

The GWAS Diversity Monitor tracks diversity by disease in real-time

Melinda C. Mills*

Charles Rahal

Leverhulme Centre for Demographic Science and Nuffield College, University of Oxford, UK

To the Editor – The Genome-Wide Association Study (GWAS) is a primary tool for the discovery of associations between genetic variants and complex phenotypes, cataloged by the NHGRI-EBI GWAS Catalog, which currently contains information on over 4,346 published studies across over 4,933 diseases and traits. Although there has been a considerable expansion in study size, phenotypes under consideration, and the number of identified variants, the typical GWAS remains predominantly focused on European-ancestry samples. Our year-by-year estimates of European-ancestry sample prevalence range from as low as 65.93% of participants at the replication stage of studies published in 2012, to 92.18% at the discovery stage in 2019. Our cumulative estimate at the time of writing currently stands at 88.45%. This is despite the recent launch of initiatives such as H3Africa, the African Genome Variation Project, and GenomeAsia 100k. Geographic and demographic diversity is also limited, with other estimates showing that 72% of participants are recruited from just 3 countries (US, UK, Iceland).¹

The transferability of GWAS results across populations depends on many factors (such as allele frequency, linkage disequilibrium, genetic architecture, epistasis and gene-environment interaction),² with single-ancestry GWAS having limited portability.^{3,4} Polygenic risk scores (PRS) are influenced by environmental factors such as historical period, country of origin,⁵ demographic composition of age or sex, and socioeconomic status of individuals.⁶ With the move towards the use of PRS derived from GWAS for clinical applications,⁷ the majority of PRS derived from GWAS would exacerbate existing global health inequalities.⁸ There is substantially more genetic variation in non-European populations, and this variation can provide a rich resource for finding new genetic associations (Supplementary Note 1).

GWAS often fail to identify variants that differ in frequency amongst populations, either under- or over-estimating risk in the understudied. The predictive performance of PRS derived from European ancestry groups is systematically lower when applied to non-European samples.⁴ The majority of identified alleles may be the ones that reached higher frequencies by chance or reflect ascertainment bias in disease mutation discovery.⁹ Neglecting genetic differences across populations can have serious consequences for drug safety and medical treatments. For instance, African-Americans have a considerably higher risk of developing chronic kidney disease. This is largely attributed to the apolipoprotein-LI (*APOL1*) gene of which two versions (G1 and G2) became more common in people from sub-Saharan Africa over the last 10,000 years.¹⁰

Although the need to prioritize greater diversity has been recognized (Supplementary Note 2, Supplementary Table 1), there is currently no tool monitoring trends or identifying gaps by disease, ancestry, and geography. To this end, we announce the release of an interactive online dashboard called the GWAS Diversity Monitor (gwasdiversitymonitor.com). Automated with regular updates drawn from the EMBL-EBI GWAS Catalog (Supplementary Note 3),¹¹ it provides quasi-real-time monitoring of GWAS diversity and represents an accessible, user-friendly range of summary statistics and interactive visualizations. Users can examine all information on all studies across

multiple figures and tabs (see Figure 1 for a static representation). The two global widgets are 'Metric' — which toggles by number of studies or participants — and 'Stage' — which shows results by the discovery or replication phase of research. Local widgets include dropdown menus which allow selection of 'Parent Term', the 'Broader' ancestry category (Supplementary Note 2 and 4; Supplementary Table 2), and 'Year' (year of study). The GWAS Diversity Monitor promotes Open Science by allowing users to download all figures, underlying datasets, and code for supplemental analyses. The code is stored on Zenodo, a general-purpose open-access repository operated by CERN which provides a persistent digital object identifier (10.5281/zenodo.3600471).

The first panel shows cumulative estimates of participant ancestries across all time periods, stages and phenotypes. The interactive bubble plot shows ancestry by phenotype, parent category, research stage and study size over time. Clicking on bubbles provides detailed study information and redirects users to the specific study on PubMed. Searching across individual EFO terms is possible by entering three or more characters to unlock a dropdown menu (Supplementary Note 5). By clicking on the ancestry buttons, tailored graphics for each term can be produced. The time-series panel details how ancestry has changed over time by either the percentage of participants or number of studies across either stage of research. A toggle designed to add robustness to the time-series plot takes unrecorded ancestries into account. The heatmap illustrates ancestry by parent terms, also over time, visualizing gaps in past and current research. A world map visualizes geographical diversity of participant recruitment (Supplementary Note 6). The doughnut chart visualizes the number of participants studied by ancestry and parent term, with an option to enable a comparison with the count of all associations at the discovery stage.

An 'Additional Information' tab provides a range of summary statistics, also dynamically updated with each new iteration of the NHGRI-EBI GWAS Catalog. This includes a summary of the: total number of studies, most recently curated study, study with largest number of participants, total number of unique study accessions, number of disease and traits analyzed, number of EFO Mapped Traits analyzed, total and average number of associations found, mean P-Value for strongest SNP risk allele, number of associations reaching the $p < 5 \times 10^{-8}$ significance threshold, journal most frequently publishing GWAS, total number of journals publishing GWAS, and the most frequently studied (non-European ancestry) trait.. An accessible 'Q&A' is available as a separate page on the dashboard.

The GWAS Diversity Monitor helps to empirically isolate combinations of ancestry and parent terms where more, or conversely limited, research has been undertaken. We can see, for instance, that liver enzyme measurement has to date never been examined at either the discovery or replication stages in African ancestry populations and only at the replication stage for all groups except for European and Asian ancestries. We can also efficiently identify the most imbalanced areas of study. At the initial discovery phase in 2019, it is Cancer research, with European ancestry groups making up 96.27% of all GWAS participants compared to 0.11% African, 0.00% African American and Afro Caribbean, and 0.50% Hispanic or Latin American ancestries. This monitor provides a real-time and easily accessible tool for funders, scientists, policymakers, industry, and patient groups to chart the trends in GWAS diversity and highlight understudied areas of research in order to enact change.

Data Availability Statement

The data described in this manuscript comes solely from the NHGRI-EBI Catalog of published genome-wide association studies (otherwise known as the 'GWAS Catalog'). It is available at <https://www.ebi.ac.uk/gwas/>. Full and unreserved attribution to the GWAS Catalog team and is made both here and on the dashboard itself, and re-use of their data is subject to the Terms of Use for EMBL-EBI Services.

Code Availability Statement

The code which powers the dashboard is available on Zenodo (DOI: 10.5281/zenodo.3600471; zenodo.org/record/3600472) and GitHub (github.com/OxfordDemSci/gwasdiversitymonitor). It is made available under an MIT License.

References

1. Mills, M.C. and Rahal, C., *Commun Biology*, **2**, 9 (2019).
2. Sirugo, G., Williams, S.M. and Tishkoff, S.A., *Cell*, **177**, 1, 26-31 (2019).
3. Martin, A.R. et al., *Am J Hum Genet.*, **100**, 4, 635-649 (2017).
4. Duncan, L. et al., *Nat Commun.*, **10**, 3328, (2019).
5. Tropf, F. et al., *Nat Hum Behav.*, **1**(10), 757-765 (2017).
6. Mostafavi, H. et al., *bioRxiv*, (2019).
7. Torkamani, A. et al., *Nat Rev Gen.*, **19**, 581-590 (2018).
8. Martin, AR. et al., *Nat Gen.*, **51**, 584-591 (2019).
9. Amorim, C.E.G. et al., *PLOS Gen.*, **13**(9), (2017).
10. Genovese, G. et al., *Science*, **329**(5993), 841-845, (2010).
11. Buniello, A. et al., *Nuc Acids Res.*, **47**, D1005-1012 (2019).

Acknowledgements

MCM is supported by the ERC grants 615603 and 835079 and both MCM and CR by The Leverhulme Trust, Leverhulme Centre for Demographic Science. CR is supported by a British Academy Postdoctoral Fellowship. Additional acknowledgements are available on the dashboard.

Author contributions

The authors jointly designed the study, CR built a prototype of the on-line dashboard which was redesigned by MCM and CR in collaboration with Global Initiative. MCM wrote the main and supplementary material manuscript which was revised jointly by both authors.

Affiliations

Leverhulme Centre for Demographic Science and Nuffield College, University of Oxford.

Corresponding Author

Correspondence to Melinda C. Mills, Leverhulme Centre for Demographic Science, University of Oxford, Oxford, OX1 1JD, UK. Email: melinda.mills@nuffield.ox.ac.uk

Competing Interests

The authors declare no competing interests.

Figure 1

Figure 1. A Static Representation of the GWAS Diversity Monitor. A description of all details and figures can be found on the 'Additional information' tab. Available at gwasdiversitymonitor.com.

Total GWAS participants diversity

Version 1.0.0. Last check for data: 2020-01-05 12:56:35 GMT.

