Improving ethnic diversity in respiratory genomics research

Martin D. Tobin[1,2] , Abril G. Izquierdo[1]

[1.] Department of Health Sciences, University of Leicester, Leicester, UK.
2. Leicester NIHR Biomedical Research Centre, Leicester, UK

Corresponding author Martin D. Tobin: martin.tobin@leicester.ac.uk

In this issue Zhu and colleagues report findings from a genomic study of lung function in the China Kadoorie Biobank (insert reference).

COPD is the most prevalent of respiratory diseases globally, with much of the current and predicted burden in low-and middle-income countries [1]. Whilst the roles of smoking, certain environmental exposures and, in around 1% of cases, alpha-1-anti-trypsin deficiency are well known, the development of disease-modifying drugs requires improved knowledge of relevant causal molecules and pathways. Genomic studies are increasingly used to inform drug development, and compounds supported by genomic evidence are more than twice as likely to be clinically developed [2]. Genomic variants mimic the effect of drugs on a particular protein or pathway, albeit over the lifespan and generally with a small effect size, such that large sample sizes are required for genome-wide association studies (GWAS). Whilst large sample sizes have been attained for GWAS of European ancestry populations, individuals of non-European ancestry are markedly under-represented, [3] with recent estimates showing that 82% of individuals studied in GWAS were of European ancestry.

Two key strategies have been adopted for identifying genomic determinants of COPD – studies of quantitative lung function, and studies of COPD cases and controls. Lung function is an independent predictor of morbidity and mortality even within the normal range and remains a key diagnostic criterion for COPD. By 2019, 279 independent genomic signals for lung function were discovered from the study of over 400,000 European ancestry participants [4, 5, 6] and 82 genomic loci through COPD case-control studies [7]. Genomic signals for lung function predicted the development of COPD – most powerfully demonstrated when combined hundreds or even thousands of genomic variants as "genetic risk scores (GRS)" or "polygenic risk scores (PRS)" [5, 8, 9] respectively. In all human genomic studies, the number of individuals of non-European ancestry studied has been very limited (**Figure 1**). Respiratory genomic studies are no exception, where the number of individuals of non-European ancestry available to study in established cohorts and case-control studies is limited, providing adequate power to show GRS and PRS associations, but with limited power to make new discoveries of individual genomic variant associations. Among resources studied to date, Burkart et al studied lung function GWAS in over 11,822 Hispanic/Latino participants [10], the COPDgene study includes 3,300 individuals of African-American ancestry used for GWAS and admixture mapping [11], and the TOPMED initiative reported new associations with lung function and COPD from whole genome sequencing of over 7,500 participants of non-European ancestry among almost 20,000 individuals studied [12].

[insert Figure 1 here]

The China Kadoorie Biobank study described in this issue represents a major step towards addressing the under-representation of studies to date. Over 100,000 individuals with genome-wide array data and lung function data were studied. Nine novel associations were reported for forced expiratory volume in 1 second (FEV_1), six for forced vital capacity (FVC) and three for FEV_1/FVC. An intronic SNP in *GPC5,* rs528366, showed an association with FEV_1 which has not been previously reported in European ancestry populations, although the allele frequencies do not differ markedly

from those in East Asian populations. This variant is associated with the levels of expression of GPC5 in plasma[13]. GPC5 is a cell surface heparan sulfate proteoglycan (glypican), a class of protein involved in cell division and growth regulation. Among other associations highlighted were those in the major histocompatibility complex (MHC) locus, including previously reported associations in *AGER* and *TNXB* [4, 14]. In European ancestry populations, signals of lung function association within the MHC were explained by classical HLA alleles, with the exception of the association with the non-synonymous coding variant, rs2070600, in *AGER*, which has been most strongly associated with FEV_1/FVC [6] and appears to alter cellular processing of the protein product, RAGE. Reduced serum levels of soluble RAGE have been described in COPD [15, 16].

Zhu and colleagues explore the relationships between lung function and obesity-related traits and their genetic determinants in different ways. First, they studied genetic variants that individually showed association with both lung function and obesity-related traits, then estimated the extent to which genome-wide associations for pairs of traits were shared. The latter estimate, genetic correlation (r_g) can theoretically range from -1 to +1 to represent complete negative or positive genetic correlation respectively. Of six trait pairs studied, the authors highlight twenty-five loci of interest including *DIS3L2* (involved in an overgrowth syndrome), *HLA-DQA1* and 12p13.2 (including *ATXN2* and *ACAD10*) showing association across different trait pairs. Negative genetic correlation was shown between FEV_1 or FVC and obesity-related trait values, which was strongest for central obesity traits and for females, though no genetic correlation was shown with FEV_1/FVC.

Second, associated genetic variants were used as tools (or "instruments") for Mendelian randomization, which aims to estimate the causal relationship between two variables without the confounding and reverse causation that can complicate interpretation in observational epidemiology [17, 18]. A higher BMI was estimated to have a causal effect on lowering FEV_1 and FVC, though not FEV_1/FVC. Finally, Zhu *et al* tested interaction between BMI and genetic influences on lung function, and a subset of findings were consistent with an effect of BMI Change on FEV_1/FVC that varies by genetic background (as captured a lung function PRS).

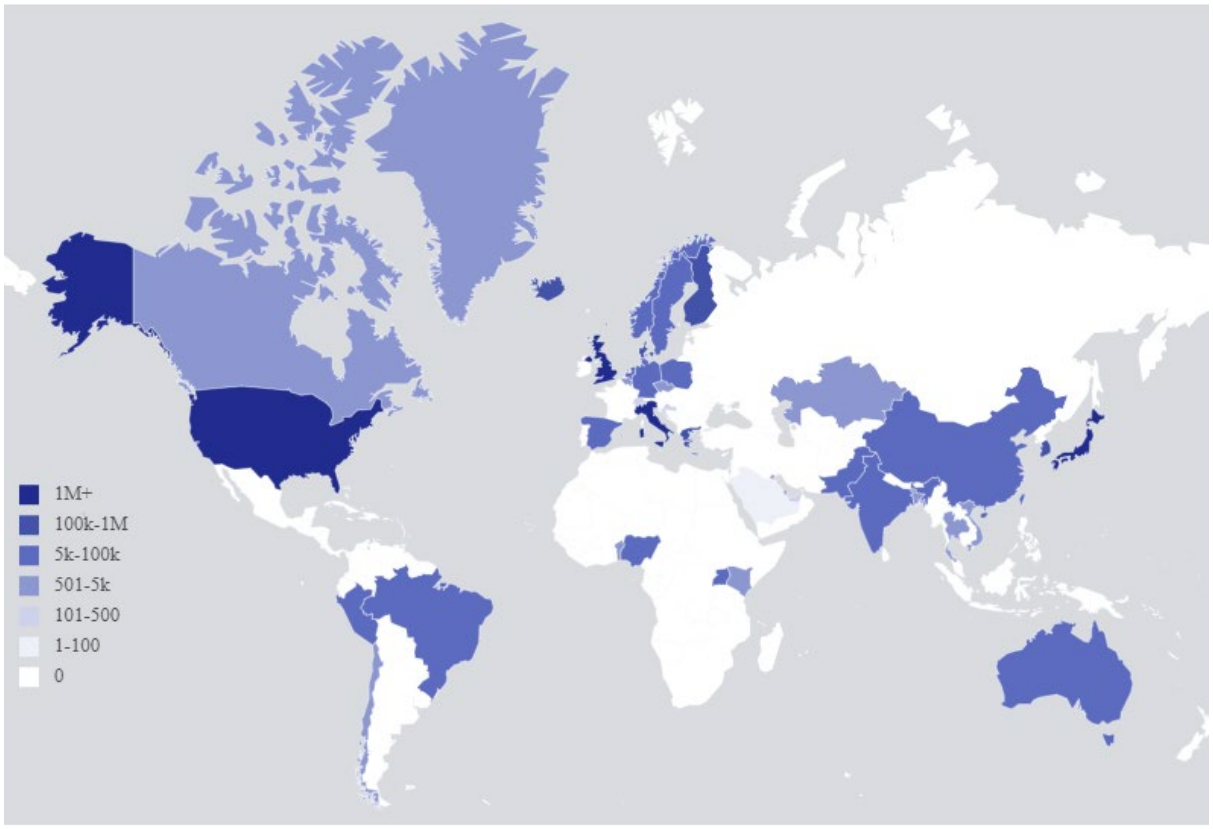
The interpretation of the inter-relationship between anthropometric and obesity-related traits, lung function and genomic variation in this study is challenging for a number of reasons, including multiple testing, comparable populations for replication studies and possible gene-environment interactions. Pathways involved in lung function and COPD susceptibility in this and other studies include those involved in growth and development, and obesity may affect lung function and also asthma through mechanical effects on ventilation and other mechanisms. Whilst an "obesity paradox" has been described in COPD[19], collider bias may account for some or all of the reported findings. In this context, these Mendelian randomization analyses add to the evidence base, showing that whilst BMI appeared causally related to lower FEV_1 and FVC, it was not causally associated with FEV_1/FVC. It would therefore seem unlikely that BMI increases susceptibility to COPD (at least in this population), but plausible that lower FEV_1 and FVC may be seen among obese patients presenting with COPD.

Overall, the findings from the China Kadoorie Biobank study are an important step forward towards informing the biology of lung function and the susceptibility to COPD. Understanding the biological relevance of these findings will be enhanced by further relating the genetic associations to genome annotations, gene expression, protein expression and other functional genomics evidence from relevant tissues and cell types. A growing number of resources now provide open access to such evidence, although these are largely based on samples of European ancestry. Comparing genetic association findings across populations of similar and differing ancestral groups is also important to

distinguish chance findings from biological association and understand which loci show heterogeneous effects across ancestries, and why [20].

Improving diversity in genomics research is needed to ensure that all populations are beneficiaries in the advances such research can bring to prediction, prevention, diagnosis and treatment. Other benefits of greater diversity include improved fine mapping of genetic signals. As not all therapeutic targets identified by genomic studies may be tractable drug targets, until the genetic architecture of relevant traits and diseases is comprehensively understood across all ethnic groups, then opportunities may be missed. Challenges in improving diversity in genomic research have included a lack of diversity in the scientific community itself, capacity-building in low-and-middle income countries, limited engagement with communities before, during and after research is conducted, and difficulty in securing funding for studies with more costly recruitment strategies [21], as well as concerns about confounding by population structure addressable through improvements in analysis methods[20, 22]. The issues have been compounded by biases in genotyping array design, the ease of accessing diverse populations for researchers in training, and the challenges of publishing findings from smaller studies or with limited replication data. We can learn from examples provided by a number of established initiatives such as H3 Africa, MalariaGen, TOPMED and new initiatives such as the Wellcome African Population Cohorts Consortium. In order to make such studies possible, the research community will need to work harder to invest in, conduct, and interpret studies across non-European ancestry populations in high-income and in low-and-middle-income countries.

(a)



(b)

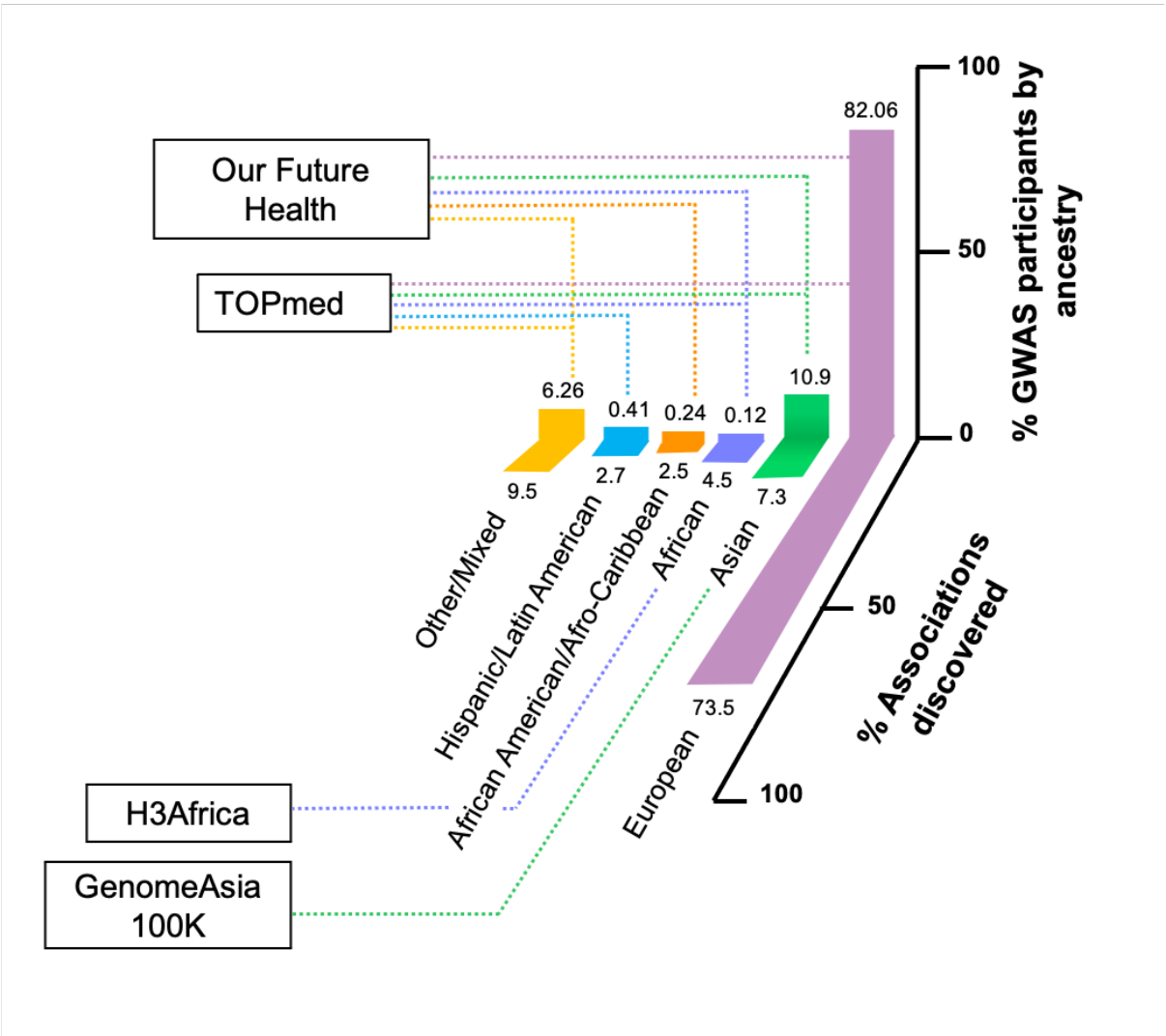


Figure 1. Participants included in genome-wide association studies (GWAS) by **(a)** Country and; **(b)** Ancestry, with examples of initiatives which could recruit substantial numbers of participants of non-European ancestry. For comparison **(b)** also shows the yield of associations discovered from GWAS of non-European ancestry. The data relate to all diseases and phenotypes and were extracted from the GWAS Diversity Monitor [23] on 7th June 2021.

Acknowledgements

M.D.T. is supported by a Wellcome Trust Investigator Award (WT202849/Z/16/Z), and an NIHR Senior Investigator Award, and is partially supported by the NIHR Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Collaborators GBDCRD. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir Med* **8**, 585-596 (2020).

2. Nelson MR*, et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856-860 (2015).

3. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26-31 (2019).

4. Repapi E*, et al.* Genome-wide association study identifies five loci associated with lung function. *Nat Genet* **42**, 36-44 (2010).

5. Shrine N*, et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet* **51**, 481-493 (2019).

6. Wain LV*, et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet* **49**, 416-425 (2017).

7. Sakornsakolpat P*, et al.* Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet* **51**, 494-505 (2019).

8. Moll M*, et al.* Relative contributions of family history and a polygenic risk score on COPD and related outcomes: COPDGene and ECLIPSE studies. *BMJ Open Respir Res* **7**, (2020).

9. Moll M*, et al.* Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. *Lancet Respir Med* **8**, 696-708 (2020).

10. Burkart KM*, et al.* A Genome-Wide Association Study in Hispanics/Latinos Identifies Novel Signals for Lung Function. The Hispanic Community Health Study/Study of Latinos. *Am J Respir Crit Care Med* **198**, 208-219 (2018).

11. Ziyatdinov A*, et al.* Mixed-model admixture mapping identifies smoking-dependent loci of lung function in African Americans. *Eur J Hum Genet* **28**, 656-668 (2020).

12. Zhao X*, et al.* Whole genome sequence analysis of pulmonary function and COPD in 19,996 multi-ethnic participants. *Nat Commun* **11**, 5182 (2020).

13. Sun BB*, et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).

14. Hancock DB*, et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet* **42**, 45-52 (2010).

15. Miller S*, et al.* The Ser82 RAGE Variant Affects Lung Function and Serum RAGE in Smokers and sRAGE Production In Vitro. *PLoS One* **11**, e0164041 (2016).

16. Pratte KA*, et al.* Soluble receptor for advanced glycation end products (sRAGE) as a biomarker of COPD. *Respir Res* **22**, 127 (2021).

17. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22 (2003).

18. Tobin MD, Minelli C, Burton PR, Thompson JR. Commentary: development of Mendelian randomization: from hypothesis test to 'Mendelian deconfounding'. *Int J Epidemiol* **33**, 26-29 (2004).

19. Iyer AS, Dransfield MT. The "Obesity Paradox" in Chronic Obstructive Pulmonary Disease: Can It Be Resolved? *Ann Am Thorac Soc* **15**, 158-159 (2018).

20. Morris AP. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* **35**, 809-822 (2011).

21. Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? *J Community Genet* **8**, 255-266 (2017).

22. Gautam Y, Ghandikota S, Chen S, Mersha TB. PAMAM: Power analysis in multiancestry admixture mapping. *Genet Epidemiol* **43**, 831-843 (2019).

23. Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet* **52**, 242-243 (2020).