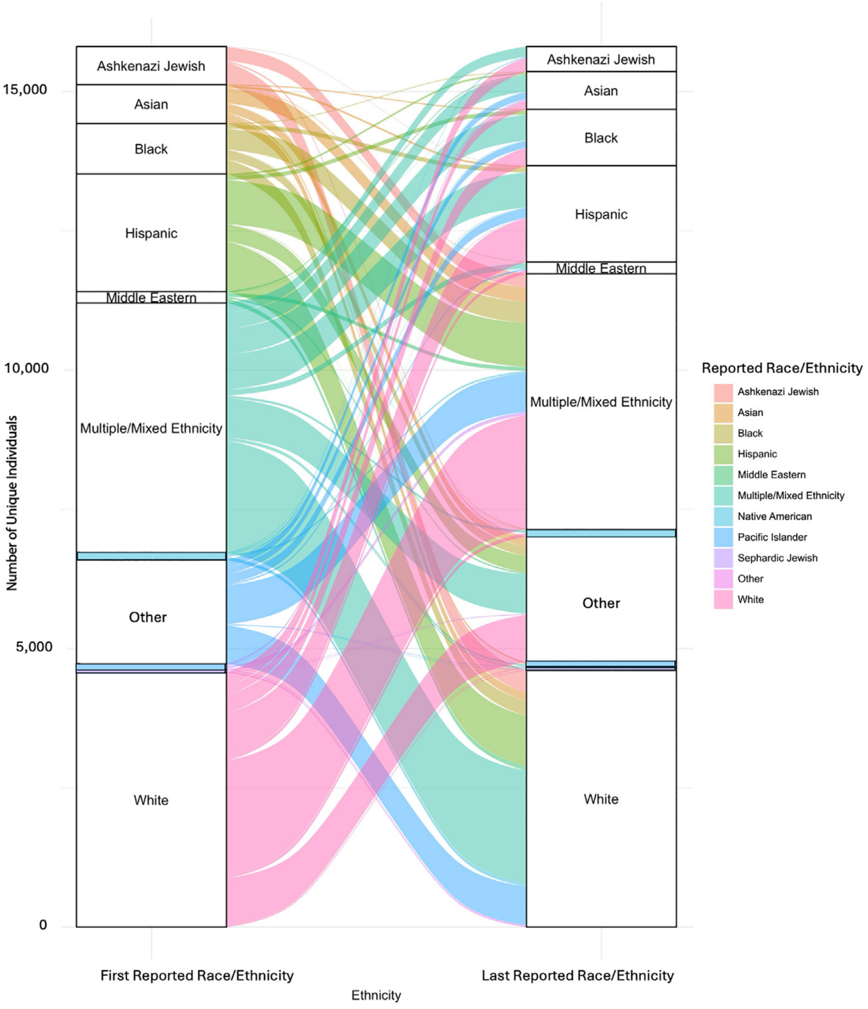

Figure 2. Changes between tests in race and/or ethnicity categories reported for 15,845 individuals (3.3% of individuals with more than 1 order [*n* = 486,584])
All changes in categories with fewer than 5 individuals (per group) are not shown. All orders with unknown race/ethnicity reported were removed. Due to the small size, the Native American, Pacific Islander, and Sephardic Jewish boxes are not labeled but are color coded, directly above or below "other" in that order. IRB approval for these heretofore unpublished results was granted by WCG IRB (study number 1167406), and all subjects provided informed consent.

clinical genetic testing based on genotyping polymorphic alleles or markers with frequencies that vary across geographic regions and thus correlate with familial ancestral origins.[19,20] These inferences may be particularly useful for those whose self-identified ancestral background is unknown or ambiguous. In order to compare self- or clinician-reported conventional population descriptors to inferred genetic ancestry groupings, our laboratory investigated the relationship between these measures in a cohort of 3,685 individuals harboring variants associated with monogenic conditions (unpublished data) using two different methods for estimating genetic ancestry: latent Dirichlet allocation followed by k-means clustering (LDA-KNN [k-nearest neighbor]) and a support vector machine classifier (SVC). Both LDA-KNN and SVC methods construct boundaries based on reference data to estimate per-individual ancestry proportions drawn from a predetermined number of reference populations.

For each type of analysis, reference population data from the 1,000 Genomes Project (phase 3 data release)[21] were used to construct five ancestry groups based on continental-level sampling. The LDA-KNN and SVC methods were in agreement with each other for 95.6% of the ancestry estimates they produced. There were 158 (4.3%) individuals who were classified differently between methods, and 113 (3.1%) individuals who could not be classified by either model were excluded from further analysis, resulting in an analysis cohort of 3,414 individuals. For 2,486 (72.8%) individuals in this cohort with available population descriptors (excluding "unknown"), the categories provided were mapped to five continental groups, defined as "super-populations" in the 1,000 Genomes Project dataset (e.g., Hispanic was mapped to admixed Americans [AMRs]). There was 91.8% overall concordance between assigned continental groups based on estimated genetic ancestry and their most closely mapped self-identified or clinician-reported conventional population descriptors.

Despite the limitations and pitfalls of using broad continental groupings to infer ancestry (and comparing these to socio-cultural identities), these efforts allowed us to estimate genetic ancestry for 27.2% of all individuals who had missing conventional population descriptors (*n* = 928), enabling the inclusion of these individuals in further downstream research. This has the double benefit of providing a larger sample size for our analyses and broadening the representation of our subject population in research, who would otherwise have been excluded.

Use of conventional and ancestry-based population descriptors in enhancing the clinical validity of genetic testing

The primary goal of clinical genomics laboratories is to report results that may inform a subject's diagnosis and/or guide clinical management. Both conventional and ancestry-based population descriptors available within large databases from clinical genomics laboratories have been used by our laboratory to improve the clinical validity of the results that are shared with subjects.

Use of conventional population descriptors to improve variant classification

Clinical genomics laboratories are well positioned to inform improvements in clinical genetics practice by utilizing large-scale internal data obtained from both affected and unaffected individuals referred for testing. Our laboratory has accumulated data from over 5 million individuals from countries around the world, the majority of whom have population descriptors. As an example of the benefits of using large-scale internal data, recent studies using large cohort data found that a *HOXB13* (MIM: 604607) gene variant (c.853del [GenBank: NM_006361.6] [p.Ter285LysextTer?]), which was initially classified by our laboratory as likely benign, is associated with an increased risk of prostate cancer (MIM: 610997) and is enriched in individuals of West African genetic ancestry,[22] as well as those characterized as Black and/or African American.[23,24] This variant has a reported minor-allele frequency of 0.2% in the Genome Aggregation Database (gnomAD) "genetic ancestry group: African/African American,"[25] which is greater than expected for a pathogenic allele and resulted in its initial clinical classification. However, our laboratory eventually reclassified it as an increased risk allele associated with a significant >2-fold increased risk for prostate cancer compared to non-carriers,[22–24] based on an internal association study of this variant and prostate cancer that accounted for population structure using subject- or clinician-reported conventional population descriptors.[26] This research effort thus improved the clinical validity of variant classification for our subject population and may facilitate the use of targeted therapies in subjects with prostate cancer and/or promote risk-reducing strategies in currently unaffected family members who are found to share this variant with the proband.

For research studies examining social determinants of health (e.g., structural racism) and evaluating healthcare disparities (e.g., differences in genetic testing referral rates between subject racial and ethnic groups), conventional population descriptors such as race and ethnicity are likely appropriate to use because these descriptors reflect the same groups subjected to socially constructed notions, policies, and practices that disadvantage some and advantage others.[10] Research insights based on conventional population descriptors can more easily be translated into clinical practice since medical records and clinicians tend to use these categories when collecting and reporting the racial and ethnic backgrounds of subjects. For example, population descriptors based on race and ethnicity are required and relevant for investigating long-standing racial disparities between Black and White subjects in the United States on the uptake of genetic services. A recent study from our laboratory illustrated that fewer at-risk relatives of self- or clinician-reported Black or African American subjects underwent cascade testing following a positive test result in a proband compared to White subjects.[27]

Conventional population descriptors are also useful for clinical studies measuring differences in the rates of uncertain and definitive results from genetic testing across population groups. For example, published data from our laboratory revealed lower rates of variants of uncertain significance (VUSs) in self-identified White subjects (relative to all other demographic categories),[28–32] consistent with results from similar studies.[33–35] More recently, we showed that the rate of VUS reclassification is particularly high among racial and ethnic groups historically underrepresented in genetic research.[36] In another study, we demonstrated that individuals of self-identified Ashkenazi Jewish descent were twice as likely to have unexpected pathogenic or likely pathogenic germline variants compared to individuals without this designation when undergoing cascade testing.[37] In a fourth study, we showed that direct-to-consumer (DTC) genetic testing, which is often restricted to only a few of the commonly recognized *BRCA1* (MIM: 113705) or *BRCA2* (MIM: 600185) pathogenic or likely pathogenic variants, is expected to miss >90% of clinically relevant variants in individuals with no self-identified Ashkenazi Jewish ancestry and ~10% among Ashkenazi Jewish-identified individuals.[38]

These studies illustrate specific scenarios in which conventional population descriptors may provide useful insights for clinical providers. When goals of research include measuring healthcare disparities, conventional population descriptors related to self-identified race and ethnicity are clearly important. Additionally, data that clinicians may use for genetic and medical counseling include ancestry categories that are informative for populations' genomic background, such as descent-associated descriptors that refer to populations with a history of bottlenecks such that founder mutations may influence allele frequencies. There are also important caveats and limitations associated with the types of data gathered in clinical genomics laboratories, which are mentioned in the studies we highlighted, consistent with recent guidance.[11]

In contrast to the above examples, which demonstrate the utility of conventional population descriptors in improving variant classification, the use of such descriptors may not be appropriate in the reproductive carrier screening setting. The use of conventional population descriptors has been standard practice when providing residual carrier risks in reproductive carrier screening for autosomal recessive and X-linked disorders. This type of risk estimate reflects the probability that an individual with a negative genetic test result may still be a carrier of the suspected genetic condition based on known frequencies in population datasets that correspond to an individual's self-reported ancestry. However, in most cases, providing residual risk estimates based on conventional population descriptors may not be clinically appropriate. The accuracy of residual risk estimates relies on disease incidence estimates for the underlying populations, which are notoriously unreliable for populations underrepresented in genomics research. This may contribute to the risk of misinterpreting reported results, especially in people with ancestral origins underrepresented in population databases, unknown family origins, or with diverse ancestral backgrounds and mixed heritage.[39] Additionally, the ancestral haplotype, or local genomic context of alleles within and in the proximity of a given disease gene, may not correspond to genome-wide estimates of ancestry groupings, which can create confusing or misleading information for clinical genetics providers, who may make decisions or interpretations based on assumptions related to "global" ancestry, without this being an appropriate proxy for the likelihood of a particular variant being found.

This is one clear example where patient ancestry for genetic testing had once been considered relevant but was subsequently set aside. The recent update to American College of Medical Genetics (ACMG) guidance on reproductive carrier screening now recommends expanded screening for carriers of hundreds of genetic disorders by next-generation sequencing for all individuals (regardless of background) based on genetic diversity in the general population and limitations of population descriptor data as a proxy for genomic background and motivated by a desire to ensure "equitable opportunity for patients to learn their reproductive risks."[40]

Use of ancestry-based population descriptors in improving variant classification

The allele frequency of a variant observed in a large, healthy population is one type of evidence used by clinical genomics laboratories to assess the likelihood that a variant is benign or pathogenic in the context of Mendelian disease.[41,42] Clinical genomics laboratories typically use databases containing genomic sequences from large numbers of individuals from many ancestral populations to assess a variant's frequency. Though most variants that are relevant in clinical genetic testing for monogenic disorders are present throughout the world's population and thus not influenced by ancestry, some clinically relevant variants are not uniformly distributed and have population frequencies that differ by genomic ancestral background.[43] Paying attention to genetic ancestry may, therefore, inform the discovery and interpretation of clinically important genomic regions and variants.

By using data in the Exome Aggregation Consortium (ExAC) database,[44] researchers from our laboratory established that, other than the few exceptional pathogenic variants with higher-than-expected allele frequencies that were already well characterized in the literature, the majority of pathogenic variants were extremely rare globally (having an allele frequency of $<0.01\%$).[26] The underlying reasons for variability in allele frequencies across populations with more specific bio-geographic characteristics are seldom considered during variant classification, which is an important and consequential oversight. A more nuanced approach is required to accurately distinguish, for example, between a benign population-specific polymorphism, a disease risk allele with variable penetrance and expressivity due to gene-environment or gene-gene interactions, and a founder mutation responsible for the elevated prevalence of a genetic disease in populations with shared ancestry.

As an example, c.3628−41_3628−17del (GenBank: NM_000256.3) ("*MYBPC3*Δ25bp") describes a 25 bp deletion in the *MYBPC3* (MIM: 600958) gene that was originally thought to be associated with hypertrophic cardiomyopathy (MIM: 115197) and was proposed as an explanation for the disease prevalence in South Asian Indians since it was observed at a higher frequency in affected individuals with this ancestral background.[45] After further inquiry into this association, however, this variant was found to be a marker for a haplotype containing both the *MYBPC3*Δ25bp allele and another variant in an intronic region of *MYBPC3* (c.1224−52G>A [GenBank: NM_000256.3]),[46,47] which has since been validated as one of the most common pathogenic variants among hypertrophic cardiomyopathy subjects.[48,49] Although c.1224−52G>A is rarer than *MYBPC3*Δ25bp, it is more common in individuals of South Asian ancestry (0.015%) compared to all individuals in ExAC's successor, gnomAD[25] (0.005%). Identification of this population-enriched genetic modifier allows clinical genomics laboratories to provide more accurate variant classifications. This example further emphasizes the importance of paying attention to the diversity of study populations and including more diverse study participants in large-scale public genomic databases.

Despite these examples of how ancestry-based population descriptors may improve the detection of clinically valid genetic test results, there are limitations to using genetic ancestry information within clinical genomics laboratories. Many clinical laboratories do not, nor are they expected to, report genetic ancestry when providing test results from routine screening or diagnostic genetic testing, even if the data for imputing genetic ancestry are available (e.g., with high-resolution SNP microarray, exome sequencing, or whole-genome sequencing). Moreover, many molecular methods used for diagnostic genetic testing or screening do not generate the data needed to infer genetic ancestry, which require comparisons of genomic variant calls to reference datasets labeled with population descriptors, often indicating the sampling location or ancestral origins of data subjects. Finally, not all conventional population descriptors with racial and ethnic categories are informative for ancestry-specific allele frequencies and indeed may obfuscate trends in genetic diversity by collapsing highly diverse populations into broad groupings, such as entire continents.

Future directions

Professional practice guidelines are needed to directly address how, in clinical genomics laboratories, population descriptors such as race, ethnicity, and ancestry are collected, stored, used, and reported. Guidance is also needed for how and when genetic ancestry inferences should be generated or obtained, incorporated into testing protocols and interpretation of results, and/or described to subjects and their healthcare providers. Guidelines for publications issued in parallel with or in response to the NASEM report state that the use of conventional population descriptors may be acceptable for certain types of research, such as research evaluating social determinants of health,[6] as long as the researchers are transparent about how the data were obtained, the specific procedures and a rationale for how the analyses were conducted, and how the use of these population descriptors may impact the interpretation of results.[11] In that regard, it is useful to recognize that when using population descriptors in certain types of analyses (e.g., investigation of variant classification discrepancies in a diverse population[31]), excluding individuals with "unknown," "other," or multiple descriptors may unintentionally perpetuate the historical underrepresentation of individuals with diverse ancestral backgrounds and non-European populations. This would be similar to excluding individuals with imputed admixed ancestry from genome-wide association study (GWAS), possibly leading to not only false positive results but also decreased generalizability and loss of statistical power.[50]

As the United States population becomes more diverse and integrated across ancestral groups, evidenced by a 276% increase in those who selected "two or more races" on the census between 2010 and 2020, categorical data from population descriptors will likely have decreasing utility, precision, and accuracy to offer researchers and clinicians alike.[51] The same limitation applies to individuals with identities that are not typically represented among multiple-choice categories provided in demographic questions. Data collection efforts may be improved by presenting more inclusive and specific categories that reflect genetic ancestry, allowing for multiple-choice selections and avoiding a single- or best-choice requirement, as well as providing an open-ended option for respondents to add custom responses without having to designate oneself as "other."[52]

Representing and including more people with diverse ancestral backgrounds to better characterize global genomic variation has many benefits, such as improving variant classification and reducing uncertain results for groups with historically high rates of VUSs.[28,30] Multiple studies have identified significant differences in the rate of VUSs and the frequency of pathogenic variants among different populations, whether stratified by population descriptors or inferred genetic ancestry groupings; these differences are likely caused by the unequal representation of populations within genomic reference databases.[29,31–33,43,53] Though not yet routine in clinical genomics laboratories, ancestry-based population descriptors could be used to inform variant classification and the return of results, indicating whether a subject's ancestral background is sufficiently represented in reference data and population allele frequency databases. This approach may improve VUS resolution in underrepresented groups, which is an achievable goal given the ongoing expansion of genomic reference databases through large-scale sequencing in the general population.[52,54]

Over the last two decades, a single reference genome (first GRCh37 and now GRCh38) has been heavily used to inform genetic testing and variant detection. These genome builds were derived largely (~70%–72%) from a single anonymous donor whose genomic profile suggested *post facto* they were male and had recent ancestry from both African and European populations. An additional ~23% of the reference sequence was derived from ~10 other individuals, and the remaining 7% was contributed from >50 DNA donors solicited through public outreach efforts in the northeast United States (https://www.ncbi.nlm.nih.gov/grc/help/faq/).[55,56] Based on these contributions of DNA from individuals in the United States, the current reference genomes used by many clinical labs (build GRCh37 or GRCh38) contain known structural errors and gaps that amount to ~8% of the linear genome reference, leading to biases in variant discovery, detection, and classification.[57,58] These errors have been mitigated over time by the addition of "patches" and other updates to the reference genome.

Recent efforts to generate a higher-quality reference genome that fills in the known gaps and errors (e.g., whole-genome, long-read sequencing efforts of the Human Pangenome Reference and Telomere-to-Telomere consortia) have eliminated hundreds of thousands of rare erroneous variants and excess fixed variants (in non-African genomes), which has led to improvements in variant calling using a pangenome reference, T2T-CHM13.[59] Having access to more complete, accurate, and diverse genome reference data improves variant calling accuracy in everyone and boosts confidence in the reported results for individuals from underrepresented populations. Eventually, the adoption of a pangenome reference for variant calling and classification is expected to improve interpretation for subjects from any and all combinations of genetic ancestral backgrounds.[60,61]

Concluding remarks

As clinical genetic testing increases for individuals with diverse ancestral backgrounds, laboratories are becoming increasingly aware of the issues raised by the application and misuse of various population descriptors, both in developing diagnostic or screening tests and reporting results from those tests.

Clinical laboratories recognize that a broadly diverse representation of genomes in public databases and the availability of a pangenome reference are expected to enable improvements in variant interpretation accuracy and reporting that is specifically tailored to the individual undergoing genetic testing. A pangenome reference will also help reduce uncertainty in the identification and clinical classification of many variants, as sequence reads from genomes that were not well represented in the linear reference will now be able to map to the pangenome reference.

Although laboratories today do not typically utilize or report genetic ancestry information during routine genetic testing, this could change in the future as exome or genome sequencing becomes standard practice. Furthermore, the development and adoption of new methods in clinical genomics laboratories (e.g., clinical variant modeling using machine learning, pangenome references, and polygenic risk scores) should be accompanied by empirical studies to test their validity and utility across different populations. This should also include validation of approaches that use population descriptors and methodologies for constructing genetic ancestry estimates to help inform clinical providers and researchers about the most relevant and accurate types of information and classifiers for specific investigative and clinical purposes. Rather than exclude population descriptors provided by subjects or their healthcare providers outright, it is important to keep in mind the utility of this information for tracking diversity, equity, and inclusion and an improved understanding of the etiology of disease, healthcare delivery, and health outcomes and to consider what would be lost if these data were no longer available to inform different types of inquiries into the access and utilization of genetic services in healthcare.

Many research studies have been reported, and many more will be conducted in the future, using expansive accumulated data from clinical genomics laboratories. Research performed on datasets from these laboratories and observations drawn from those datasets are powerful ways to inform clinical decision-making because of the approximation of such cohorts to real-world populations. Feero et al.[11] encourage researchers to document how population descriptors are collected, how they might influence interpretation of results and the conclusions drawn, and their potential limitations. Since genetic ancestry does not capture an individual's sociopolitical, economic, geographic, or cultural environment, all of which may play a role in their experience of clinical genetic testing and subsequent healthcare utilization, it may be important to continue collecting population descriptors related to social and cultural identities, such as race and ethnicity. This will ensure that healthcare utilization metrics, as collected by health disparities researchers, can continue to track differences in access to genetic services as well as the impact of genetic testing on patient populations. Improved standards for the collection and use of population descriptors may also directly enhance inclusivity in genomics research studies and provide greater insights into the generalizability of results while providing more precise and relevant information about participants in ways that are more respectful of individual identities.

Data and code availability

The previously unpublished data supporting the current study have not been deposited in a public repository but are available from the corresponding author upon request.

Acknowledgments

This work was funded by Invitae Corp. (now Labcorp Genetics, Inc.).

Author contributions

Project conception, S. Aradhya, K.E.H., and F.M.F.; data analysis, K.E.H., S.R.P., E.M.R., T.J.W., and A.B.P.; drafting of original manuscript, K.E.H., S.R.P., E.M.R., T.J.W., R.E.E., F.M.F., S. Aguilar, and S. Aradyha; critical review and revision of manuscript, E.D.E., R.L.N., and A.B.P.

Declaration of interests

K.E.H., S.R.P., E.M.R., T.J.W., R.E.E., F.M.F., S. Aguilar, and E.D.E are all former employees of Invitae Corp. and current employees of Labcorp Genetics (on August 5th, 2024, Labcorp Genetics, Inc., acquired select assets of Invitae Corp.). R.L.N. is a former employee of Invitae Corp. S. Aradhya is a former employee of Invitae Corp. and a current employee of Illumina. E.D.E. is an advisor and stockholder of Taproot Health, Exir Bio, and ROMTech.

Web resources

GenBank, https://www.ncbi.nlm.nih.gov/genbank/

Genome Reference Consortium Frequently Asked Questions, https://www.ncbi.nlm.nih.gov/grc/help/faq/

Measuring Race And Ethnicity Across The Decades: 1790–2010, https://www.census.gov/newsroom/blogs/random-samplings/2015/11/measuring-race-and-ethnicity-across-the-decades-1790-2010.html

OMIM, https://www.ncbi.omim.org

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2025.01.008.

References

1. Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., Richter, D.J., Lander, E.S., and Altshuler, D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. *32*, 135–142. https://doi.org/10.1038/ng947.
2. Shriver, M.D., Mei, R., Parra, E.J., Sonpar, V., Halder, I., Tishkoff, S.A., Schurr, T.G., Zhadanov, S.I., Osipova, L.P., Brutsaert, T.D., et al. (2005). Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum. Genomics *2*, 81–89. https://doi.org/10.1186/1479-7364-2-2-81.
3. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation.

Science *319*, 1100–1104. https://doi.org/10.1126/science.1153717.

4. Committee on the Use of Race and Ethnicity in Biomedical Research, Board on Health Sciences Policy, Board on Population Health and Public Health Practice, Board on Health Care Services, Health and Medicine Division, and National Academies of Sciences, Engineering, and Medicine (2024). Rethinking Race and Ethnicity in Biomedical Research (Preprint at National Academies Press). https://doi.org/10.17226/27913.
5. 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. https://doi.org/10.1038/s41436-021-01109-w.
6. Borrell, L.N., Elhawary, J.R., Fuentes-Afflick, E., Witonsky, J., Bhakta, N., Wu, A.H.B., Bibbins-Domingo, K., Rodríguez-Santana, J.R., Lenoir, M.A., Gavin, J.R., 3rd., et al. (2021). Race and genetic ancestry in medicine - A time for reckoning with racism. N. Engl. J. Med. *384*, 474–480. https://doi.org/10.1056/NEJMms2029562.
7. Gombault, C., Grenet, G., Segurel, L., Duret, L., Gueyffier, F., Cathébras, P., Pontier, D., Mainbourg, S., Sanchez-Mazas, A., and Lega, J.-C. (2023). Population designations in biomedical research: Limitations and perspectives. HLA *101*, 3–15. https://doi.org/10.1111/tan.14852.
8. Krainc, T., and Fuentes, A. (2022). Genetic ancestry in precision medicine is reshaping the race debate. Proc. Natl. Acad. Sci. USA *119*, e2203033119. https://doi.org/10.1073/pnas.2203033119.
9. Cerdeña, J.P., Grubbs, V., and Non, A.L. (2022). Genomic supremacy: the harm of conflating genetic ancestry and race. Hum. Genomics *16*, 18. https://doi.org/10.1186/s40246-022-00391-2.
10. National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Health and Medicine Division; Committee on Population; Board on Health Sciences Policy; Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research (2023). Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field (National Academies Press (US)). https://doi.org/10.17226/26902.
11. Feero, W.G., Steiner, R.D., Slavotinek, A., Faial, T., Bamshad, M.J., Austin, J., Korf, B.R., Flanagin, A., and Bibbins-Domingo, K. (2024). Guidance on use of race, ethnicity, and geographic origin as proxies for genetic ancestry groups in biomedical publications. Am. J. Hum. Genet. *111*, 621–623. https://doi.org/10.1016/j.ajhg.2024.03.003.
12. Flanagin, A., Frey, T., Christiansen, S.L.; and AMA Manual of Style Committee (2021). Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. JAMA *326*, 621–627. https://doi.org/10.1001/jama.2021.13304.
13. Jackson, C.S., Turner, D., June, M., and Miller, M.V. (2023). Facing Our History-Building an Equitable Future. Am. J. Hum. Genet. *110*, 377–395. https://doi.org/10.1016/j.ajhg.2023.02.005.
14. Popejoy, A.B., Crooks, K.R., Fullerton, S.M., Hindorff, L.A., Hooker, G.W., Koenig, B.A., Pino, N., Ramos, E.M., Ritter, D.I., Wand, H., et al. (2020). Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. Am. J. Hum. Genet. *107*, 72–82. https://doi.org/10.1016/j.ajhg.2020.05.005.
15. Popejoy, A.B., Ritter, D.I., Crooks, K., Currey, E., Fullerton, S.M., Hindorff, L.A., Koenig, B., Ramos, E.M., Sorokin, E.P., Wand, H., et al. (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum. Mutat. *39*, 1713–1720. https://doi.org/10.1002/humu.23644.
16. Jensen, E., Jones, N., Rabe, M., Pratt, B., Medina, L., Orozco, K., and Spell, L. (2021). The Chance That Two People Chosen at Random Are of Different Race or Ethnicity Groups Has Increased Since 2010. 2020 U.S. Population More Racially and Ethnically Diverse than Measured in 2010. https://www.census.gov/library/stories/2021/08/2020-united-states-population-more-racially-ethnically-diverse-than-2010.html.
17. Friedl, J.E.F. (2006). Mastering Regular Expressions, 3rd ed. (O'Reilly Media).
18. Popejoy, A., Morales, A., and McKnight, D. (2023). P537: Reliability of clinician entries for patient self-identified race, ethnicity, and ancestry in clinical genetic testing. Genetics in Medicine Open *1*, 100584. https://doi.org/10.1016/j.gimo.2023.100584.
19. Liu, C.-C., Shringarpure, S., Lange, K., and Novembre, J. (2020). Exploring population structure with admixture models and principal component analysis. Methods Mol. Biol. *2090*, 67–86. https://doi.org/10.1007/978-1-0716-0199-0_4.
20. Jordan, I.K., Sharma, S., Nagar, S.D., Valderrama-Aguirre, A., and Mariño-Ramírez, L. (2022). Genetic ancestry inference for pharmacogenomics. Methods Mol. Biol. *2547*, 595–609. https://doi.org/10.1007/978-1-0716-2573-6_21.
21. Auton, A., Abecasis, G.R., Brooks, L.D., Korbel, J.O., Kang, H.M., Garrison, E.P., Abecasis, G.R., McCarthy, S., McVean, G.A., et al.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
22. Darst, B.F., Hughley, R., Pfennig, A., Hazra, U., Fan, C., Wan, P., Sheng, X., Xia, L., Andrews, C., Chen, F., et al. (2022). A Rare Germline HOXB13 Variant Contributes to Risk of Prostate Cancer in Men of African Ancestry. Eur. Urol. *81*, 458–462. https://doi.org/10.1016/j.eururo.2021.12.023.
23. Kanayama, M., Chen, Y., Rabizadeh, D., Vera, L., Lu, C., Nielsen, S.M., Russell, E.M., Esplin, E.D., Wang, H., Isaacs, W.B., et al. (2024). Clinical and Functional Analyses of an African-ancestry Gain-of-function HOXB13 Variant Implicated in Aggressive Prostate Cancer. Eur. Urol. Oncol. *7*, 751–759. https://doi.org/10.1016/j.euo.2023.09.012.
24. Na, R., Wei, J., Sample, C.J., Gielzak, M., Choi, S., Cooney, K.A., Rabizadeh, D., Walsh, P.C., Zheng, L.S., Xu, J., and Isaacs, W.B. (2022). The HOXB13 variant X285K is associated with clinical significance and early age at diagnosis in African American prostate cancer patients. Br. J. Cancer *126*, 791–796. https://doi.org/10.1038/s41416-021-01622-4.
25. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. Nature *625*, 92–100. https://doi.org/10.1038/s41586-023-06045-0.
26. Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S.E., and Topper, S.E. (2017). Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med. *9*, 13. https://doi.org/10.1186/s13073-017-0403-7.
27. Kassem, N.M., Althouse, S.K., Monahan, P.O., Hayes, L., Nielsen, S.M., Heald, B., Esplin, E.D., Hatchell, K.E., and Ballinger,

T.J. (2023). Racial disparities in cascade testing for cancer predisposition genes. Prev. Med. *172*, 107539. https://doi.org/10.1016/j.ypmed.2023.107539.

28. Chen, E., Facio, F.M., Aradhya, K.W., Rojahn, S., Hatchell, K.E., Aguilar, S., Ouyang, K., Saitta, S., Hanson-Kwan, A.K., Capurro, N.N., et al. (2023). Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing. JAMA Netw. Open *6*, e2339571. https://doi.org/10.1001/jamanetworkopen.2023.39571.
29. Shore, N., Gazi, M., Pieczonka, C., Heron, S., Modh, R., Cahn, D., Belkoff, L.H., Berger, A., Mazzarella, B., Veys, J., et al. (2023). Efficacy of National Comprehensive Cancer Network Guidelines in Identifying Pathogenic Germline Variants Among Unselected Patients with Prostate Cancer: The PROCLAIM Trial. Eur. Urol. Oncol. *6*, 477–483. https://doi.org/10.1016/j.euo.2023.07.008.
30. Coughlin, S.E., Heald, B., Clark, D.F., Nielsen, S.M., Hatchell, K.E., Esplin, E.D., and Katona, B.W. (2022). Multigene Panel Testing Yields High Rates of Clinically Actionable Variants Among Patients With Colorectal Cancer. JCO Precis. Oncol. *6*, e2200517. https://doi.org/10.1200/PO.22.00517.
31. Appelbaum, P.S., Burke, W., Parens, E., Zeevi, D.A., Arbour, L., Garrison, N.A., Bonham, V.L., and Chung, W.K. (2022). Is there a way to reduce the inequity in variant interpretation on the basis of ancestry? Am. J. Hum. Genet. *109*, 981–988. https://doi.org/10.1016/j.ajhg.2022.04.012.
32. Wyatt Castillo, R.B., Nielsen, S.M., Chen, E., Heald, B., Ellsworth, R.E., Esplin, E.D., and Tomlinson, G.E. (2024). Disparate rates of germline variants in cancer predisposition genes in African American/Black compared with non-Hispanic White individuals between 2015 and 2022. JCO Precis. Oncol. *8*, e2300715. https://doi.org/10.1200/PO.23.00715.
33. Sorscher, S., LoPiccolo, J., Heald, B., Chen, E., Bristow, S.L., Michalski, S.T., Nielsen, S.M., Lacoste, A., Keyder, E., Lee, H., et al. (2023). Rate of Pathogenic Germline Variants in Patients With Lung Cancer. JCO Precis. Oncol. *7*, e2300190. https://doi.org/10.1200/PO.23.00190.
34. Jones, J.C., Golafshar, M.A., Coston, T.W., Rao, R., Wysokinska, E., Johnson, E., Esplin, E.D., Nussbaum, R.L., Heald, B., Klint, M., et al. (2023). Universal Genetic Testing vs. Guideline-Directed Testing for Hereditary Cancer Syndromes Among Traditionally Underrepresented Patients in a Community Oncology Program. Cureus *15*, e37428. https://doi.org/10.7759/cureus.37428.
35. Pan, E., Shaya, J., Madlensky, L., Randall, J.M., Javier-Desloges, J., Millard, F.E., Rose, B., Parsons, J.K., Nielsen, S.M., Hatchell, K.E., et al. (2022). Germline alterations among Hispanic men with prostate cancer. Prostate Cancer Prostatic Dis. *25*, 561–567. https://doi.org/10.1038/s41391-022-00517-6.
36. Kobayashi, Y., Chen, E., Facio, F.M., Metz, H., Poll, S.R., Swartzlander, D., Johnson, B., and Aradhya, S. (2024). Clinical variant reclassification in hereditary disease genetic testing. JAMA Netw. Open *7*, e2444526. https://doi.org/10.1001/jamanetworkopen.2024.44526.
37. Heald, B., Pirzadeh-Miller, S., Ellsworth, R.E., Nielsen, S.M., Russell, E.M., Beitsch, P., Esplin, E.D., Nussbaum, R.L., Pineda-Alvarez, D.E., Kurian, A.W., and Hampel, H. (2024). Cascade testing for hereditary cancer: comprehensive multigene panels identify unexpected actionable findings in relatives. J. Natl. Cancer Inst. *116*, 334–337. https://doi.org/10.1093/jnci/djad203.
38. Desai, N.V., Barrows, E.D., Nielsen, S.M., Hatchell, K.E., Anderson, M.J., Haverfield, E.V., Herrera, B., Esplin, E.D., Lucassen, A., Tung, N.M., and Isaacs, C. (2023). Retrospective Cohort Study on the Limitations of Direct-to-Consumer Genetic Screening in Hereditary Breast and Ovarian Cancer. JCO Precis. Oncol. *7*, e2200695. https://doi.org/10.1200/PO.22.00695.
39. Nussbaum, R.L., Slotnick, R.N., and Risch, N.J. (2021). Challenges in providing residual risks in carrier testing. Prenat. Diagn. *41*, 1049–1056. https://doi.org/10.1002/pd.5975.
40. Gregg, A.R., Aarabi, M., Klugman, S., Leach, N.T., Bashford, M.T., Goldwaser, T., Chen, E., Sparks, T.N., Reddi, H.V., Rajkovic, A., et al. (2021). Screening for autosomal recessive and X-linked conditions during pregnancy and preconception: a practice resource of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *23*, 1793–1806. https://doi.org/10.1038/s41436-021-01203-z.
41. 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. https://doi.org/10.1038/gim.2015.30.
42. Ghosh, R., Harrison, S.M., Rehm, H.L., Plon, S.E., Biesecker, L.G.; and ClinGen Sequence Variant Interpretation Working Group (2018). Updated recommendation for the benign stand-alone ACMG/AMP criterion. Hum. Mutat. *39*, 1525–1530. https://doi.org/10.1002/humu.23642.
43. Venner, E., Patterson, K., Kalra, D., Wheeler, M.M., Chen, Y.-J., Kalla, S.E., Yuan, B., Karnes, J.H., Walker, K., Smith, J.D., et al. (2024). The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. Commun. Biol. *7*, 174. https://doi.org/10.1038/s42003-023-05708-y.
44. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291. https://doi.org/10.1038/nature19057.
45. Chowdry, A.B., Mandegar, M.A., Benton, G.M., Naughton, B.T., and Conklin, B.R. (2003). Population sampling and in vitro modeling of a 25bp deletion in MYBPC3 associated with hypertrophic cardiomyopathy. https://blog-api.23andme.com/wp-content/uploads/2012/11/HCM-ASHG-TTAM.pdf.
46. Viswanathan, S.K., Puckelwartz, M.J., Mehta, A., Ramachandra, C.J.A., Jagadeesan, A., Fritsche-Danielson, R., Bhat, R.V., Wong, P., Kandoi, S., Schwanekamp, J.A., et al. (2018). Association of Cardiomyopathy With MYBPC3 D389V and MYBPC3Δ25bpIntronic Deletion in South Asian Descendants. JAMA Cardiol. *3*, 481–488. https://doi.org/10.1001/jamacardio.2018.0618.
47. Schwäbe, F.V., Peter, E.K., Taft, M.H., and Manstein, D.J. (2021). Assessment of the Contribution of a Thermodynamic and Mechanical Destabilization of Myosin-Binding Protein C Domain C2 to the Pathomechanism of Hypertrophic Cardiomyopathy-Causing Double Mutation MYBPC3Δ25bp/D389V. Int. J. Mol. Sci. *22*, 11949. https://doi.org/10.3390/ijms222111949.
48. Chumakova, O.S., and Baulina, N.M. (2023). Advanced searching for hypertrophic cardiomyopathy heritability in

real practice tomorrow. Front. Cardiovasc. Med. *10*, 1236539. https://doi.org/10.3389/fcvm.2023.1236539.

49. Harper, A.R., Bowman, M., Hayesmoore, J.B.G., Sage, H., Salatino, S., Blair, E., Campbell, C., Currie, B., Goel, A., McGuire, K., et al. (2020). Reevaluation of the South Asian *MYBPC3* $^{\Delta 25bp}$ intronic deletion in hypertrophic cardiomyopathy. Circ. Genom. Precis. Med. *13*, e002783. https://doi.org/10.1161/circgen.119.002783.
50. Tan, T., and Atkinson, E.G. (2023). Strategies for the Genomic Analysis of Admixed Populations. Annu. Rev. Biomed. Data Sci. *6*, 105–127. https://doi.org/10.1146/annurev-biodatasci-020722-014310.
51. Jones, N., Marks, R., Ramirez, R., and Ríos-Vargas, M. (2021). 2020 Census Illuminates Racial and Ethnic Composition of the Country, *12* (United States Census Bureau).
52. All of Us Research Program Genomics Investigators (2024). Genomic data in the All of Us Research Program. Nature *627*, 340–346. https://doi.org/10.1038/s41586-023-06957-x.
53. Abou Alaiwi, S., Nassar, A.H., Adib, E., Groha, S.M., Akl, E.W., McGregor, B.A., Esplin, E.D., Yang, S., Hatchell, K., Fusaro, V., et al. (2021). Trans-ethnic variation in germline variants of patients with renal cell carcinoma. Cell Rep. *34*, 108926. https://doi.org/10.1016/j.celrep.2021.108926.
54. Turnbull, C., Scott, R.H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F.B., Halai, D., Baple, E., Craig, C., Hamblin, A., et al. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ *361*, k1687. https://doi.org/10.1136/bmj.k1687.
55. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722. https://doi.org/10.1126/science.1188021.
56. Miga, K.H., and Wang, T. (2021). The need for a human pangenome reference sequence. Annu. Rev. Genomics Hum. Genet. *22*, 81–102. https://doi.org/10.1146/annurev-genom-120120-081921.
57. Wagner, J., Olson, N.D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.-C., Gupta, R., Wenger, A.M., Rowell, W.J., et al. (2022). Curated variation benchmarks for challenging medically relevant autosomal genes. Nat. Biotechnol. *40*, 672–680. https://doi.org/10.1038/s41587-021-01158-1.
58. Miller, C.A., Walker, J.R., Jensen, T.L., Hooper, W.F., Fulton, R.S., Painter, J.S., Sekeres, M.A., Ley, T.J., Spencer, D.H., Goll, J.B., and Walter, M.J. (2022). Failure to Detect Mutations in U2AF1 due to Changes in the GRCh38 Reference Sequence. J. Mol. Diagn. *24*, 219–223. https://doi.org/10.1016/j.jmoldx.2021.10.013.
59. Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., et al. (2022). A complete reference genome improves analysis of human genetic variation. Science *376*, eabl3533. https://doi.org/10.1126/science.abl3533.
60. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446. https://doi.org/10.1038/s41586-022-04601-8.
61. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324. https://doi.org/10.1038/s41586-023-05896-x.