GWAS diversity by disease requires real-time monitoring

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

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

**To the Editor** – Genome-Wide Association Studies (GWAS) are the primary tool for the discovery of genetic variants, with over 4,000 studies published across 3,500 diseases or traits (1). Although there has been a rapid expansion in cohort size, phenotyping and number of identified variants, GWAS remain predominantly based on European-ancestry samples, ranging between 96% (2007) with a decrease of 72% (2012) and subsequent increase to 88% (2017). Geographic diversity is also limited with 72% of discoveries from samples emanating from just 3 countries (US, UK, Iceland) (1).

The transferability of GWAS results to other populations depends on many factors (i.e., allele frequencies, linkage disequilibrium, genetic architecture, epitasis, gene-environment interaction) (2), with single-ancestry GWASs having limited portability (3). Polygenic risk scores (PRS) for complex traits are also influenced by the heterogeneity of environmental factors, such as historical period and country of origin (4). With the move towards clinical applications of polygenic risk scores (PRS) derived from GWAS (5), application of PRS from the majority of GWAS would systematically provide greater health improvement for European-ancestry populations, thereby exacerbating health inequalities (6).

Meta-analyzes in GWAS often fail to identify variants that differ in frequency amongst populations with PRS, either under- or overestimating risk in understudied populations (2). A study that applied a PRS for schizophrenia based on a European-ancestry sample found a 10-fold higher risk in African and African-American populations, showing this lack of portability (7). Neglecting genetic differences can also have serious consequences for drug safety. A failure to account for G6PD deficiency-causing mutations in sub-Saharan Africa led to the withdrawal of an effective antimalarial drug combination, with the drug being safe in G6PD non-enzymatically deficient individuals (2,8).

ADD SOMETHING ABOUT DIFFERENCE BETWEEN DISCOVERY AND REPLICATION

The incidence of disease also differs across populations. Cystic fibrosis, for instance has a high prevalence in Europe (1 in 2000/3000 births) compared to being rare in African Americans (1 in 17,000 births) (2). Whilst the most common causal allele in European population ΔF508 accounts for more than 70% of the cases, it is only 29% for African diaspora populations (2,9). A description of the current gaps in GWAS by disease and ancestry group is therefore also urgently needed to cover this vast uncharted territory.

The need for prioritization of greater diversity in genetic studies across a broader number of diseases has been recognized, but to date not easily monitored. To improve the monitoring of diversity and identify gaps in areas of research, we developed an on-line platform called the GWAS Diversity Monitor (gwasdiversitymonitor.com) to facilitate the visualization, analysis and monitoring of GWAS Diversity. The platform integrates almost 4,000 peer-reviewed published GWA studies across XXXX unique traits and XXXX phenotypes indexed by the EMBL-EBI GWAS Catalog. It is updated daily, providing a virtual real-time monitoring of GWAS diversity.

ADD STAT THAT CURRENTLY WE KNOW 10000% MORE ABOUT DISEASE X IN EUROPEANS COMPARED TO Africans, etc.

Users can use the GWAS Diversity Monitor to visualize all GWA studies from 2005 to now across this five figure, five widget and two-tab interactive dashboard. They can also click and hover on the graphs for additional information. In addition the GWAS Diversity Monitor is highly extendable in that it allows user to download all datasets for additional analyses.

DON’T ADD ALL OF THESE BUT SUMMARIZE SOME

 There are a total of 4131 studies in the Catalog.

 Earliest study in catalogue was PubMedID 15761122 on 2005-03-10 by Klein RJ et al.

 Most recent study in the catalogue was PubMedID 31295674 on 2019-07-05 by Condreay LD et al.

 Accession with biggest sample is PubMedID 30643256 (N=2370390) by Baselmans BML et al.

 There are a total of 7407 unique study accessions.

 There are a total of 4429 unique diseases raits studied.

 There are a total of 2974 unique EBI "Mapped Traits".

 The total number of associations found is 152652.

 The average number of associations found is 20.61.

 Mean P-Value for the strongest SNP risk allele is: 1.010E-6.

 The number of associations reaching the 5e-8 threshold: 99763.

 The journal to feature the most GWAS studies is: Nat Genet.

 Total number of different journals publishing GWAS is: 529.

 Most frequently studied (Non-European) "Diseases Trait": Type 2 diabetes.

With the continuous and rapid publication of GWAS, and a recognition of lack of diversity and gaps in knowledge about diseases within certain populations, the GWAS Diversity Monitor provides a user-friendly interface to visualize and monitor GWAS diversity by populations and diseases across time.

TEXT OUT TO ADD BACK IN

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/).

**Data availability**

The GWAS Diversity Monitor uses data provided by the NHGRI-EBI Catalog (and their licensing information can be found [here](https://www.ebi.ac.uk/about/terms-of-use)).

**Code availability**

The dashboard is written in Bokeh 1.2.0 in Python and hosted on Heroku (a PaaS).

Source code is hosted on GitHub: <https://github.com/crahal/GWASDiversityMonitor>

**References**

1. MC Mills and C Rahal, “A scientometric review of genome-wide association studies”, *Communications Biology*, **2**, (2019)
2. Sirugo, G, Williams, SM & SA Tishkoff, “The missing diversity in human genetic studies,” Cell, 177, 26-31.
3. AR Martin et al., “Human demographic history impacts genetic risk predictions across diverse populations,” *Am J Hum Genet.* **100**, 635-649, 2017.
4. Tropf et al. “Hidden heritability due to heterogeneity across seven populations,” *Nat Hum Beh*. **1** 757-765 (2017).
5. Torkamani, A. et al., “The personal and clinical utility of polygenic risk scores”, *Nat Rev Gen*. **19**, 581-590 (2018).
6. AR Martin et al., “Clinical use of current polygenic risk scores may exacerbate health disparities,” *Nat Gen.*, **51**, 584-591, 2019.
7. Curtis, D. “Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia,” *Psychiatr. Genet*. **28**, 85-89.
8. Luzzatto, L. “The rise and fall of the antimalarial Lapdap: a lesson in pharmacogenetics,” *Lancet* **376**, 739-741.
9. Stewart, C. and MS Pepper, “Cystic fibrosis in the African Diaspora,” *Ann. Am. Thorac. Soc*. **14**, 1-7.

**Acknowledgements**

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

**Author information**

The authors jointly designed the study, CR built the on-line dashboard.

**Affilations**

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

**Corresponding authors**

Correspondence to Melinda C. Mills, Nuffield College, New Road, Oxford, OX1 1NF, UK Email: melinda.mills@nuffield.ox.ac.uk

**Ethics declarations**

The authors declare no competing interests.