Using the Data We Have: Improving Diversity in Genomic Research

Teri A. Manolio[1,*]

The shortage of genomic research data in persons of non-European ancestry is impeding our ability to use genomics in the clinical care of non-European individuals. Improved efforts to utilize data on non-European populations will increase the quality of genomic research and the inferences drawn from it for people of all backgrounds.

The lack of ethnic diversity and concomitant heavy emphasis on European ancestry (EA) populations in human genomic research have been well documented and widely decried.[1] Euro-centricity of genomic research has serious implications for the health and medical care of non-EA populations, as it increases the likelihood non-EA individuals will receive inconclusive results of genetic testing or, worse, erroneous interpretations of genomic variants.[1,2] Inadequate data on risk allele frequencies and their associated risks in non-EA populations reduces the accuracy of both monogenic and polygenic risk predictions and limits the potential for applying them across different ancestry groups.[1]

The benefits of increased diversity in human genomic databases, however, extend beyond the healthcare of non-EA individuals. Association analyses in persons of non-European ancestry have been shown to identify a disproportionately larger number of associated alleles,[3] which enhances the quality of gene-disease association research for everyone. Finding high frequencies of presumed risk alleles in populations without a high prevalence of disease is also powerful evidence against inferences of disease causality, as has been demonstrated for variants initially inferred to be pathogenic for hypertrophic cardiomyopathy but later shown to be too frequent in black Americans to be disease causing.[1,2]

Reasons given for the lack of diversity in genomic studies include limited numbers of persons of non-European ancestry in existing genomic research studies and lack of trust in the biomedical research enterprise among non-EA individuals.[3] Non-EA cohorts that do exist outside the U.S. and Europe often have difficulty obtaining funding for genomic research, despite continued generous funding of EA cohorts. This is beginning to change, particularly with investments in non-EA minority populations within the U.S. and establishment of large-scale capacity-building programs internationally,[3] but more such investments are needed.

Concerns regarding potentially spurious findings due to correlated but non-causal differences in disease burden and allelic frequencies across populations (population stratification) have also led to focusing analyses on presumably more homogeneous populations,[3] the largest of which in genomic studies has almost always been that of European ancestry. Under-representation of investigators who are themselves of non-European ancestry has also been implicated (particularly in the U.S.),[3] not only because such populations may be of greater personal interest to them, but more importantly because of the unique perspectives they bring to health research in non-EA populations.

Using the Data We Have

Proposed solutions to this problem have largely—and correctly—focused on increased efforts to recruit and study much larger numbers of persons of non-European ancestry and facilitate integration of their data in commonly used databases, with appropriate consent and protections for privacy and confidentiality.[3] Much less attention has been paid to analyzing fully the limited numbers of non-EA participants currently included in many large studies, either due to concerns of population stratification (which can largely be identified and corrected) or because their numbers are vastly overshadowed by the numbers of EA participants.

Perhaps the most notable example of this can be found in published analyses of genomic data from the UK Biobank, a study of more than 500,000 British residents ages 40–69 recruited in 2006–2010. Genome-wide array data were assayed and imputed to more than 90 million variants in more than 487,000 individuals,[4] 88% of whom self-identified as being of white British ethnic background and another 6% as other white background. While the numbers of participants from other ethnic groups in the remaining 6% are dwarfed by these numbers (despite reflecting the UK population as a whole), they include more than 9,400 persons of self-reported Asian or Asian British ethnic background (mostly from the Indian subcontinent) and more than 7,600 Black or Black British. Not so long ago a genome-wide association study (GWAS) of 7,000 or 9,000 persons would have been considered sizeable; indeed, the NHGRI-EBI GWAS catalog

[1]National Human Genome Research Institute, Bethesda, MD 20817, USA
*Correspondence: manolio@nih.gov
https://doi.org/10.1016/j.ajhg.2019.07.008.

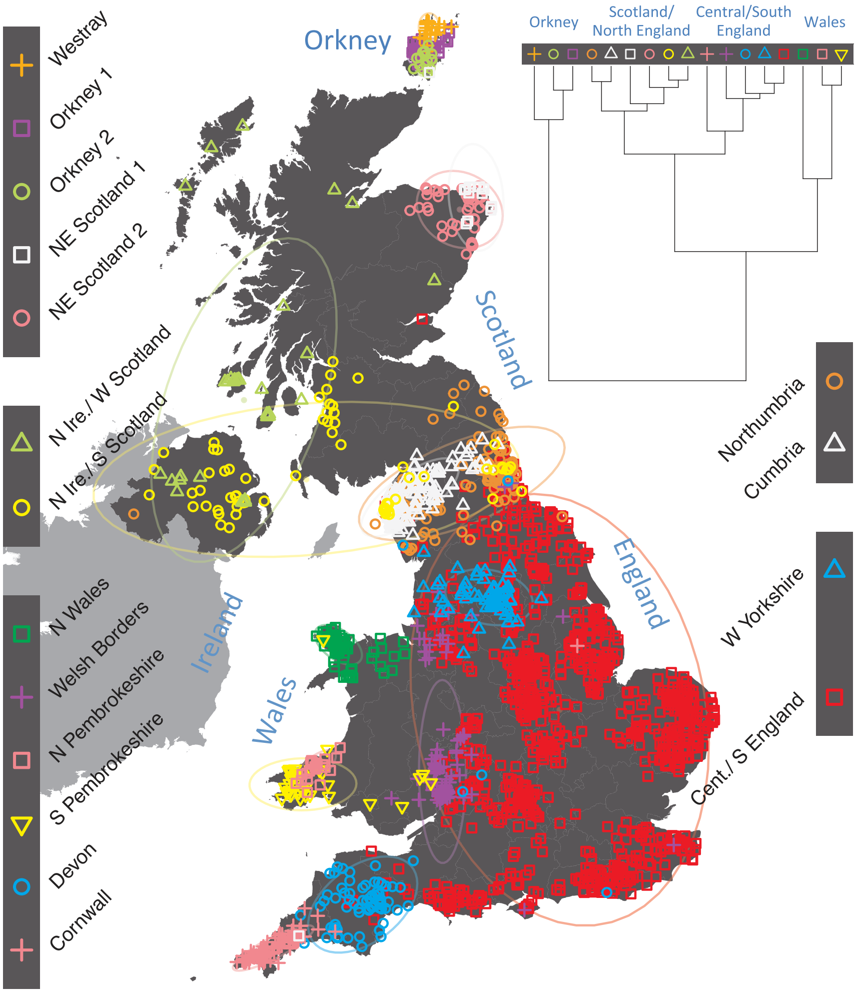

Figure 1. Clustering of the 2,039 UK Individuals into 17 Clusters Based Only on Genetic Data

For each individual, the colored symbol representing the genetic cluster to which the individual is assigned is plotted at the centroid of their grandparents' birthplaces. Cluster names are in side-bars and ellipses give an informal sense of the range of each cluster. The tree (top right) depicts the order of the hierarchical merging of clusters (adapted from Leslie et al.[7]).

includes 118 papers studying 6,000–9,000 persons, 103 of which examined continuous traits that lend themselves well to population studies.[5] This is a size range that might reasonably be expected from the two largest non-EA UK Biobank subgroups after exclusions for relatedness or sample quality. More than 20 such papers were published in the 14 months from January 2018 to February 2019 alone (the latest dates available in the catalog), suggesting that scientific interest remains strong in smaller GWASs. Yet every one of the 27 papers in the GWAS catalog with “UK Biobank” in the title, and 2 others presently in the catalog curation queue, limited their analyses to EA subgroups varying described as “White British,” “British,” “European,” “White European,” “Caucasian,” or “White.” Most cited a desire to avoid population stratification and many cited precedents set by earlier papers utilizing the UK Biobank resource.

This should in no way be considered a criticism of the creators of the UK Biobank, who have produced a scientific resource of inestimable value and made it widely available and easy to use. Their sharing of imputation and ancestry estimations developed by investigators such as Bycroft et al.[4] has greatly enriched the resource and facilitated the work of countless others. UK Biobank is used as an example here precisely because it is so widely used, and it is also relatively easy to identify in published work. But the near-uniform adherence by users of the resource to European-only genomic analyses is disconcerting, particularly for a field that prides itself on bold new approaches to data exploration and analysis.

This situation is reminiscent of the massive under-representation of gene-disease associations on the X chromosome, largely due to GWAS analyses often being limited to the autosomes.[6] Only 242 of 743 GWAS papers (33%) published in 2010 and 2011, for example, included X chromosome analyses.[6] For the calendar year 2018, only 152 associations were reported on the X chromosome, compared to 2,827 on chromosome 7, a chromosome of similar size, while only 19% of GWASs published in March 2019 included X chromosome analyses (A. Wise, personal communication). Reasons for this under-representation include the added, though minor, complexity of chromosome X association tests and possibly the slight lag in producing X chromosome imputation panels while investigators were busily analyzing imputed autosomal data. But perhaps most importantly,

> …most initial GWAS reports produced many useful autosomal findings while excluding the X chromosome from analysis, perhaps setting expectations that autosomal data alone were sufficient for high profile publications. Challenges in analyzing and interpreting X chromosome data, combined with the plethora of findings obtainable from the autosomes alone, might therefore lead many investigators to underutilize X chromosome data given that important associations can often be found without it.[6]

The parallels to the current exclusion of non-EA data from large-scale analyses are striking, and worrisome. If we are not to look back in 8 years

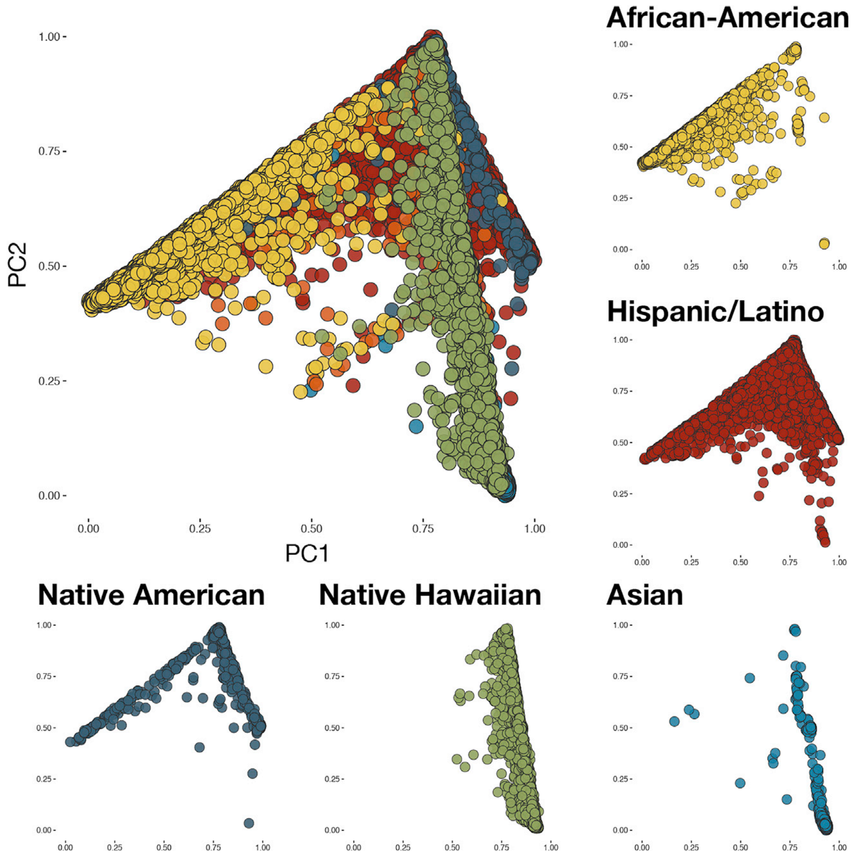

Figure 2. Population Substructure Present in the Multi-ethnic Sample of PAGE
The population substructure present in the multi-ethnic sample of PAGE (n = 49,839), showing a continuum along principal components 1 and 2 that prevents meaningful stratification (adapted from Wojcik et al.[10]).

and see a string of EA-only analyses of UK Biobank and similar large, multi-ethnic studies, we must resolve to do something about it now.

Methods Development

This is not to suggest that population structure is easily resolved—when numbers are very large, structure can be identified in even presumably "homogeneous" groups such as the white British (Figure 1[7]). The potential for standard corrections for population structure to be insufficient was recently demonstrated in an analysis of multiple GWAS datasets that had shown strong evidence of polygenic adaptation for height, signals of selection that were attenuated or absent in the UK Biobank white British.[8] Computationally efficient methods for correcting for population structure such as BOLT-LMM are increasingly being used, with appreciable increases in study power.[9] That reliable multi-ethnic analyses are both feasible and productive, albeit requiring meticulous attention to detail, was recently demonstrated by the identification of 27 novel loci for a variety of traits in nearly 50,000 non-European individuals in the Population Architecture using Genomics and Epidemiology (PAGE) study.[10] Multi-ethnic joint analyses were particularly important in PAGE, where the genetic ancestry of participants fell along a continuum rather than being clustered into discrete populations (Figure 2), as is likely the case in most modern societies. Continued efforts to apportion people to discrete ethnic groups for genetic analysis (which BOLT-LMM and similar methods do not require), unless these groups represent small and truly isolated populations, seem both counterproductive and unnecessary, and may increasingly lead us astray. Instead we should embrace genetic ancestry for the continuous variable that it is, and further develop methods to make the most of it in genetic analyses. Larger reference panels from diverse ancestry groups are also needed to improve imputation and are increasingly being produced.[3]

The Role of the Scientific Community, Funders, Peer Reviewers, and Journals

As a scientific community we need to pursue the exciting questions that multi-ethnic analyses can address, leaving behind outmoded perceptions of humans as clustering neatly into discrete population groups. Many large population studies include mixtures of ethnic groups, and as Wojcik et al.[10] have shown, they gain power and precision when appropriately combined. Specific studies of migrants and their descendants, such as the South Asian and Black subgroups in UK Biobank, and comparison to the populations from which they are derived can also yield useful insights. Funders should ensure that genomic research funding emphasizes non-European ancestry populations and should encourage and expect full use of the multi-ethnic data they have invested in collecting. New tools are needed for such analyses, as well as new explorations of where we might have been misled by past Euro-centric studies. Increased data sharing and simplified data access, as exemplified by resources such as UK Biobank, will facilitate such work, as would integration of data from multiple diverse ancestry groups into user-friendly and accessible resources. We should demand of ourselves and our colleagues the intellectual rigor to seek genomic knowledge that will benefit us all and not be satisfied with possibly easier but definitely less complete analyses with limited applicability to more than three-quarters of the world's population.

This spirit of intellectual curiosity and diligence should carry over into peer review, where reviewers should fully recognize the limitations of "single" ancestral group analyses and instead favor studies that fully utilize the data and samples provided by all participants, where such analyses are almost always scientifically appropriate. Even small numbers of persons of differing ancestries can be of enormous value, as noted by Manrai et al.[2] in their identification of false positive cardiomyopathy variants. By their estimate, inclusion of even small numbers of black Americans, such as are often set aside, would likely have prevented these false conclusions.[2] Examination of, and correction for, population substructure should be an integral part of all genomic analyses (aside from family-based analyses, which are generally considered immune to it[8]), even those considered to be from a "homogeneous" group. Reviewers should insist on compelling justifications for excluding any study subject solely on the basis of ancestry, and accept it only when a convincing case can be made. While there may be specific scientific questions for which an extremely conservative approach to minimizing confounding (by focusing on a single ethnicity) is warranted, these are likely quite infrequent, and might also benefit from further evaluation in a broader analysis.

Journals have great potential to shift the tide toward more complete analyses, since at present they seem so willing to publish studies that exclude participants based on ancestry alone. This may be hard to assess for every study, but for the better-known ones, reviewers and editors should carefully review exclusions based on ancestry and require very strong justification for them. As suggested by Hindorff et al.,[3] stronger publication standards should emphasize the need for diversity in research design and execution, and prioritize genomic studies that provide it. Existing recommendations for medical journal publications (Web Resources) should be broadened to include descriptions of the ancestral diversity of participants and/or explanations for a lack of diversity. One would hope such expectations would be widely publicized and not fall out of the blue on the shoulders of the next harried graduate student trying to work with invaluable resources such as the UK Biobank. We should expect senior scientists and those most familiar with such studies to pave the way by exemplar analyses that demonstrate the benefits and complexities of utilizing all participants' data to the fullest.

And while we're at it, let's look at the X chromosome.

Acknowledgments

The author gratefully acknowledges helpful input from Eric Boerwinkle, Vence Bonham, Rory Collins, Peter Donnelly, Eric Green, Lucia Hindorff, Jonathan Marchini, and Jonathan Pritchard.

Declaration of Interests

The author is a member of the UK Biobank International Scientific Advisory Board and was a founder, and is now a funder, of the NHGRI-EBI GWAS Catalog and the PAGE Study.

Web Resources

ICMJE Recommendations, http://www.icmje.org/icmje-recommendations.pdf

References

1. 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.; Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG) (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum. Mutat. *39*, 1713–1720.
2. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. N. Engl. J. Med. *375*, 655–665.
3. Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E.C., Hutter, C.M., Manolio, T.A., and Green, E.D. (2018). Prioritizing diversity in human genomics research. Nat. Rev. Genet. *19*, 175–185.
4. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
5. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47* (D1), D1005–D1012. 10.1093/nar/gky1120.
6. Wise, A.L., Gyi, L., and Manolio, T.A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am. J. Hum. Genet. *92*, 643–647.
7. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al.; Wellcome Trust Case Control Consortium 2; and International Multiple Sclerosis Genetics Consortium (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.
8. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., and Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. eLife *8*, e39725.
9. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.
10. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.