Introducing the GWAS Diversity Monitor

Melinda C. Mills\* (ORCID: 0000-0003-1704-0001)

Charles Rahal (ORCID: 0000-0002-1764-2697)

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

Submitted as a ‘Letter’ to Science, 22nd August, 2019

\*Correspondence: Melinda C. Mills, Nuffield College, New Road, Oxford, OX1 1NF, UK Email: melinda.mills@nuffield.ox.ac.uk. Tel: +44(0) 1865 278500.

The proliferation of large biobanks and direct-to-consumer data has led to previously unimaginable sample sizes becoming available for Genome Wide Association Studies (‘GWAS’). The objective of such GWAS is to identify statistical associations between a set of genetic variants across different individuals (Single Nucleotide Polymorphisms, or ‘SNPs’) and specific traits of interest, often leading to substantial clinical findings. Advances have ranged widely, relating to diseases such as breast cancer and Alzheimer’s to anthropometric and behavioral traits. The first such study to be indexed by the EMBL-EBI GWAS Catalog ⁠— published in Science on April 15th, 2005 (1) ⁠— utilized 96 European ancestry cases and 50 European ancestry controls.

While sample sizes have increase substantially, the uniformity in ancestries utilized for such studies has mostly not. While the largest accession indexed to date contains 2,083,151 individuals for the ‘discovery’ stage and 287,239 for the replication stage (2), the largest 252 of all 14,005 Study Accessions are comprised of exclusively European samples (904 of the largest 1,000). Including a diverse pool of participants is critical in our search for understanding the genetic heterogeneity in disease phenotypes as we move towards an era of personalized medicine (such as in the discovery of rare, but statistically significant genetic variants). We also reemphasize the limited portability of polygenic scores across different populations, where prediction accuracies *within* ancestries are currently as high as 24.4% 43 for height (3) and up to 13% for educational attainment (4).

For that reason, we extend our former work (5) and introduce the GWAS Diversity Monitor: a five figure, five widget, two tab interactive dashboard with associated summary statistics which utilizes the GWAS Catalog. The back-end of the code checks for new updates daily, refreshing the dashboard when new updates are made. For more information, please see [gwasdiversitymonitor.com](http://gwasdiversitymonitor.com/).

1. RJ Klein *et al.*, “Complement Factor H Polymorphism in Age-Related Macular Degeneration”, *Science*, **308**, 5270, pp 385-389 (2005).
2. BML Baselmans *et al*., “Multivariate genome-wide analyses of the well-being spectrum”, *Nature Genetics*, **51**, 3, pp. 445-451 (2019).
3. L Yengo *et al*., “Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry”, *Human Molecular Genetics*, **27**, 20, pp. 3641-49 (2018).
4. JJ Lee *et al*., “Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals”, *Nature Genetics*, **58**, 8, pp.112-1121 (2018).
5. MM Mills and C Rahal, “A scientometric review of genome-wide association studies”, *Communications Biology*, **2**, 9, (2019)