Received: 1 May 2018 | Revised: 17 August 2018 | Accepted: 30 August 2018

DOI: 10.1002/humu.23644

SPECIAL ARTICLE

WILEY HGVS HUMAN GENOME VARIATION SOCIETY

The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics

Alice B. Popejoy[1] | Deborah I. Ritter[2] | Kristy Crooks[3,4] | Erin Currey[5] | Stephanie M. Fullerton[6] | Lucia A. Hindorff[5] | Barbara Koenig[7] | Erin M. Ramos[5] | Elena P. Sorokin[1] | Hannah Wand[1] | Mathew W. Wright[1] | James Zou[1] | Christopher R. Gignoux[4] | Vence L. Bonham[8] | Sharon E. Plon[2] | Carlos D. Bustamante[1] | on behalf of the Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG)

[1]Department of Biomedical Data Science, Stanford University, Standford, California

[2]Department of Pediatrics, Texas Children's Hospital, Baylor College of Medicine, Houston, Texas

[3]Department of Pathology, University of Colorado, Anschutz Medical Campus, Aurora, Colorado

[4]Department of Medicine, Division of Bioinformatics and Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado

[5]Division of Genomics and Society, National Human Genome Research Institute (NHGRI), Bethesda, Maryland

[6]Department of Bioethics & Humanities, University of Washington, Seattle, Washington

[7]Department of Anthropology, History, and Social Medicine, University of California, San Francisco

[8]Social and Behavioral Research Branch, National Human Genome Research Institute (NHGRI), Bethesda, Maryland

Correspondence
Alice B. Popejoy, Department of Biomedical Data Science, Stanford University, 1265 Welch Road, Stanford, CA 94305.
Email: apopejoy@stanford.edu

Funding information
National Human Genome Research Institute, Grant/Award Numbers: U41HG009649, U41HG009650

For the ClinGen/ClinVar Special Issue

Abstract

The Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group highlights the need to develop guidance on race, ethnicity, and ancestry (REA) data collection and use in clinical genomics. We present quantitative and qualitative evidence to characterize: (1) acquisition of REA data via clinical laboratory requisition forms, and (2) information disparity across populations in the Genome Aggregation Database (gnomAD) at clinically relevant sites ascertained from annotations in ClinVar. Our requisition form analysis showed substantial heterogeneity in clinical laboratory ascertainment of REA, as well as marked incongruity among terms used to define REA categories. There was also striking disparity across REA populations in the amount of information available about clinically relevant variants in gnomAD. European ancestral populations constituted the majority of observations (55.8%), allele counts (59.7%), and private alleles (56.1%) in gnomAD at 550 loci with "pathogenic" and "likely pathogenic" expert-reviewed variants in ClinVar. Our findings highlight the importance of implementing and supporting programs to increase diversity in genome sequencing and clinical genomics, as well as measuring uncertainty around population-level datasets that are used in variant interpretation. Finally, we suggest the need for a standardized REA data collection framework to be developed through partnerships and collaborations and adopted across clinical genomics.

KEYWORDS

ancestry, clinical genomics, diversity, ethnicity, populations, race

1 | BACKGROUND

Global efforts, such as the National Institutes of Health (NIH)–funded Clinical Genome Resource (ClinGen, https://www.clinicalgenome.org), seek to greatly increase curation activities and develop resources for clinicians and scientists to study the relationships among diseases, genes, and variants. As such, it is critical to understand how race, ethnicity, and genetic ancestry (REA) should be ascertained, communicated, and applied in clinical genomics. Clinical sequencing is now being routinely used for diagnosis, prognosis, and treatment. This has spilled

over into the direct-to-consumer (DTC) market, where ancestry and disease products are expanding in popularity among the general public in the United States (Ramos & Weissman, 2018; Roberts et al., 2017). As a consequence of tens of millions of people already having access to their genetic data, physicians are faced with diverse populations of patients asking about the clinical significance of their genetic information, and this trend will continue to grow in the future. The collection of REA data from patients is often implemented in self-reported contexts (Bonham et al., 2017), for which there are no established best practices, and challenges include complexities of ancestral admixture and international differences in the use of racial, ethnic, and ancestral categories (Morales et al., 2018; Smart, Bolnick, & Tutton, 2017; Travassos & Williams, 2004; Yu, Taylor, Edwards, & Fullerton, 2012).

At present, the "precision" of precision medicine is based on a foundation of evidence mostly from European ancestry populations (Landry, Ali, Williams, Rehm, & Bonham, 2018; Morales et al., 2018, Popejoy & Fullerton, 2016). Patients of non-European ancestry are more likely to receive ambiguous genetic test results (variants of unknown or uncertain significance) (Caswell-Jin et al., 2018), false positive diagnoses (Manrai et al., 2016), and false negative diagnoses in the absence of a robust genomic evidence-base (personal correspondence with E.P. Sorokin, C.R. Gignoux and E. Kenny on behalf of the Population Architecture Using Genomics and Epidemiology, or PAGE, Consortium) and variant reclassification over time (Hiatt et al., 2018). Thus, investigators and professionals in the clinical and scientific research communities have a responsibility to increase diversity in clinical and research sequencing, as well as to understand and harmonize the presentation and utilization of REA in clinical genomics. In this article, we address effects on clinical genomics from the lack of diversity in databases and a lack of clarity in terms used to describe people and populations.

The ClinGen Ancestry and Diversity Working Group (ADWG) formed in November 2017 to improve our understanding of the role of genomic diversity across populations in a clinical genomics setting, with a focus on the United States. Our work includes contributing to the knowledgebase around scientifically and culturally complex issues of self-reported race, ethnicity, geographic and genomic ancestry, particularly in the context of interpreting variants identified via clinical sequencing. The ADWG is multidisciplinary, composed of investigators and clinicians from genomic medicine, bioethics, health disparities, public health, genetic epidemiology, and statistical and population genetics. The diverse expertise of our members, our collaboration with other ClinGen working groups, and the strong commitment of the National Human Genome Research Institute (NHGRI) to promote diversity and research on ancestry (Green, Guyer, & Institute, 2011; Hindorff et al., 2018; Manolio et al., 2017) results in the ADWG being well-poised to tackle such complex issues. We also recognize that in order for these efforts to be successful, collaborations with researchers, communities, and groups outside of ClinGen are crucial. Our current projects involve collaborations with external bodies including private and public entities, and we are seeking further engagement through partnerships and outreach.

We describe here the initial phase of ADWG efforts, focusing on pilot studies that evaluate two core areas in clinical genomics: (1) ascertainment of patient race, ethnicity, and ancestry (REA); and (2) information disparity across populations for genetic variants that are clinically relevant. We are also surveying clinical genomics practitioners to investigate knowledge, beliefs, and use of REA in variant curation and clinical applications. Our findings are meant to provide a baseline understanding in the clinical genomics community about how race, ethnicity, and ancestry are conceptualized and utilized in current practice, with the future aim of developing standards and recommendations to guide how these concepts inform sequence variant interpretation and subsequent decision making by clinical practitioners. While these baseline efforts are focused on clinical laboratories and practitioners in the United States, in the future we intend to partner with external groups and expand our analyses to other countries, which will build on lessons learned from these US-based pilot studies. Expanding our analyses to genomic databases and clinical settings outside the United States may provide more diverse data and will help us to better understand the use of REA in a more global context.

The striking information disparity between European ancestry populations and all others in terms of available data underscores the importance of continued efforts to diversify the genomic evidence base. It also strongly suggests that uncertainty around allele frequency estimates is likely to be much higher for non-European ancestry populations, as well as downstream analyses, such as disease risk predictions may be less accurate for individuals from these populations also.

2 | METHODS

2.1 | Clinical ascertainment

To establish a baseline understanding of how race, ethnicity, and ancestry are ascertained in a clinical genomics setting, we analyzed genetic test requisition forms (RFs) from 10 clinical laboratories that have contributed the most variant interpretations to ClinVar and have analyzed more than two genes and conditions. For this analysis, *Clinvar Miner* (Henri et al., 2018) was accessed on February 3, 2018 to identify a list of the top 10 submitters to ClinVar filtering on criteria for collection method (clinical testing) and submission review status (evidence provided with the assertion from at least one submitter). Laboratory websites were manually accessed to download RFs between February 3, 2018 and June 3, 2018. Preference was given to prenatal or carrier testing, hereditary cancers, and general RFs. In one case, laboratory RFs were not available online, but were located through websites of other laboratories that use them. In another case, the laboratory RFs were not available from the laboratory website to nonclinicians, so the lab sent us a patient history form, which asks the same REA question as their test RFs. However, most laboratory RFs were available online and provided consistent categories across forms; in total, we analyzed forms from 10 different labs that were among the top submitters to ClinVar and had RFs publicly available.

If consistency in the RF options for REA was observed among at least three RFs, then one was chosen as the "representative" RF for that laboratory. Only one of the 10 laboratories analyzed provided different REA options to choose from across the RFs viewed, and in

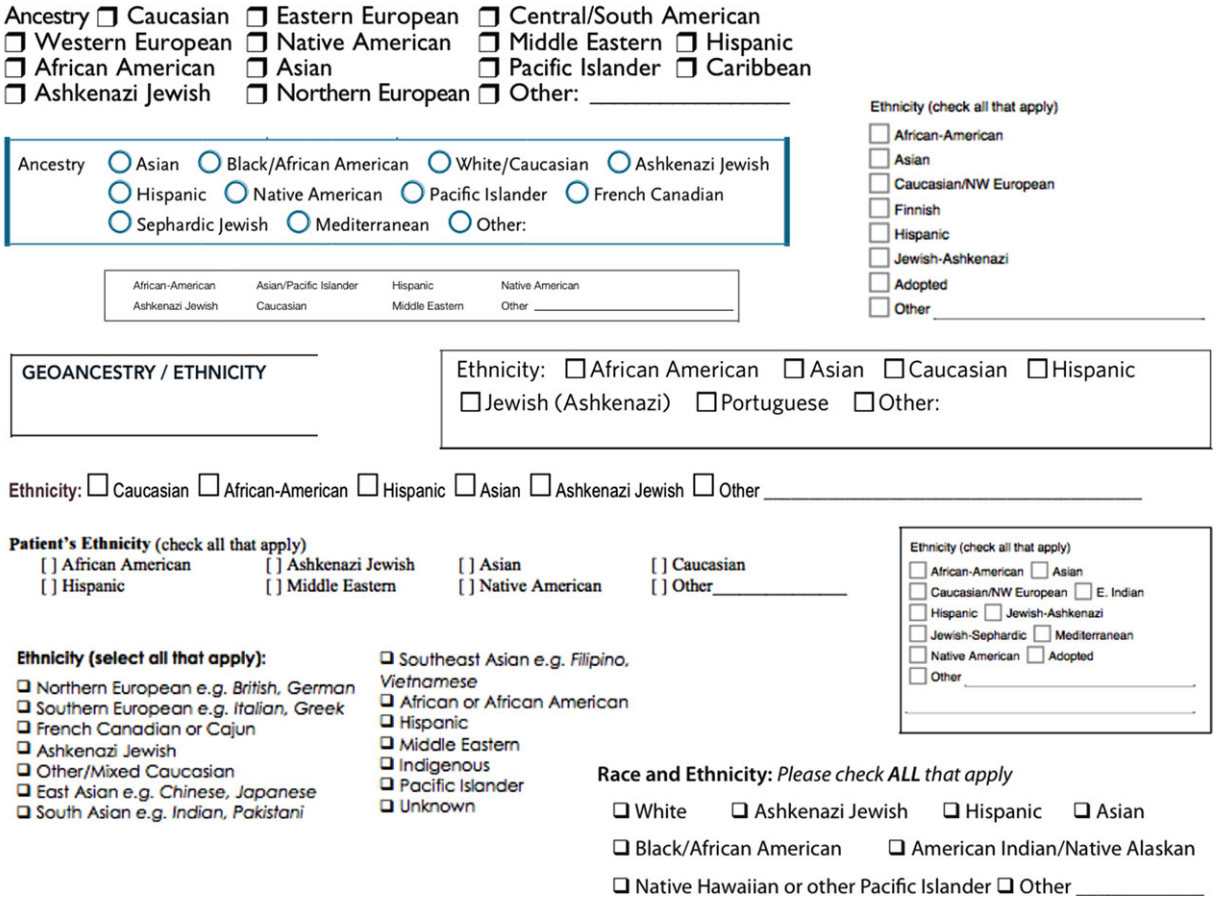

FIGURE 1 Screenshots of race, ethnicity, and ancestry questions on clinical laboratory requisition forms

that case the RF with the maximal number of REA options was used. Screenshots of the relevant questions and possible answers were captured and logged in a working document, examples of that are shown in Figure 1. REA options provided on different forms were recorded in an Excel file, in addition to labels used to designate that section on the RF (such as "Ancestry" or "Race and Ethnicity"), which differed across forms. Finally, we tabulated the number of times each specific REA option was observed in any RF and associated it with a representative category from the population ontology used by MedGen (https://www.ncbi.nlm.nih.gov/medgen/), which is the ontology utilized in ClinVar.

2.2 | Information disparity

To create the variant set for an information disparity analysis, we obtained a .csv file of all unique GRCh37 (hg19) "pathogenic" or "likely pathogenic" entries in ClinVar for single nucleotide variants that have been reviewed by an expert panel (N = 1,661). We further restricted the analysis to genes with at least 10 variants observed in ClinVar; 96% of the initial variants at unique sites were found within eight genes (N = 1,585 variants). Population-level data including allele counts, number of private alleles, and numbers of observations were obtained from population groupings in gnomAD (accessed April 2018): African/African American (AFR), Ashkenazi Jewish (ASJ), East Asian (EAS), non-Finnish European (NFE), Finnish (FIN), Latino or admixed American (AMR), and South Asian (SAS). We further restricted our analysis to variants identified from ClinVar that had corresponding entries in gnomAD exomes or genomes (N = 550) to determine the level of information disparity at those sites.

3 | RESULTS

3.1 | Ascertainment heterogeneity and ambiguity

Our analysis of RFs revealed considerable heterogeneity across clinical laboratories in the way REA data are ascertained. Out of questions asked on RFs from 10 different laboratories, only one asked an open-ended question with a blank "Geoancestry/Ethnicity" field. All others provided predetermined selections in a multiple-choice or best-choice response format. Across forms from the same clinical lab, the questions and possible responses were usually identical, regardless of the type of test. Among labs, however, the combination and number of preset selections varied widely. Some specific race and ethnicity categories were consistent across most forms, such as "African American" (9/10), "Ashkenazi Jewish" (8/10), "Asian" (8/10), and "Hispanic" (9/10). Variable REA terms were used to describe European-ancestry populations, such as "White" (2/10) and "Caucasian" (7/10). Many forms also offered more granular European categories based on geographic origin, such as: "Northern European, for example, British, German," "Southern European, for example, Italian, Greek," "Western European," and "Eastern European." Similarly,

categories based on geography were used to define various Asian populations on one form: "East Asian, for example, Chinese, Japanese," "South Asian, for example, Indian Pakistani," and "Southeast Asian, for example, Filipino, Vietnamese." In contrast, large continents and highly diverse populations were described using racial or cultural categories, such as "Black/African American" (1/10), "African or African American" (1/10), "Native American" (6/10), and "Indigenous" (1/10) with no geographic specificity. Clinical laboratory RFs thus varied in how they identified populations, which included continental ancestral population, country of origin, self-identified race, and ethnicity, often without distinguishing between the different types of population identifiers.

To organize these heterogeneous REA labels into broad categories, we identified the MedGen population codes used in ClinVar and mapped each of the terms to one of those categories (Table 1). Among the nine RFs that offered preset REA selections, there were 38 unique terms used to describe roughly eight large continental populations or ancestral designations found in MedGen, and the majority of those terms (28/38) were used only once on a single RF. Ten of the unique descriptors observed on RFs are variations of a European ancestry population, typically a specific geographic region. In contrast, there were only three very similar descriptors available for African Americans, none of which refers to any geographic region, and all three of which included the term "African American." There were also three instances of terms used to describe Ashkenazi Jewish ancestry or ethnicity, and all three of them included "Jewish." Finally, there were several terms used to designate groups that did not clearly correspond to any MedGen categories, including "Caribbean," "Central/South American," "French Canadian or Cajun," and "Jewish-Sephardic," each of which were each seen only once, and "Middle Eastern," which was seen four times.

Overall, no two clinical laboratories provided the same descriptive categories to designate a group or population on their RFs. Nor was there consistency in the way laboratories described the category of information being sought about REA. Metadata headers (for the REA section of lab RFs) included terms such as "ancestry," "ethnicity," "race and ethnicity," or blank (preselected race and ethnicity options, field was not named). Table 1 shows all of the terms used to describe various groups and populations, organized by their corresponding MedGen category as determined by the authors. We acknowledge that there may be differences of opinions with how each of these terms is associated (or not) with the MedGen categories and this presentation is not meant to suggest that this is the best framework for mapping those associations. Rather, it is intended to organize the diverse categories we observed and clearly show the heterogeneity and ambiguity with which people are grouped in a clinical genomics setting, illuminating the importance of efforts to improve the ascertainment of self-reported race and ethnicity.

TABLE 1 Terms used to describe human racial, ethnic, and ancestry groups on requisition forms for highly productive ClinVar-submitting laboratories

African American; MedGen:C0085756 African American (6) African or African American (1) Black/African American (2)	**Hispanic Americans; MedGen:C0019576** Hispanic (9)
American Indian or Alaska Native; MedGen:C1515945 American Indian/Native American (1) Native American (5) Indigenous (1)	**Native Hawaiian or Other Pacific Islander; MedGen:C1513907** Native Hawaiian or other Pacific Islander (1) Pacific Islander (3)
Ashkenazi Jew; MedGen:CO337704 Ashkenazi Jewish (7) Jewish (1) Jewish Ashkenazi (1)	**Chinese; MedGen:C0152035/South East Asian; MedGen:C0238697** Asian (7) Asian/Pacific Islander (1) E. Indian (1) East Asian, for example, Chinese, Japanese (1) South Asian, for example, Indian Pakistani (1) Southeast Asian, for example, Filipino, Vietnamese (1)
Caucasians; MedGen:C0043157 / European Caucasoid; MedGen:C0682087 Caucasian (5) Caucasian/North Western European (1) Eastern European (1) Mediterranean (1) Northern European (1) Northern European, for example, British, German (1) Portuguese (1) Southern European, for example, Italian, Greek (1) Western European (1) White (1) White/Caucasian (1)	**Mixed Ethnic Group; MedGen:C0682086** Other/mixed Caucasian (1)
	None/Unspecified Adopted (1) Other: _____ (8) Unknown (1)
	(Other: No MedGen category clearly applies) Caribbean (1) Central/South American (1) French Canadian or Cajun (1) Jewish Sephardic (1) Middle Eastern (4)

Each term has a number in parentheses after it, which represents the number of forms on which the corresponding term was observed. Theme headers are the MedGen categories used as population descriptors in ClinVar, and terms are matched to these categories.

3.2 | European-biased clinical genomic evidence

It is well-established that genomic databases on which clinical variant interpretations are made are biased toward European ancestry, but the amount of information from different REA populations specific to clinically relevant sites is less clear. We conducted an analysis to quantify the information disparity across populations in two databases that are typically used for clinical variant curation: ClinVar (Landrum et al., 2014) and the Genome Aggregation Database (gnomAD) (Lek et al., 2016).

Of the 1,805 variants that have been reported to ClinVar and determined by an expert panel to be "pathogenic" or "likely pathogenic," 1,661 of these are found at unique sites in the genome. Of these unique

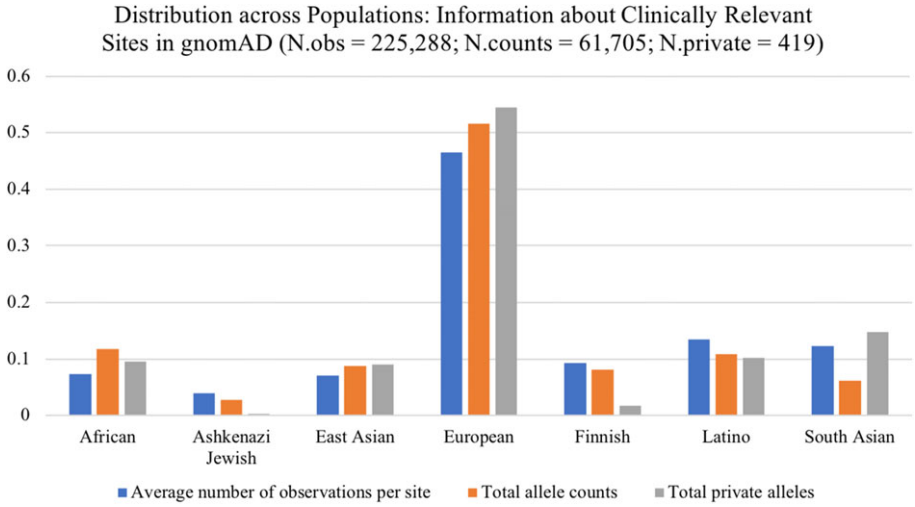

FIGURE 2 Information disparity at clinically relevant sites across populations. Bar graph shows the distribution of information that is available for different ancestral populations in gnomAD, for variants in ClinVar that are annotated by expert review as "pathogenic" or "likely pathogenic." The percentage of average total observations (blue, $N = 225,288$) is 7.4% African (AFR), 4% Ashkenazi Jewish (ASJ), 7.1% East Asian (EAS), 46.5% European (NFE), 9.3% Finnish (FIN), 13.5% Latino (AMR), and 12.3% South Asian (SAS). The percentage of total allele counts (orange, $N = 61,705$) observed across populations is 11.8% AFR, 2.7% ASJ, 8.1% EAS, 51.6% NFE, 8.2% FIN, 10.9% AMR, and 6.2% SAS. Private alleles are those found only in a single population (gray, $N = 419$); the percentage of all private alleles observed was 9.5% in AFR, 0.2% ASJ, 9.1% EAS, 54.4% NFE, 1.7% FIN, 10.3% AMR, and 14.8% SAS

sites, 1,585 (96%) are found in eight genes [*BRCA1* (388), *BRCA2* (473), *CFTR* (154), *MLH1* (232), *MSH2* (195), *MSH6* (72), *MYH7* (44), and *PMS2* (27)]. Among the 1,585 clinically relevant sites in these genes, 550 (35%) had a corresponding variant entry in gnomAD, not surprisingly, as most pathogenic variants represent rare alleles in population databases (Lek et al., 2016). We determined what proportion of total observations, allele counts, and private (population-specific) alleles in a particular gene were attributed to each population and illustrate the resulting distribution in Figure 2. Our primary finding is that all three of these (nonindependent) metrics show the majority of information is from European-ancestry individuals.

Overall, roughly 50% of all information about clinically relevant variants (number of observations at each site, observed allele counts, and alleles that are observed in only one population, "private alleles") came from non-Finnish European (NFE)-ancestry individuals. It is worth noting that closer to 60% of all information can be attributed to "White" racial groups, including NFE, FIN, and ASJ. Every population except for NFE contributed an additional approximately 10% of the total information. Latino or admixed Americans (AMR) had the highest number of total observations among non-Europeans (13.5%) while EAS had the lowest (7.1%). After NFE, SAS had the highest proportion of population-specific alleles (14.8% that are specific to South Asia) while ASJ and FIN had noticeably fewer (0.2% for Ashkenazi Jewish and 1.7% for Finnish, respectively). All comparisons are highly significant due to the depth of sequencing available in gnomAD (pairwise comparisons of counts: total observations, all $P < 1\times 10^{-4}$).

4 | DISCUSSION

Our findings from both the clinical laboratory RF analysis and the quantification of information disparity across populations at clinically relevant genetic loci reveal an underlying deficiency in clinical genomics research and implementation in diverse populations. How clinical labs ascertain information about REA is so highly variable that large NIH-supported consortia, such as the Clinical Sequencing Evidence-Generating Research (CSER, https://cser-consortium.org) are actively engaged in discussions to harmonize these measures across clinical sites in the United States. Indeed, the marked heterogeneity we found here in REA ascertainment by clinical laboratories may have far-reaching downstream effects for research and clinical practice in the era of genomic technology. There is a growing literature describing the downstream effects of a European-biased evidence base for genomic medicine (Landry & Rehm, 2018). For example, a variant that had been determined pathogenic based on observations in mostly European-ancestry individuals led to pervasive false positive diagnoses of African Americans for hypertrophic cardiomyopathy (Manrai et al., 2016). Subsequent invasive treatments were determined to have been unnecessary after the variant was observed in a sufficient number of healthy African Americans, then redesignated as benign. As demographics shift toward a more diverse population in the United States, clinical genomics must focus on ways to improve variant interpretation and reclassification for individuals of admixed and non-European ancestry.

Our results showed prevalent information disparities among populations at clinically relevant sites, and only one-third of sites in the genome with "pathogenic" and "likely pathogenic" variants in ClinVar have readily available population-level data. This has important implications for variant curation and interpretation using the American College of Medical Genetics (ACMG) Guidelines (Richards et al., 2015) as several of those criteria rely on allele frequencies. One criterion for moderate evidence of pathogenicity is specifically related to the absence of a variant from "population databases" (PM2 evidence code). However, absence of a variant is difficult to distinguish

from incomplete sampling across all underrepresented populations in genomic databases (currently applies to all non-European ancestry groups). Also, pathogenic variants tend to be rare, so many more observations from diverse groups in large population databases will be necessary. Further complicating the issue of using allele frequency data in clinical genomics is the nomenclature incongruity between RF REA collection and major control database populations (i.e., "Asian" on clinical laboratory RFs vs. "South Asian" or "East Asian" in ExAC and gnomAD). Such inconsistencies across population labels make it challenging to directly relate variant information in the clinic to population allele frequencies. Equitable sampling across all geographic regions and REA groups would begin to alleviate these complications, particularly as they relate to the PM2 evidence code for variant interpretation in the ACMG guidelines, since absence of a variant from "population databases" would be more informative.

The issue of whether and how to categorize ancestral populations in clinical genomics for improved healthcare, in laboratory RFs and online public sequencing repositories, is an admittedly controversial topic. Some argue that whole-genome and whole-exome sequencing will alleviate the need to ascertain self-reported REA measures, since genetic ancestry can be directly inferred from DNA; however, disease outcomes vary between racial and ethnic groups to a larger extent than can be explained by genetic differences and this needs to be accounted for in genomic medicine (Burchard et al., 2003). Thus, carefully collecting these data is necessary, and conceptualizing the various meanings of "race," "ethnicity," and "ancestry" is an active and ongoing debate (Hunt & Megyesi 2008). Historical and social science research documents how REA terms have varied meaning and are fluid constructs (Race, Ethnicity, and Genetics Working Group, 2005; Smedley & Smedley, 2012; Yudell, 2014). Terms such as "Hispanic," which was present on all RFs that offered an REA choice, are overall not particularly descriptive, despite its usage as an ethnicity metric in the US Census. Additionally, terms such as "Adopted," "Other/mixed Caucasian," "Unknown," or "Indigenous" may not be at all informative for clinical variant interpretation. Categories describing smaller geographic regions, religious groups, or other genetically isolated populations ("French Canadian/Cajun," "Ashkenazi Jewish", etc.) may help to raise awareness of population bottlenecks and can inform variant interpretation of unique allele frequencies due to founder effects or genetic drift. Conversely, whether large geographic/continental regions, such as "African," "Asian," "Western European," and "South/Central American" have clinical utility as population designators is largely unexamined in the clinical genomics space.

Social and political issues are prevalent in population identity, especially for groups that have been traditionally marginalized and/or subjected to discrimination. For example, "Native American" may be used colloquially but on official documents, "American Indian or Alaska Native" is typically the preferred terminology, since this designation is used in all treaties with the US government. Likewise, in anecdotal accounts, patients have reported hesitation to self-identify as "Jewish," "Ashkenazi," or "Sephardic" because such labels have been used for political persecution. How this information should be used in clinical genomics needs further research. Deciding the most optimal format and terminology for research purposes and clinical applications as well as harmonizing across laboratories, online repositories of genetic data and sequencing projects will require a dedicated collaborative effort across public and private sectors.

Our analysis of information disparity at pathogenic sites substantiats the necessity of increasing diversity in genome sequencing. Historically, methodologies and resources in genomics research have been designed for and implemented on mostly White, European-ancestry individuals, and this legacy is perpetuated in current research. Some methodologists are working to leverage unique characteristics of ancestrally diverse genomes to promote discovery (Aschard, Gusev, Brown, & Pasaniuc, 2015; Park et al., 2018; Wojcik et al., 2017), and others are exploring effective incentives to sequence more diverse populations. Currently, national efforts are underway to increase genomic diversity in clinical sequencing and research, including NIH-supported All of Us (https://allofus.nih.gov), CSER (https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-16-011.html), Genome Sequencing Program (https://www.genome.gov/10001691/nhgri-genome-sequencing-program-gsp/), Population Architecture Using Genomics and Epidemiology (Matise et al., 2011; Wojcik et al., 2017) (NHGRI), and Trans-Omics for Precision Medicine (TOPMed, https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program) (National Heart, Lung, and Blood Institute). Additionally, attempts to standardize ancestry categories for research have been undertaken, such as *Ancestro* for the Catalog of Genome-Wide Association Studies (GWAS Catalog) (Morales et al., 2018). If adopted, existing frameworks will need to be harmonized and built upon in the development of standard REA categories for genomic databases, and other applications in clinical genomics, such as electronic health records and laboratory RFs. The ClinGen ADWG is positioned to influence this process of developing and spearheading adoption of these harmonized frameworks, and its members are actively working with European Bioinformatics Institute (EBI–EMBL), which hosts the GWAS Catalog, as well as other consortium projects to lay the groundwork for this effort.

Additionally, there are deeply entrenched systems that inhibit broader inclusion of diverse populations, such as structural and institutional racism (Parsa et al., 2013), historical misuse of data (Garrison, 2013), and subsequent mistrust of the biomedical research community (Ferrera et al., 2016; Gamble, 1997). As such, there is an inherent tension between the need to acknowledge past harms to communities of color and encouraging participation in genomics research, as well as data sharing, to advance the equitable distribution and effectiveness of precision medicine in all REA populations. Addressing racial and ethnic disparities in genome interpretability requires sharing of data across millions of sequenced genomes and exomes with rich metadata, including context of sequencing (for care, research, DTC, etc.) and REA labels. NHGRI-funded research sequencing efforts are a start, since they promise hundreds of thousands of ethnically diverse exomes and genomes and resulting population allele frequency resources by 2021. Meanwhile, millions of genomes, exomes, and large panel tests are likely to be sequenced in the context of clinical care. Therefore, continued development of data standards and secure sharing protocols for clinical genomics are critical for ameliorating population-level disparities in clinical genomics.

Additionally, investments in public health interventions and community resources must be prioritized in parallel to genomic sequencing and database development, to address disparities in disease risk and health outcomes that are due to environmental inequities in underserved populations.

As scientists and clinicians, we need to carefully balance the utility of REA information to improve research and clinical care while not reifying harmful, unfounded beliefs about biological differences between racial groups (Fitzgerald, 2014; Roberts, 2011). Race, ethnicity, and ancestry impact health risks and outcomes, as in certain diseases, such as asthma, kidney, heart disease, and diabetes (Estrada et al., 2014; Landry & Rehm, 2018; Manrai et al., 2016; Parsa et al., 2013; Torgerson et al., 2011) as well as pharmacogenomics (Ramamoorthy, Pacanowski, Bull, & Zhang, 2015). However, determining the degree to which genetic and/or shared environmental factors (each of which is differently associated with REA measures) influence complex traits is difficult, and will take both genomic and public health solutions to elucidate. The challenge for the clinical genomics community is to determine how to responsibly usher in the era of precision medicine for all patients (Bonham, Callier, & Royal, 2016), and the ClinGen ADWG is working toward the development of recommendations to do just that.

5 | CONCLUSION

Inclusion of diverse patients and research participants in clinical genomics is a critical piece of addressing the information disparity in our clinical genomic evidence base; however, development and standardization of genetic ancestry as well as self-reported race and ethnicity measures are also key components. While large-scale whole-genome and -exome sequencing may someday fully elucidate the genomic architecture of diverse ancestral populations, genomic data alone do not account for population-level differences in disease outcomes, even for classic Mendelian traits. Furthermore, these technologies do not address multifactorial (social, economic, environmental, etc.) components of complex disease etiology. This is where carefully determined REA measures can provide additional layers of understanding of disease-gene and -variant associations, and they must be harmonized across large-scale research consortia, public and private clinical laboratories, and clinical points of care. Future research of the ClinGen ADWG will seek to generate more data and provide guidance on the implications of uncertainty around allele frequency estimates in diverse populations; and in the meantime, those involved in variant curation and interpretation efforts using population-level data should be mindful of information disparities across populations. Ultimately, the clinical genomics community must agree on how to appropriately handle the collection and use of REA, as well as the utility of this information. The ClinGen ADWG is called upon by NHGRI to provide guidance on these issues. We welcome others to co-create solutions with us as we actively seek collaborations for the next phase of our efforts in partnership with existing research and governing bodies both locally and globally.

ACKNOWLEDGMENTS

Additional members of the ClinGen Ancestry and Diversity Working Group (ADWG) who are not listed as authors include: Berg, J., Dwight, S., Kenny, E., Lee, S.S., Ormond, K.E., Oh, S., Prahbu, S., Rivas, M., Shields, A.E., Thornton, T., and Williams, D., many of whom participated in conversations that led to the development of these projects and/or provided minor edits on the manuscript. This publication was supported by the National Human Genome Research Institute of the National Institutes of Health through the following grants and contracts: U41HG009649 and U41HG009650. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID

Alice B. Popejoy http://orcid.org/0000-0003-4976-0628
Mathew W. Wright http://orcid.org/0000-0002-2650-2426

REFERENCES

Aschard, H., Gusev, A., Brown, R., & Pasaniuc, B. (2015). Leveraging local ancestry to detect gene-gene interactions in genome-wide data. *BMC Genetics*, *16*, 124. https://doi.org/10.1186/s12863-015-0283-z

Bonham, V. L., Callier, S. L., & Royal, C. D. (2016). Will precision medicine move us beyond race? *New England Journal of Medicine*, *374*(21), 2003–2005.

Bonham, V. L., Umeh, N. I., Cunningham, B. A., Abdallah, K. E., Sellers, S. L., & Cooper, L. A. (2017). Primary care physicians' collection, comfort, and use of race and ethnicity in clinical practice in the United States. *Health Equity*, *1*(1), 118–126.

Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., … Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, *348*(12), 1170–1175.

Caswell-Jin, J. L., Gupta, T., Hall, E., Petrovchich, I. M., Mills, M. A., Kingham, A. E., … Kurian, A. W. (2018). Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. *Genetics in Medicine*, *20*(2), 234–239.

Ferrera, M. J., Feinstein, R. T., Walker, W. J., & Gehlert, S. J. (2016). Embedded mistrust then and now: Findings of a focus group study on African American perspectives on breast cancer and its treatment. *Critical Public Health*, *26*(4), 455–465.

Fitzgerald, K. J. (2014). The continuing significance of race: Racial genomics in a postracial era. *Humanity & Society*, *38*(1), 49–66.

Gamble, V. N. (1997). Under the shadow of tuskegee: African Americans and health care. *American Journal of Public Health*, *87*(11), 1773–1778.

Garrison, N. A. (2013). Genomic justice for native Americans: Impact of the havasupai case on genetic research. *Science Technology & Human Values*, *38*(2), 201–223.

Green, E. D., & Guyer, M. S.; National Human Genome Research Institute. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, *470*(7333), 204–213.

Henri, A., Hemphill, S. E., Ruiz-Schultz, N., Cushman, B., DiStefano, M. T., Azzariti, D., … Eilbeck, K. (2018). ClinVar miner: Demonstrating utility of a web-based tool for viewing and filtering ClinVar data. *Human Mutation*, *39*(8):1051–1060. https://doi.org/10.1002/humu.23555

Hiatt, S. M., Amaral, M. D., Bowling, K. M., Finnila, C. R., Thompson, M. L., Gray, D. E., … Cooper, G. M. (2018). Systematic reanalysis of

genomic data improves quality of variant interpretation. *Clinical Genetics*, *94*(1):174–178. https://doi.org/10.1111/cge.13259

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., & Green, E. D. (2018). Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, *19*(3), 175–185.

Hunt, L. M., & Megyesi, M. S. (2008). The ambiguous meanings of the racial/ethnic categories routinely used in human genetics research. *Social Science & Medicine*, *66*(2), 349–361.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(Database issue), D980–D985.

Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs*, *37*(5), 780–785.

Landry, L. G., Rehm, H. L. (2018). Association of racial/ethnic categories with the ability of genetic tests to detect a cause of cardiomyopathy. *JAMA Cardiology*, *3*(4), 341–345.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*, 285.

Manolio, T. A., Fowler, D. M., Starita, L. M., Haendel, M. A., MacArthur, D. G., Biesecker, L. G., … Bult, C. (2017). Bedside back to bench: Building bridges between basic and clinical genomic research. *Cell*, *169*(1), 6–12.

Manrai, A. J., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., … Kohane, I. S. (2016). Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, *375*, 655–665.

Matise, T. M., Ambite, J. L., Buyske, S., Carlson, C. C., Cole, S. A., Crawford, D. C., … PAGE Study. (2011). The next page in understanding complex traits: Design for the analysis of population architecture using genetics and epidemiology (PAGE) study. *American Journal of Epidemiology*, *174*(7), 849–859.

Morales, J., Welter, D., Bowler, E. H., Cerezo, M., Harris, L. W., Hindorff, L. A., … MacArthur, J. A. L. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biology*, *19*(1), 21.

Park, D. S., Eskin, I., Kang, E. Y., Gamazon, E. R., Eng, C., Gignoux, C. R., … Zaitlen, N. (2018). An ancestry-based approach for detecting interactions. *Genetic Epidemiology*, *42*(1), 49–63.

Parsa, A., Kao, W. H., Xie, D., Astor, B. C., Li, M., Hsu, C. Y., … Appel, L. J. (2013). APOL1 risk variants, race, and progression of chronic kidney disease. *New England Journal of Medicine*, *369*(23), 2183–2196.

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*, 161–164.

Race, Ethnicity, and Genetics Working Group. (2005). The use of racial, ethnic and ancestral categories in human genetics research. *American Journal of Human Genetics*, *77*(4), 519–532.

Ramamoorthy, A., Pacanowski, M. A., Bull, J., & Zhang, L. (2015). Racial/ethnic differences in drug disposition and response: Review of recently approved drugs. *Clinical Pharmacology and Therapeutics*, *97*(3), 263–273.

Ramos, E., & Weissman, S. M. (2018). The dawn of consumer-directed testing. *American Journal of Medical Genetics*, *178*(1), 89–97.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (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. *Genetics in Medicine*, *17*(5), 405–424.

Roberts, D. (2011). *Fatal invention: How science, politics, and big business re-create race in the twenty-first century*. New York, NY: The New Press.

Roberts, J. S., Gornick, M. C., Carere, D. A., Uhlmann, W. R., Ruffin, M. T., & Green, R. C. (2017). Direct-to-consumer genetic testing: User motivations, decision making, and perceived utility of results. *Public Health Genomics*, *20*(1), 36–45.

SIGMA Type 2 Diabetes Consortium, Estrada, K., Aukrust, I., Bjørkhaug, L., Burtt, N. P., Mercader, J. M., García-Ortiz, H., … MacArthur, D. G. (2014). Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA*, *311*(22), 2305–2314.

Smart, A., Bolnick, D. A., & Tutton, R. (2017). Health and genetic ancestry testing: Time to bridge the gap. *BMC Med Genomics*, *10*(1), 3. https://doi.org/10.1186/s12920-016-0240-3

Smedley, A., & Smedley, B. D. (2012). *Race in North America: Origin and evolution of a worldview*. Emeryville, CA: Avalon Publishing.

Torgerson, D. G., Ampleford, E. J., Chiu, G. Y., Gauderman, W. J., Gignoux, C. R., Graves, P. E., … Nicolae, D. L. (2011). Meta-analysis of genome-wide association studies of asthma in ethnically diverse north American populations. *Nature Genetics*, *43*(9), 887–892.

Travassos, C., & Williams, D.R. (2004). The concept and measurement of race and their relationship to public health: A review focused on Brazil and the United States. *Cad. Saúde Pública, Rio de Janeiro*, *20*(3), 660–678.

Williams, D. R., & Mohammed, S. A. (2013). Racism and health I: Pathways and scientific evidence. *American Behavioral Scientist*, *57*(8), 1152–1173.

Wojcik, G., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., … Carlson, C. S. (2017). Genetic diversity turns a new page in our understanding of complex traits. *bioRxiv Preprint*, https://doi.org/10.1101/188094

Yu, J. H., Taylor, J. S., Edwards, K. L., & Fullerton, S. M. (2012). What are our AIMs? Interdisciplinary perspectives on the use of ancestry estimation in disease research. *AJOB Primary Research*, *3*(4), 87–97.

Yudell, M. (2014). *Race unmasked: Biology and race in the twentieth century*. New York, NY: Columbia University Press.

How to cite this article: Popejoy AB, Ritter DI, Crooks K, et al. The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Human Mutation*. 2018;39:1713–1720. https://doi.org/10.1002/humu.23644