




A genome-wide association study identifies distinct variants associated with pulmonary function among European and African ancestries from the UK Biobank

Musalula Sinkala ^{1,3}, Samar S. M. Elsheikh^{2,3}, Mamana Mbiyavanga¹, Joshua Cullinan¹ & Nicola J. Mulder ¹

Pulmonary function is an indicator of well-being, and pulmonary pathologies are the third major cause of death worldwide. We analysed the UK Biobank genome-wide association summary statistics of pulmonary function for Europeans and individuals of recent African descent to identify variants associated with the trait in the two ancestries. Here, we show 627 variants in Europeans and 3 in Africans associated with three pulmonary function parameters. In addition to the 110 variants in Europeans previously reported to be associated with phenotypes related to pulmonary function, we identify 279 novel loci, including an *ISX* intergenic variant rs369476290 on chromosome 22 in Africans. Remarkably, we find no shared variants among Africans and Europeans. Furthermore, enrichment analyses of variants separately for each ancestry background reveal significant enrichment for terms related to pulmonary phenotypes in Europeans but not Africans. Further analysis of studies of pulmonary phenotypes reveals that individuals of European background are disproportionately overrepresented in datasets compared to Africans, with the gap widening over the past five years. Our findings extend our understanding of the different variants that modify the pulmonary function in Africans and Europeans, a promising finding for future GWASs and medical studies.

¹Computational Biology Division, Faculty of Health Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Anzio Rd, Observatory 7925 Cape Town, South Africa. ²Pharmacogenetics Research Clinic, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada. ³These authors contributed equally: Musalula Sinkala, Samar S. M. Elsheikh. ✉email: musalula.sinkala@uct.ac.za

Pulmonary function measures using the spirometer are indicators of respiratory health and predict morbidity and mortality^{1,2}. However, these parameters, which include the force expiratory volume in 1-second (FEV₁), forced vital capacity (FVC), and peak expiratory capacity (PEF), vary significantly among populations of different ancestry backgrounds³ and show strong evidence of genetic and environmental influences^{1,4}.

During the last decade, large-scale genome-wide association studies (GWASs) have used various pulmonary parameters to evaluate the genomic loci associated with pulmonary function and related traits that have yielded hundreds of associated variants^{5–10}. These and other studies indicate that genomic loci associated with pulmonary function overlap with chronic obstructive pulmonary disease, asthma, pulmonary fibrosis, lung cancer, and other pulmonary phenotypes^{2,8–10}. For example, a recent GWAS based on the UK Biobank cohort ($N = 50,008$), including heavy smokers and never smokers, identified six loci associated with low FEV₁¹⁰. Another study of individuals ($N = 48,943$) sampled from the extremes of pulmonary function distribution in the UK Biobank identified 95 variants strongly associated with chronic obstructive pulmonary disease susceptibility⁸. Importantly, these previous studies have applied the analyses to a selected population group of primarily European ancestry.

The UK Biobank cohort contains data on 389,449 individuals, providing an opportunity to use GWAS approaches to identify variants associated with pulmonary function among individuals of European and recent African descent by allowing large-scale comparisons of lung function parameters¹¹. Furthermore, by integrating the genetic association of FEV₁, PEF, and FVC, a list of shared loci that collectively modify pulmonary function could be identified. We hypothesise that different genetic variants are associated with pulmonary function in Africans. Thus, their identification will provide additional information relevant to understanding pulmonary function in physiology and disease in district populations. However, to our knowledge, no GWAS study

has been performed to compare the SNPs associated with the full range of FEV₁, FVC, and PEF parameters across the entire UK Biobank cohort and separately among Africans and Europeans.

Here, we compare variations in pulmonary function parameters among individuals of African and European ancestry represented in the UK biobank. First, we used the genome-wide associated summary statistics for three UK Biobank-defined continuous pulmonary function parameters: FEV₁, FVC, and PEF. Then, we conducted further analyses to identify genes, regions, and gene sets associated with each pulmonary phenotype. Furthermore, we evaluate the candidate phenotype variants in relation to published GWAS results. Overall, this approach allows us to report credible loci associated with pulmonary function among Africans and Europeans, which were enriched across many plausible genes and gene sets involved in pulmonary function or related phenotypes.

Results

UK Biobank pulmonary function demographics. There were 389,449 participants, comprising Europeans ($N = 383,471$) and Africans ($N = 5978$). The average participant age at recruitment was 56.8 years (standard deviation = 8.0 years) for Europeans and 51 years (7.9) for Africans, respectively. This difference was statistically significant (Welch test: $t = -41.3$, $p = 9.07 \times 10^{-300}$) (see Supplementary Fig. 1).

Lung function parameters vary between individuals of European and African ancestry. We assessed the mean FVC, FEV₁, and PEF between Europeans ($N = 383,471$) and Africans ($N = 5978$) represented in the UK Biobank datasets. We found that the mean FVC was significantly higher in the Europeans (mean = 3.73 L) compared to the Africans (mean = 2.95 L), (Welch test: $t = 48.35$, $p < 1 \times 10^{-320}$; Fig. 1a). Furthermore, we found that the FEV₁ and the PEF were both significantly higher in Europeans (mean FEV₁ = 2.82 L, mean PEF = 389.6 L/min)

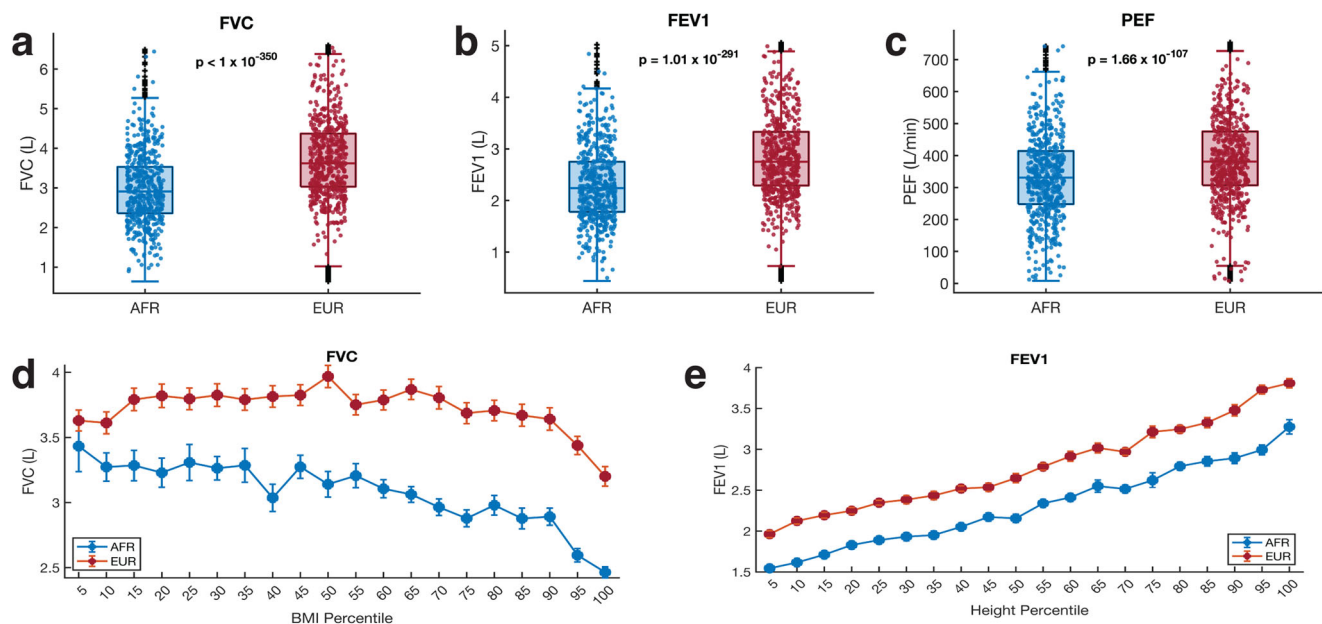


Fig. 1 Comparison of the pulmonary function parameter among Africans and Europeans. The boxplots indicate the distribution of (a) FEV₁, (b) FVC, and (c) PEF in Europeans ($n = 383,471$) and Africans ($n = 5978$). The p -values shown for each comparison were calculated from Welch's t -test. On each box, the central mark indicates the median, and the left and right edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. To make the visualisation clearer, the filled circle mark showing the distribution only includes 1000 randomly sampled points from the total sample size of each group. Error bars showing the variation in (d) FVC and (e) FEV₁ across BMI percentiles and height, respectively, among Africans and Europeans. The middle point indicates the mean FVC or FEV₁/FVC, and the error bars indicate the standard error of the mean at the BMI percentile.

than those measured in the Africans (mean FEV1 = 2.28 L, mean PEF = 332.7 L/min), FEV1; $t = 42.60$, $p = 1.0 \times 10^{-291}$ (Fig. 1b) and PEF; $t = 24.06$, $p = 1.7 \times 10^{-107}$; Fig. 1c. According to a recent systematic review, “Whites” have higher pulmonary function parameters than other ethnic groups (including Africans)¹². About 50% of these articles cited inherent factors and anthropometric differences to explain the observed differences. However, similar to other studies^{13–15}, our findings show that these variations in pulmonary function measures exist across various ages, heights, and BMI percentiles (Fig. 1 and Supplementary Fig. 2a–i). However, using a generalised linear model, we found that the observed higher FVC, FEV1, and PEF in Europeans compared to Africans is not due to the age difference between the two groups, even though the FVC ($t = -19.26$, $p = 1.0 \times 10^{-82}$), FEV1 ($t = -16.68$, $p = 1.88 \times 10^{-62}$), and PEF ($t = -11.91$, $p = 1.01 \times 10^{-32}$), tend to reduce with age (see Supplementary Table 1 and Supplementary Note 1). Recently, a lack of knowledge among healthcare workers concerning variations in pulmonary function measures among ethnic groups has been suggested to impact the assessment of minority patients’ recovery from COVID-19¹⁵. However, no studies have identified major genetic variants that vary by ethnic groups that can explain the disparities in lung function^{15,16}.

Previous studies show that the FVC, FEV1, and PEF vary with age, body mass index (BMI), and height of individuals^{17–20}. Here, we also found that FVC, FEV1, and PEF tend to reduce with age, an increase in BMI is observed at the 50 percentile, and all three parameters increase along with the height of the individuals (Fig. 1d, e, and Supplementary Fig. 2a, i). However, unlike age and height, we found that the relationship between pulmonary function parameters and BMI appears to be associated with overweight/obesity, with a threshold effect and not a simple linear relationship (see Supplementary Note 1). Furthermore, we observed that the FEV1/FVC levels are conversely higher in Africans than Europeans across the BMI percentiles (Supplementary Fig. 2f).

Genetic variant associated with FVC, FEV1 and PEF among Europeans and Africans. Since the FVC, FEV1, and PEF values were significantly higher in Europeans than in Africans, we presumed that a genome-wide association analysis would identify the genetic variants associated with each of these pulmonary function parameters in each group. Therefore, we collected the GWAS summary statistics for each pulmonary function parameter within each ethnic group (see the “Methods” section). In these data, we discovered 1 variant in Africans and 67,855 variants in Europeans that were associated (GWA p -values 5×10^{-8}) with FEV1, 6 in Africans and 79,132 in Europeans that were associated with FVC, and zero (0) in Africans and 26,432 in Europeans that were associated with PEF (Supplementary Fig. 3a–c). The total number of significant variants discovered for each pulmonary function parameter, including those in substantial linkage disequilibrium ($R^2 > 0.4$), and the intersection of these variants are displayed in Supplementary Fig. 3a–f.

We applied fine mapping²¹ to identify 310 (credible set) causal variants significantly associated (p -values $< 5 \times 10^{-8}$ and causal probability > 0.1 ; see “Methods” section) with FVC in Europeans and 2 significant associations in Africans (Fig. 2a, b). For FEV1, we found 308 significant causal variant associations in Europeans and 1 in Africans (Fig. 2c, d). Furthermore, for PEF, we identified 374 significant causal variant associations in Europeans and none (0) in Africans (Fig. 2e, f). Overall, we identified 820 unique credible SNPs associated with the three pulmonary functions. Surprisingly, the significant SNPs associated with FVC, FEV1, and PEF were unique to each ancestral group (Fig. 2g–i and Supplementary Data 1).

Next, we evaluated the independent SNPs associated with the three pulmonary function parameters while considering the population’s linkage disequilibrium structure (see the “Methods” Section). Here, we identified 630 independent SNPs from the 820 credible sets of causal SNPs associated with all three pulmonary function parameters. Finally, we compared the 627 independent SNPs in Europeans with the 3 SNPs in Africans significantly associated with the three pulmonary function parameters and found no common variants between the two sets. Conversely, we found that 164 SNPs were associated with FVC and FEV1 in Europeans (Fig. 2k). However, there was no overlap in the associated SNPs among Africans (Fig. 2j). Finally, it should be noted that smoking impacts pulmonary function, but the effect of smoking was not accounted for in the GWA analyses. Therefore, this is probably a limitation of our findings.

Since the SNPs significantly associated with pulmonary function were unique for Europeans and Africans, we next relaxed the GWAS significance threshold to a suggestive cut-off p -value²² of 1×10^{-6} . Then, we compared the significant SNPs in Europeans and Africans for FVC, FEV1, and PEF. For all three pulmonary function metrics, even when using a less strict significance criterion, we were unable to discover any shared SNPs between Africans and Europeans (Supplementary Fig. 3d–f). Furthermore, we found that the most statistically significant SNPs in Africans had relatively larger beta estimates in Africans than Europeans for the FVC, FEV1, and PEF (see Supplementary Fig. 4 and Supplementary Notes 2). In addition, we have provided an interactive online visualisation that allows the user to evaluate the significance of SNPs in each group using an arbitrary significance threshold and compare the SNPs on different chromosomes, linkage disequilibrium loci, and genes, for FVC (Supplementary Figs. 5, 6), FEV1, and PEF (see the Supplementary Notes: Comparison of variants associated with pulmonary function).

We compared the minor allele frequency of SNPs in the UK Biobank between Europeans and Africans for the combined 820 SNPs (817 in Europeans plus 3 in Africans) associated with pulmonary function. We found that 788 out of 820 SNPs differed significantly in frequency between Africans and Europeans (Supplementary Data 2). The top-three variants that exhibited the most significantly higher frequencies in Europeans compared to Africans were rs2042395 (frequency in Europeans = 0.77, in Africans = 0.19, Fisher test p -value = 4.94×10^{-323}), rs3748400 (Europeans = 0.78, Africans = 0.19, $p = 6.92 \times 10^{-323}$), rs8045843 (Europeans = 0.78, Africans = 0.17, $p = 8.89 \times 10^{-323}$), see Supplementary Data 2 and Supplementary Fig. 2g. Interestingly, the variants rs2042395 and rs8045843 have been previously associated with the “well-being spectrum”²³ and “sensitivity to environmental stress and adversity”²⁴, respectively, in individuals of European ancestry. Conversely, the top variants with higher frequency in Africans compared to Europeans were rs143384 (Europeans = 0.40 and Africans = 0.92, $p = 2.0 \times 10^{-323}$), rs3133084 (Europeans = 0.23 and Africans = 0.65, $p = 8.4 \times 10^{-323}$), and rs7853063 (Europeans = 0.20 and Africans = 0.60, $p = 6.4 \times 10^{-323}$), see Supplementary Fig. 2g. Among these, the variant rs143384 has been reported to be associated with FVC, lung function, and PEF²⁵, and among anthropometric traits in Europeans²⁶.

Altogether, these analyses revealed that different SNPs may be associated with FVC, FEV1, and PEF among Europeans and Africans and that the frequency of these SNPs significantly varies between these populations.

Pathway and GWAS catalog enrichments of the SNPs. We assessed the enrichment of GWAS Catalog²⁷ annotation terms for the genes containing SNPs associated with lung function

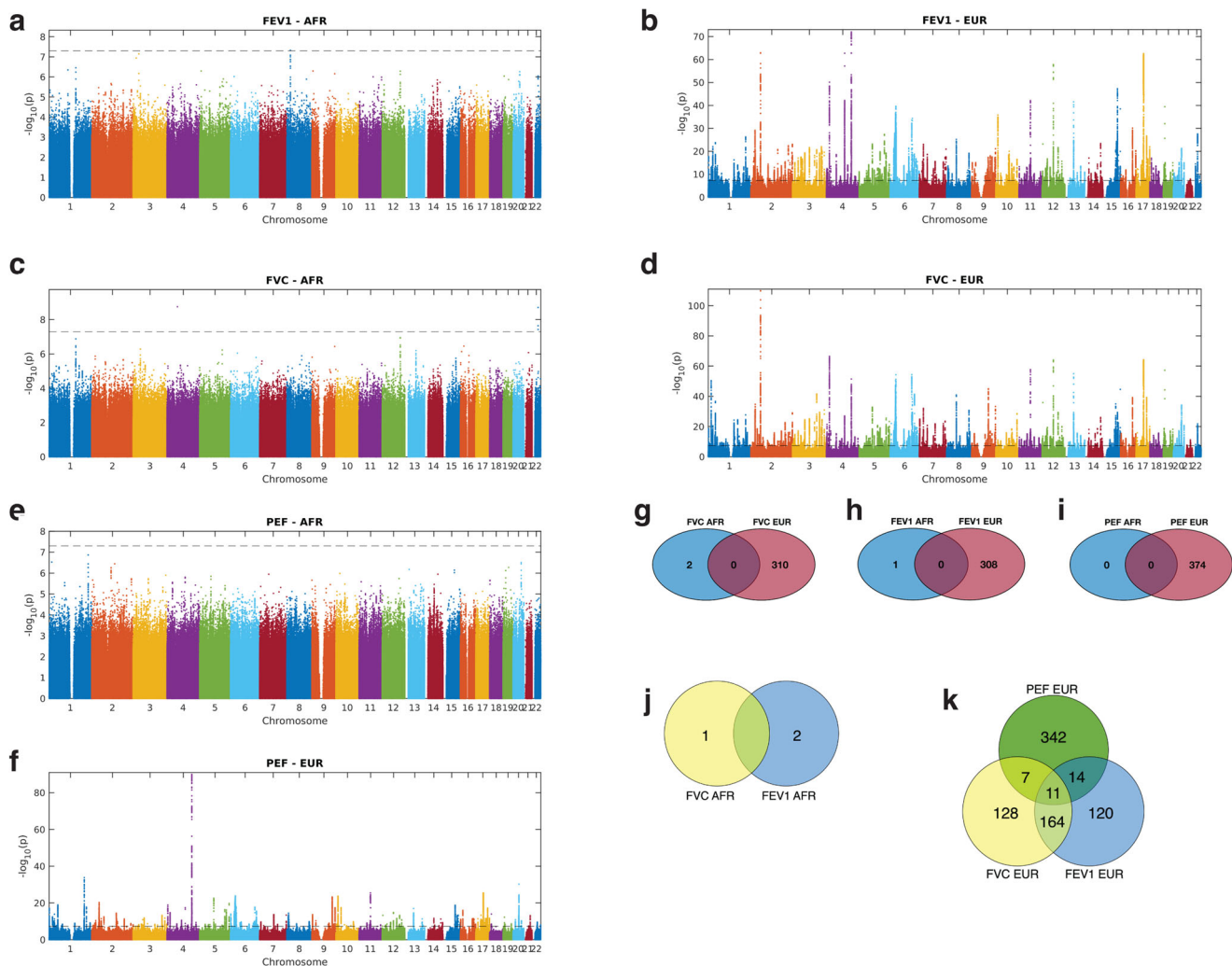


Fig. 2 Manhattan plots and Venn diagrams of the SNP associations. The Manhattan plots include those of the SNPs associated with (a) FEV1 in Africans and (b) FEV1 in Europeans, (c) FVC in Africans and (d) FVC in Europeans, and (e) PEF in Africans and (f) PEF in Europeans for each chromosome. The Venn diagrams show the overlap among the significant causal SNPs associated with (g), FVC (h), PEF, and (i) FEV1 in Africans and Europeans. The distribution of genetic variants associated with three pulmonary function parameters among the (j) Africans and (k) Europeans. Refer to Supplementary Data 1 for details concerning individual SNPs and their frequencies among Africans and Europeans.

(suggestive cut-off p -value of 1×10^{-6}) in each study population (see Supplementary Data 2).

The GWAS Catalog term analyses revealed that in Europeans, the genes were significantly enriched for GWAS terms associated with “Height” (hypergeometric test; $p = 1.06 \times 10^{-93}$), “Lung function (FEV1)” ($p = 5.4 \times 10^{-25}$), “Pulmonary function interaction” ($p = 2.33 \times 10^{-19}$) among others (Fig. 3a and Supplementary Data 3). In Africans, we found that the genes were significantly enriched for GWAS terms associated with “Subcutaneous adipose tissue” ($p = 1.2 \times 10^{-07}$), “Birth weight” ($p = 3.7 \times 10^{-04}$), “Cognitive decline rate in late mild cognitive impairment” ($p = 7.3 \times 10^{-04}$), among others (Fig. 3b and Supplementary Data 3). Overall, these results show that the SNPs identified among Europeans are in genes known to play roles in many phenotypes, most notably those related to pulmonary function or GWAS phenotypes related to pulmonary function. Conversely, the SNPs we identified associated with pulmonary function among Africans fall within genes that are not enriched for pulmonary function-related terms.

Variant spanning loci associated with pulmonary function among Europeans and Africans. Many of the associated SNPs

may simply reflect the linkage disequilibrium structure of the populations^{28,29} (see Supplementary Data 4). For example, we found 10 variants associated with FEV1 and FVC in Europeans within loci 12q14.3, and upon fine mapping²¹, we found that the most likely causal SNP within the loci was rs1351394 (Probabilistic Identification of Causal SNPs²¹, causal probability value = 0.7243), a 3-prime untranslated region variant located in the gene *HMG2* (Fig. 4a). The variant rs1351394 has previously been associated with variations that affect FEV1 capacity, including height^{30,31} and birth length³². Furthermore, *HMG2* is involved in lung development³³.

At locus 19q13.42, we found that the most likely causal SNP is rs147110934 (causal probability = 0.83), associated with FEV1 and FVC in Europeans (Fig. 4b, also see Supplementary Data 4). rs147110934 is a predicted missense variant that falls within the *ZNF628* gene. In addition, whilst rs147110934 has not been previously associated with pulmonary function, we found it is associated with height³⁴ and body weight^{35,36}, both of which are associated with FVC and FEV1.

Furthermore, we found several SNPs in the loci 9q22.32 associated with pulmonary function (Fig. 4c). Here, the lead and predicted causal (causal probability = 1) variant is rs16909898,

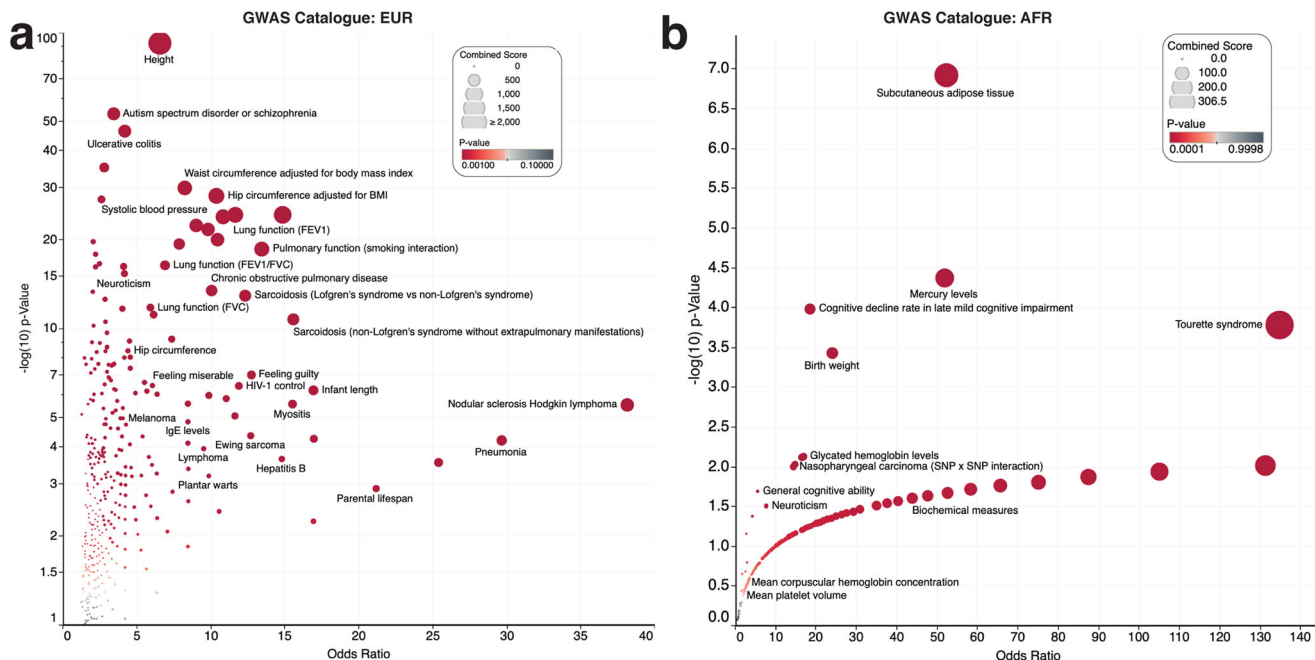


Fig. 3 GWAS catalog enrichment analysis plots. Volcano plots of the GWAS Catalog enrichment analysis for genes where significant SNPs are located for (a) Europeans and (b) Africans. All four plots show the adjusted p -value on the y-axis and the odds ratio of the enrichment score on the x-axis. Each circle represents a GWAS Catalog term or Elsevier pathway. The circles are coloured based on the levels of statistical significance, with the redder colours showing a greater degree of significance. Each circle is sized based on the combined enrichment score of the term represented by the circle.

located in the *PTCH1* gene previously identified to modify pulmonary function parameters^{37,38} and height³¹.

In addition, for individuals of African ancestry, at the locus 5q32, the lead SNP among the four associated with pulmonary function was rs369476290 (causal probability = 0.67), an intergenic variant located near the gene *ISX*. rs369476290 has not been previously linked to pulmonary function or disease (Fig. 4d).

Since the SNPs significantly associated with pulmonary function were unique for Europeans and Africans, we next set to compare the estimated beta values for all SNPs with a GWA significance of <0.05 . Here, we found that the most statistically significant SNPs in Africans had relatively larger beta estimates in Africans than Europeans for the FCV, FEV1, and PEF (Supplementary Fig. 4). Overall, this finding showed that the SNPs significantly associated with pulmonary function in Africans demonstrated larger effect sizes than in Europeans. Conversely, we found thousands of variants associated with pulmonary function in Europeans that tended toward statistical significance in Africans (see Supplementary Notes: Comparison of variants associated with pulmonary function).

Furthermore, we aimed to replicate the causal variants associated ($p < 5 \times 10^{-8}$) with pulmonary function in Europeans in Africans at a p -value of less than 0.05. Interestingly, we found 56 independent variants that could be associated with pulmonary function in both Europeans and Africans (see Supplementary Note 3). These include, among others, the loci near the gene *MECOM*, where the causal SNP rs11709963 was associated with FEV1 (p -value = 5.3×10^{-19}) in Europeans. There was some evidence for an association within the region for Africans (rs1362771, $r^2 = 0.51$ the causal SNP rs11709963 in Europeans) was associated with FVC (replication $p = 0.02$), see Supplementary Data 4 and Supplementary Fig. 7. Furthermore, a *SATB2* variant, rs77064030 (p -value in Europeans = 6.7×10^{-11}) that is in linkage disequilibrium with rs78696503 ($r^2 = 0.8$), associated with FEV in Africans (replication p -value in Africans = 0.007), see Supplementary Fig. 8. Among variants

associated with PEF, is the *FAM132A* variants rs79361800 (p -values; Europeans = 9.20×10^{-10} and Africans = 1.02×10^{-5}), see Supplementary Fig. 9.

Therefore, we suggest that our findings may be due to both the difference in the sample size (which is associated with the statistical power to identify the causal variants) and the existence of different variants associated with pulmonary function among European and African individuals.

Comparison to variants previously associated with pulmonary function. Next, we aimed to identify the previously described and novel SNPs among the significant SNPs that were also predicted to be causal within a particular linkage disequilibrium block (see the “Methods” section). Here, we grouped the SNPs into four ordinal categories based on confidence: (1) SNPs reported to be associated with pulmonary function, (2) SNPs related to phenotypes correlated to pulmonary function (e.g., height, see Supplementary Fig. 1), (3) SNPs that fall within genes reported to be associated with pulmonary function and/or disease, (4) SNPs that are expression quantitative trait loci (eQTLs) in the lung, and (5) the novel SNPs.

Interestingly, we found that among our list, 97 variants in Europeans and none (0) in Africans have been previously associated with pulmonary function (see Table 1 and Supplementary Data 4). These include variants in the genes *PLEKHM1*, *HMG2A*, *KDM2A*, and *SYTL2* (Table 2). Likewise, we found that 69 variants in Europeans, and none (0) of the variants in African ancestry individuals had previously been associated with a phenotype correlated to pulmonary function (see Supplementary Data 4). Furthermore, we found that 178 variants in Europeans and 0 variants in Africans are located within genes associated with various pulmonary function phenotypes and diseases, and 4 variants in Europeans and none in Africans are significant eQTLs in the lungs. These four variants affect the expression of *CAMLG*, *PHF15*, *RNF40*, and *MLLT6*. Finally, we found 206 novel variants in Europeans and 3 in Africans associated with pulmonary

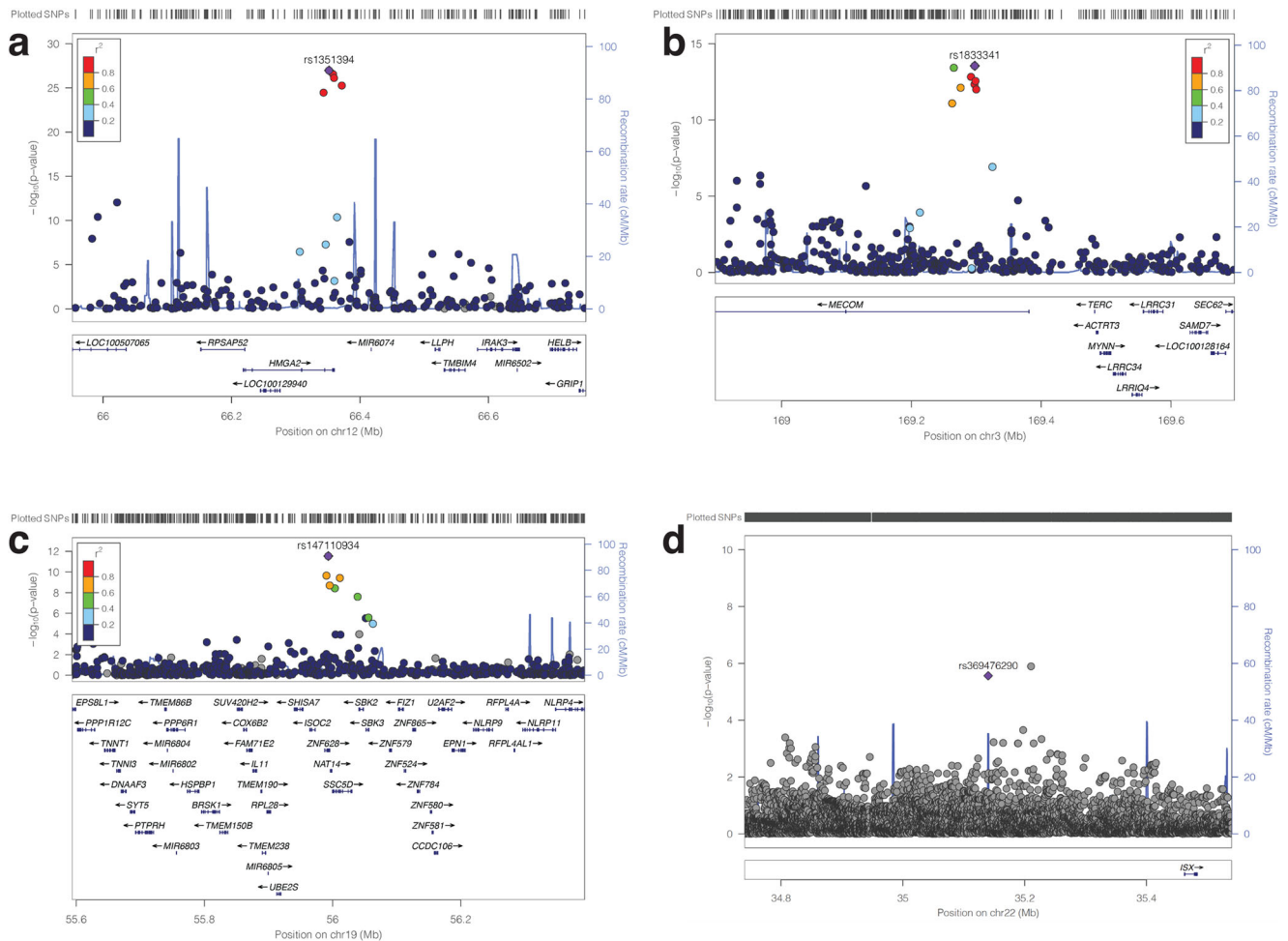


Fig. 4 Regional association plots for genome-wide significant pulmonary function. These include the loci for the lead SNPs (a) rs1351394 at loci 12q14.3, (b) rs16909898 at 9q22.32, (c) rs147110934 at loci 19q13.42, and (d) rs369476290 on chromosome 22. The genes within the chromosomal loci are shown in the lower panel. The blue line indicates the recombination rate. The filled circles show the position of the SNPs along the region on the x-axis and the negative logarithm of the association *p*-value on the y-axis. The lead SNP is shown in purple, and the SNPs within the locus are coloured based on the linkage disequilibrium correlation value (*r*²) with the lead SNP based on the European HapMap haplotype (in panels a, b, and c) and African HapMap haplotype (panel d) from the 1000 genome project.

Table 1 Known and novel variants associated with pulmonary function.

Ethnicity	Pulmonary Function	Pulmonary Function Associated Traits	Lung Disease Associated	eQTLs*	Novel
Africans	0	0	0	0	3
Europeans	97	69	178	4	279

*Expression quantitative trait loci.

function; see Supplementary Data 4 for the complete list of significant variants and the studies reporting the known variants. Among the novel discoveries, in Europeans, 104, 101 and 136 were associated with FVC, FEV1, and PEF, respectively, whereas in Africans, 2, 1 and 0 were associated with FVC, FEV1, and PEF, respectively.

We focused on the genes in which the novel SNPs associated with pulmonary function among Europeans were located to perform enrichment analyses based on the Disease Gene Network database³⁹, and the Phenotype and Genotype Integrator database⁴⁰. Here, our Disease Gene Network analysis revealed that the novel genes are enriched for terms related to pulmonary function, including “Forced expiratory volume function” ($p = 9.7 \times 10^{-13}$) and body measures that modify pulmonary function, including “Body Height”

($p = 1.33 \times 10^{-15}$), see Supplementary Fig. 10. Similarly, our phenotype and genotype integrator enrichment analysis revealed that the genes are enriched for pulmonary function-related terms, including Forced Expiratory Volume ($p = 2 \times 10^{-4}$) and phenotypes associated with pulmonary function, including Body Height ($p = 4.2 \times 10^{-07}$), see Supplementary Fig. 10. These findings show that despite the SNPs being novel among Europeans, the genes within which the SNPs are located are known to be associated with pulmonary function.

Bias in GWAS studies explains why few SNPs were previously associated with pulmonary function in Africans. Since none of the SNPs we identified as being associated with pulmonary

Table 2 Top significant variants associated with pulmonary function.

Variant	Nearest Genes	Ethnicity	Measure	Chrom:Pos	GWAS p	Evidence
rs536516159	<i>LZTS1</i>	AFR	FEV1	8: 20598779	4.8×10^{-08}	Novel
rs8756	<i>HMGA2, AC090673.2</i>	EUR	FEV1	12: 66359752	8.4×10^{-58}	Pulmonary Function
rs55663797	<i>PLEKHM1</i>	EUR	FEV1	17: 43544379	7.3×10^{-52}	Lung Disease Assoc.
rs1828591	<i>HHIP</i>	EUR	FEV1	4:145480780	8.3×10^{-52}	Pulmonary Function
rs571481915	<i>LPHN3</i>	AFR	FVC	4: 61183218	1.8×10^{-09}	Novel
rs369476290	<i>ISX</i>	AFR	FVC	22: 35140134	2.0×10^{-09}	Novel
rs8756	<i>HMGA2</i>	EUR	FVC	12: 66359752	1.2×10^{-63}	Pulmonary Function
rs2696624	<i>KANSL1</i>	EUR	FVC	17: 44326845	1.8×10^{-62}	Novel
rs7952436	<i>KDM2A, ADRBK1</i>	EUR	FVC	11: 67024534	9.3×10^{-58}	Pulmonary Assoc.
rs6829956	<i>HHIP</i>	EUR	PEF	4:145440288	1.1×10^{-90}	Pulmonary Function
rs1342062	<i>SLC26A9</i>	EUR	PEF	1:205912786	1.6×10^{-34}	Lung Disease Assoc.
rs143384	<i>GDF5</i>	EUR	PEF	20: 34025756	5.8×10^{-31}	Pulmonary Function

function among Africans has been reported in the literature, we queried the GWAS Catalog²⁷ for previous studies of pulmonary function or phenotypes related to pulmonary function (such as asthma) across various ancestry backgrounds. We found those studies to be significantly biased toward individuals of European ancestry (Fig. 5a). Also, despite the number of studies conducted on individuals of African ancestry increasing over the last five years, the gap is widening between the number of studies reported on Europeans compared to Africans during the same time interval (Fig. 5a). Overall, among the 235 GWAS studies reported on pulmonary function or phenotypes related to pulmonary function, only eight were conducted on Africans or African Americans. In comparison, we found that 120 studies have been conducted exclusively on individuals of European ancestry (Fig. 5b). Furthermore, in the same studies, the cumulative sample size of the Europeans in 2021 (10,633,660 individuals) is approximately 235 times greater than that of the Africans (45,189 individuals; see Fig. 5c).

Discussion

We analysed variations in pulmonary function and the associated genetic variants among individuals of African and European ancestry in the UK Biobank. Here, we report differences in FEV1, FVC, and PEF parameters among Africans and Europeans. Previous studies have examined the pulmonary function parameters between Africans and Europeans, with most reporting the differences we observed^{3,41–43}. However, there has been no explanation for the genetic basis of these observed differences.

Here, we showed that the SNPs associated with pulmonary function differed between Europeans and Africans. Others have reported that the genetic variants associated with various phenotypes may differ among individuals of different ancestry^{44–47}. For example, we found that the SNPs associated ($p < 5 \times 10^{-8}$) with pulmonary function in African individuals were non-significant in Europeans, even at a p-value cut-off threshold of 0.05 (see Supplementary Note 2). Our findings confirmed that different variants might be associated with pulmonary function among Africans and Europeans. Despite this observed difference between the two ancestral groups, we are also cognizant that the number of individuals of African ancestry represented in the UK Biobank is much lower than that of Europeans. To some extent, the smaller calculated beta estimates with larger standard errors in the African group compared to the European group are explained by the difference in the sample size (see the interactive plot here). Therefore, the smaller sample size of Africans may have resulted in us missing some common associations among the groups^{48,49}. It would be interesting to evaluate our findings based on a larger sample of individuals of African ancestry.

Given that the frequency of SNPs, primarily those we found associated with pulmonary function, varies between Africans and Europeans, it is apparent why different variants are associated with these traits⁴⁸. For example, we found that rs12925700 is approximately 21 times more frequent, and rs11205303 is 14 times more frequent in Europeans than Africans, and both SNPs are reported elsewhere^{50,51} and here as being associated with pulmonary function in Europeans. Furthermore, the frequency of genetic variants among individuals of a particular ancestry affects the penetrance of disease and phenotype associated with the alternate alleles^{48,52–55}. For example, non-alcoholic fatty liver disease⁵⁶, serum uric acid levels⁵⁷, white blood cell count⁵⁸, fatty acid desaturases⁵⁹, and other phenotypes^{60–62} are associated with different alleles among Africans and Europeans. These alleles are sometimes located on the same gene, but their frequencies vary between ancestral groups.

Our enrichment analyses demonstrated a link between the significant SNPs and GWAS Catalog terms associated with pulmonary function in Europeans, with several results showing plausible biological mechanisms. Whereas it was apparent that the significantly enriched terms in Europeans were mainly associated with pulmonary function and related phenotypes (Fig. 3), we found that the top-ranking terms among SNPs in Africans are not related to pulmonary function. This finding exemplifies the bias in previous GWAS studies that have not picked up genes associated with pulmonary function in Africans. We believe that more GWAS on larger groups of Africans than those presented here are needed to identify the variants that modify pulmonary function and other traits.

We also showed that genetic association studies of pulmonary function, pulmonary physiology, and pathology are significantly biased toward individuals of European ancestry. Even in cases where individuals of African ancestry are included in the studies or studied separately, the number of participants is lower than that of individuals of European ancestry. Furthermore, the trend shows that this gap has widened vis-à-vis how Africans and Europeans are studied over the last few years (see Fig. 5).

In summary, we have revealed the extent of variations between Africans and Europeans in the pulmonary function parameters: FEV1, FVC, and PEF. In addition, we have identified the different genetic variants associated with pulmonary function among individuals of African and European ancestry. Our integrative analysis of the causal genetic variants, together with the GWAS phenotypes and diseases associated with the genes in which the variants fall, indicates that the significant SNPs are associated with pulmonary function and related phenotypes in Europeans. Therefore, more genetic association studies focusing on people of African ancestry are evidently needed to identify and validate additional causal variants for these traits and other diseases.

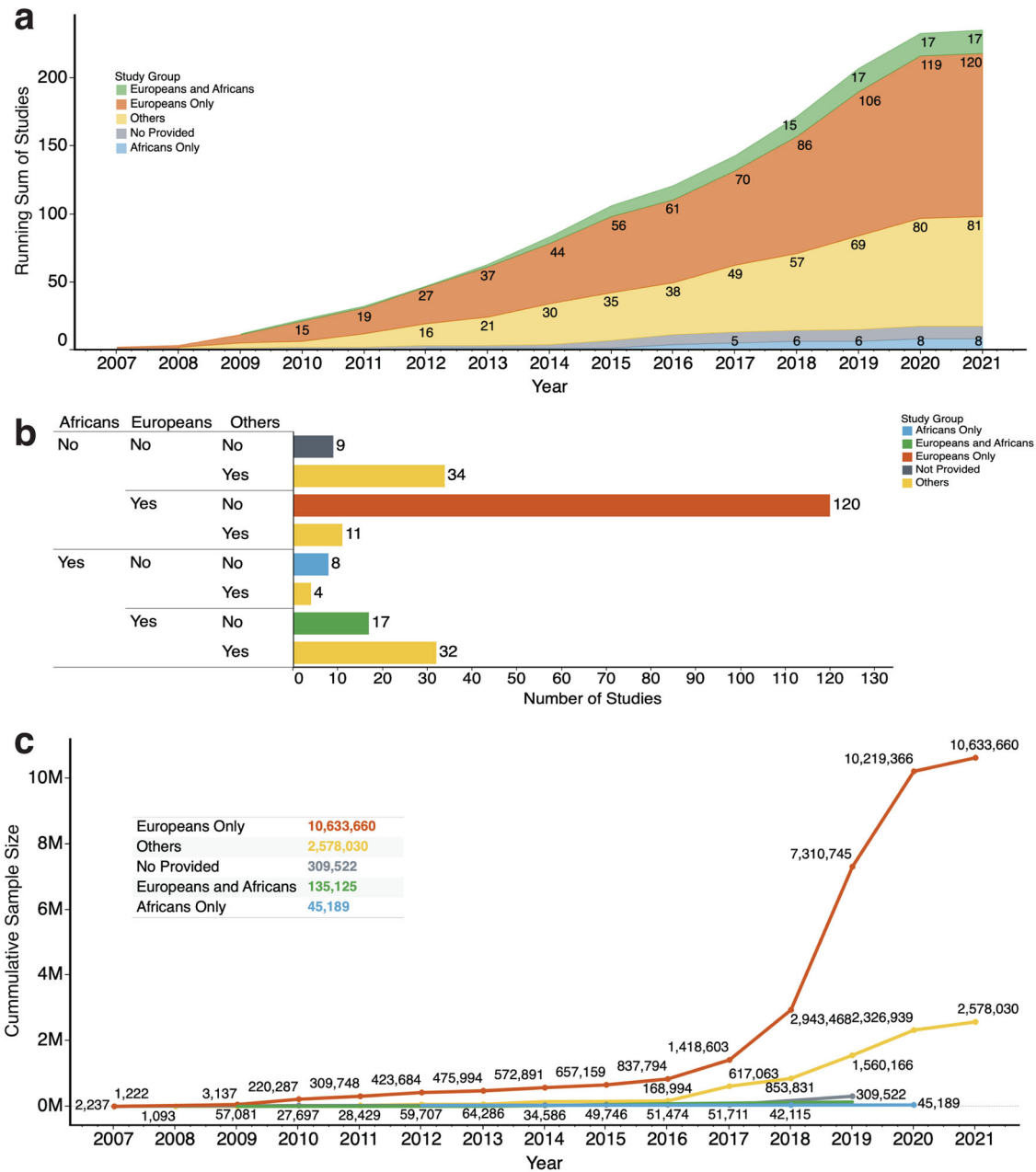


Fig. 5 The GWAS catalog of pulmonary function and lung phenotypes studies. **a** The plot of the running sum of GWAS studies reported from 2007 to 2021. The colours show details about the race/ancestry groups: Africans only, Europeans only, Europeans and Africans, Others, and those for which the race/ancestry group is “Not provided”. **b** The total number of GWAS studies reported for each race/ancestry group combination. The colours depict information about race and ethnic groups. **c** The trend of the cumulative sum of participants (on the y-axis) of studies from 2007 to 2021. The colours show details about the race/ancestry groups. The marks are labelled by the cumulative sum of participants. The figure insert shows the total number of participants by race/ancestry group.

Methods

We analysed a UK Biobank¹¹ dataset of 383,471 individuals of European ancestry (designated as White, British, Irish, and “any other white background”) and 5978 individuals of recent African ancestry. The UK biobank obtained all participant samples and body measurements from consenting individuals. Information on the UK biobank ethics policy and approval can be found here: <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>. The demographics of the UK Biobank participants are extensively described elsewhere¹¹. The data elements we analysed include genotyping array data of imputed SNPs, anthropometric measurements, and pulmonary function parameters: FVC, FEV1, and PEF. The ancestry groups were initially defined by self-identification. Then, a principal component analysis was performed, followed by a random forest on the projected principal component analysis data to reassign the initial self-defined ancestries of individuals with a membership posterior probability >0.5. Other individuals with a posterior probability less than 0.5 for any given ancestry group were dropped from further analysis.

Comparison of pulmonary function parameters in Europeans and Africans. We compare the mean values of the pulmonary function parameters FVC, FEV1, and PEF between 383,471 Europeans and 5978 Africans using the Welch t-test. Furthermore, to evaluate how FVC, FEV1, PEF, and FEV1/FVC values vary with the participant’s body mass index, height, and age, we calculated the 10th percentile bins of each anthropometric measurement and visualised the trend using error bars plotted for each percentile.

Genome-wide identification of genetic variants and associations. The methods applied for genotyping participants in the UK Biobank are reported elsewhere^{11,63}. Furthermore, the genotyping quality control implemented for the analyses is described at the following link <https://pan.ukbb.broadinstitute.org/docs/qc>. We obtained the GWAS summary statistics computed by the UK Biobank project for each pulmonary function parameter. The methods used to perform the GWA

analyses are described elsewhere^{64,65}. Briefly, the GWAS was performed for the pulmonary function phenotypes and ancestry groups using the Scalable and Accurate Implementation of Generalized Mixed Model Approach 65, using a linear or mixed logistic model including a kinship matrix as a random effect and covariates as fixed effects. The covariates included the participant's age, sex, age multiplied by sex, the square of the age, the square of the age multiplied by the sex, and the first 10 principal components calculated from the genotype datasets. The Manhattan plots were produced in MATLAB using the software described here⁶⁶. Furthermore, we used the Probabilistic Identification of Causal SNPs software with default settings to fine-map SNPs to identify the most credible causal SNPs within each linkage disequilibrium block while conditioning on the lead SNP signal in each locus ± 500 kb²¹.

Identification of unique and common variants. We applied the following approach to identify the unique variants associated with pulmonary function traits in Africans and Europeans. First, we extracted all the credible sets of causal variants associated with pulmonary function (FVC, FEV₁, and PEF) within ± 500 kb of the most statistically significant variant within a particular linkage disequilibrium block. Then, the linkage disequilibrium structure of the populations was estimated using the UK Biobank and the same individuals used in the analysis. If a causal variant associated with one pulmonary function parameter (e.g., FVC) was associated with another pulmonary function measure (e.g., p -values $< 5 \times 10^{-8}$ for FEV₁) or in linkage disequilibrium ($r^2 > 0.4$) with a variant associated with another pulmonary function parameter (e.g., FEV₁), then we return the most statistically significant variant (i.e., the variant with the smallest GWA estimated p -value). This approach allowed us to remove 190 non-independent variants from the 820 (FVC = 310, FEV₁ = 309, and PEF = 374) causal variants, leaving 630 independent (credible set) causal variants (FVC = 256, FEV₁ = 233, and PEF = 297) associated with pulmonary function.

Replication of variants in Africans. We attempted to replicate the significant finding from Europeans in Africans because the variants associated ($p < 5 \times 10^{-8}$) with pulmonary function in Europeans were not associated with pulmonary function in Africans. Here, for variants significantly associated with a pulmonary function parameter, we extracted all the variants linked (linkage disequilibrium: $r^2 > 0.4$) to the causal variant. The linked variants in Europeans were then assessed for their association with the trait in Africans by extracting the estimate of GWA p -values in Africans and adjusting the p -values using the Benjamini and Hochberg procedure. Finally, we considered all variants with the adjusted p -values < 0.05 , within each linkage disequilibrium block, as evidence of local replication.

Pathways and enrichment analyses. We used NBCI's dbSNP^{67,68} to ascribe the significant variants associated (suggestive cutoff p -value of 1×10^{-6})²² with pulmonary function identified using GWAS to specific genes. This yielded a list of genes associated with pulmonary function in Europeans or Africans. Finally, using these two gene lists (for Europeans and Africans), we separately performed gene set enrichment analysis⁶⁹ using Enrichr⁷⁰ to identify the Elsevier pathways⁷⁰, Disease Gene Network database³⁹, Phenotype and Genotype Integrator database, and GWAS Catalog²⁷ ontology terms that are significantly enriched for (see Supplementary Data 3).

GWAS literature, disease phenotypes, and eQTLs. We retrieved data from the previous GWAS of pulmonary function and pulmonary function-related phenotypes from GWAS Catalog²⁷. This information was subset into two categories: "pulmonary reported"; for those studies that reported pulmonary function phenotype, and "pulmonary associated" for those that reported associations related to pulmonary function-related phenotypes (see Supplementary Data 4). We used the approach described above to identify variants previously reported to be associated with pulmonary function or pulmonary disease in GWA studies in the GWAS Catalog to identify novel variants associated with pulmonary function separately for Europeans and Africans. Briefly, for each variant we found associated with pulmonary function, we searched for variants in the GWAS Catalog that are in strong linkage disequilibrium ($r^2 > 0.4$) with the variant. If any variant meets this criterion, we consider the associated variant in our study to have been previously reported elsewhere or otherwise novel. Furthermore, we obtained information on diseases associated with the genes in which the variants are located from the Pharos database⁷¹. Finally, information on SNPs that are expression quantitative trait loci in the lungs was obtained from the Genotype-Tissue Expression consortium database⁷².

Statistics and reproducibility. We performed the statistical analyses in R programming language, MATLAB 2021a and Bash. We used the Welch test, Wilcoxon rank-sum test and the one-way analysis of variance to compare continuous measures among groups. All statistical tests were considered significant if the two-sided p -value was < 0.05 for single comparisons. The multiple hypotheses tests were corrected by calculating a two-sided q -value (False Discovery Rate) for each group/comparison using the Benjamini and Hochberg procedure⁷³.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets that support the results presented in this manuscript are available from: the UK Biobank; <https://www.ukbiobank.ac.uk> and <https://pan-ukb-us-east-1.s3.amazonaws.com>, dbSNP; <https://www.ncbi.nlm.nih.gov/snp>, the GWAS Catalog; <https://www.ebi.ac.uk/gwas>, and interactive Manhattan visualisations and regional plots of chromosomes found at <https://public.tableau.com/app/profile/musalula.sinkala7788/viz/FEV1VariantsbyChromosome/ChromosomeFilter>, <https://public.tableau.com/app/profile/musalula.sinkala7788/viz/FVCVariantsbyChromosome/ChromosomeFilter>, and <https://public.tableau.com/app/profile/musalula.sinkala7788/viz/PEFVariantsbyChromosome/ChromosomeFilter>. The source data underlying Figs. 1, 3, and 5 are presented in Supplementary Data 5. Furthermore, the GWA summary statistics derived by the Pan-UK Biobank project's⁶⁴ for the three pulmonary function parameters are available via the Amazon Web Services links: FCV: https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files/continuous-3062-both_sexes-irmt.tsv.bgz FEV₁: https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files/continuous-3063-both_sexes-irmt.tsv.bgz PEF: https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files/continuous-3064-both_sexes-irmt.tsv.bgz.

Code availability

Custom code written in MATLAB for processing and analysis of the data presented here is freely available at <https://doi.org/10.5281/zenodo.751661874>.

Received: 20 February 2022; Accepted: 9 January 2023;

Published online: 14 January 2023

References

1. Reilly, J. J. COPD and declining FEV₁—time to divide and conquer? *N. Engl. J. Med.* **359**, 1616–1618 (2008).
2. Lange, P. et al. Lung-function trajectories leading to chronic obstructive pulmonary disease. *N. Engl. J. Med.* **373**, 111–122 (2015).
3. Braun, L. Race, ethnicity and lung function: a brief history. *Can. J. Respir. Ther.* **51**, 99–101 (2015).
4. Bui, D. S. et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. *Lancet Respir. Med.* **6**, 535–544 (2018).
5. Soler Artigas, M. et al. Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat. Commun.* **6**, 8658 (2015).
6. Loth, D. W. et al. Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat. Genet.* **46**, 669–677 (2014).
7. Cho, M. H. et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir. Med.* **2**, 214–225 (2014).
8. Wain, L. V. 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).
9. Hobbs, B. D. et al. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* **49**, 426–432 (2017).
10. Wain, L. V. et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med.* **3**, 769–781 (2015).
11. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
12. Braun, L., Wolfgang, M. & Dickerson, K. Defining race/ethnicity and explaining difference in research studies on lung function. *Eur. Respir. J.* **41**, 1362–1370 (2013).
13. Whitrow, M. J. & Harding, S. Ethnic differences in adolescent lung function: anthropometric, socioeconomic, and psychosocial factors. *Am. J. Respir. Crit. Care Med.* **177**, 1262–1267 (2008).
14. Whittaker, A. L., Sutton, A. J. & Beardsmore, C. S. Are ethnic differences in lung function explained by chest size? *Arch. Dis. Child Fetal Neonatal Ed.* **90**, F423–F428 (2005).
15. Anderson, M. A., Malhotra, A. & Non, A. L. Could routine race-adjustment of spirometers exacerbate racial disparities in COVID-19 recovery? *Lancet Respir. Med.* **9**, 124–125 (2021).
16. Braun, L. Race correction and spirometry: why History matters. *Chest* **159**, 1670–1675 (2021).
17. Buchman, A. S. et al. Pulmonary function, muscle strength and mortality in old age. *Mech. Ageing Dev.* **129**, 625–631 (2008).

18. Shapira, N. et al. Determinants of pulmonary function in patients undergoing coronary bypass operations. *Ann. Thorac. Surg.* **50**, 268–273 (1990).
19. Schoenberg, J. B., Beck, G. J. & Bouhuys, A. Growth and decay of pulmonary function in healthy blacks and whites. *Respir. Physiol.* **33**, 367–393 (1978).
20. Park, J. E., Chung, J. H., Lee, K. H. & Shin, K. C. The effect of body composition on pulmonary function. *Tuberc. Respir. Dis. (Seoul.)* **72**, 433–440 (2012).
21. Taylor, K. E., Ansel, K. M., Marson, A., Criswell, L. A. & Farh, K. K. PICS2: Next-generation fine mapping via probabilistic identification of causal SNPs. *Bioinformatics* **37**, 3004–3007 (2021).
22. Hammond, R. K. et al. Biological constraints on GWAS SNPs at suggestive significance thresholds reveal additional BMI loci. *Elife* **10** (2021).
23. Baselmans, B. M. L. et al. Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).
24. Nagel, M., Speed, D., van der Sluis, S. & Ostergaard, S. D. Genome-wide association study of the sensitivity to environmental stress and adversity neuroticism cluster. *Acta Psychiatr. Scand.* **141**, 476–478 (2020).
25. 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).
26. Hatzikotoulas, K. et al. Genome-wide association study of developmental dysplasia of the hip identifies an association with GDF5. *Commun. Biol.* **1**, 56 (2018).
27. Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
28. Sloan, D. B., Fields, P. D. & Havird, J. C. Mitonuclear linkage disequilibrium in human populations. *Proc. Biol. Sci.* **282**, 20151704 (2015).
29. Mangin, B. et al. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb.)* **108**, 285–291 (2012).
30. Berndt, S. I. et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45**, 501–512 (2013).
31. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
32. van der Valk, R. J. et al. A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Hum. Mol. Genet.* **24**, 1155–1168 (2015).
33. Singh, I. et al. Hmga2 is required for canonical WNT signaling during lung development. *BMC Biol.* **12**, 21 (2014).
34. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
35. Warrington, N. M. et al. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat. Genet.* **51**, 804–814 (2019).
36. Hubel, C. et al. Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **180**, 428–438 (2019).
37. Hancock, D. B. et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42**, 45–52 (2010).
38. Kim, W. et al. Genome-wide gene-by-smoking interaction study of chronic obstructive pulmonary disease. *Am. J. Epidemiol.* **190**, 875–885 (2021).
39. Pinero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
40. Ramos, E. M. et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **22**, 144–147 (2014).
41. Klimentidis, Y. C. et al. Heritability of pulmonary function estimated from pedigree and whole-genome markers. *Front. Genet.* **4**, 174 (2013).
42. Mak, A. C. Y. et al. Lung function in African American children with asthma is associated with novel regulatory variants of the KIT ligand KITLG/SCF and gene-by-air-pollution interaction. *Genetics* **215**, 869–886 (2020).
43. Harik-Khan, R. I., Fleg, J. L., Muller, D. C. & Wise, R. A. The effect of anthropometric and socioeconomic factors on the racial difference in lung function. *Am. J. Respir. Crit. Care Med.* **164**, 1647–1654 (2001).
44. Barnes, K. C. Genomewide association studies in allergy and the influence of ethnicity. *Curr. Opin. Allergy Clin. Immunol.* **10**, 427–433 (2010).
45. Chan, S. L., Jin, S., Loh, M. & Brunham, L. R. Progress in understanding the genomic basis for adverse drug reactions: a comprehensive review and focus on the role of ethnicity. *Pharmacogenomics* **16**, 1161–1178 (2015).
46. Ueta, M. et al. Genome-wide association study using the ethnicity-specific Japonica array: identification of new susceptibility loci for cold medicine-related Stevens-Johnson syndrome with severe ocular complications. *J. Hum. Genet.* **62**, 485–489 (2017).
47. Jorgenson, E. et al. Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol. Psychiatry* **22**, 1359–1367 (2017).
48. Asif, H. et al. GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Mol. Psychiatry* **26**, 2048–2055 (2021).
49. Ball, R. D. Designing a GWAS: power, sample size, and data structure. *Methods Mol. Biol.* **1019**, 37–98 (2013).
50. Lutz, S. M. et al. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet.* **16**, 138 (2015).
51. Moll, M. et al. A systematic analysis of protein-altering exonic variants in chronic obstructive pulmonary disease. *Am. J. Physiol. Lung Cell Mol. Physiol.* **321**, L130–L143 (2021).
52. Emison, E. S. et al. Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. *Am. J. Hum. Genet.* **87**, 60–74 (2010).
53. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15**, 765–776 (2014).
54. Katsanis, N. The continuum of causality in human genetic disorders. *Genome Biol.* **17**, 233 (2016).
55. Minikel, E. V. et al. Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).
56. Dobson, A. J. An introduction to generalized linear models, vii, 225 p. (Chapman & Hall/CRC, Boca Raton, 2002).
57. Rule, A. D. et al. Association between SLC2A9 transporter gene variants and uric acid phenotypes in African American and white families. *Rheumatol. (Oxf.)* **50**, 871–878 (2011).
58. Reiner, A. P. et al. Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet.* **7**, e1002108 (2011).
59. Buckley, M. T. et al. Selection in Europeans on fatty acid desaturases associated with dietary changes. *Mol. Biol. Evol.* **34**, 1307–1318 (2017).
60. Batai, K. et al. Common vitamin D pathway gene variants reveal contrasting effects on serum vitamin D levels in African Americans and European Americans. *Hum. Genet.* **133**, 1395–1405 (2014).
61. Mathias, R. A. et al. A combined genome-wide linkage and association approach to find susceptibility loci for platelet function phenotypes in European American and African American families with coronary artery disease. *BMC Med. Genomics* **3**, 22 (2010).
62. Larkin, E. K. et al. A candidate gene study of obstructive sleep apnea in European Americans and African Americans. *Am. J. Respir. Crit. Care Med.* **182**, 947–953 (2010).
63. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P. R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
64. P-U, T. Quality Control (QC) | Pan UKBB n.d. <https://pan-dev.ukbb.broadinstitute.org/docs/qc/index.html>. (2021).
65. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
66. H, G. Manhattan Plots for visualisation of GWAS results - File Exchange - MATLAB Central n.d. https://www.mathworks.com/matlabcentral/fileexchange/69549-manhattan-plots-for-visualisation-of-gwas-results?s_tid=srchtitle. (2021).
67. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res.* D29–D38 (2022).
68. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
69. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
70. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
71. Nguyen, D. T. et al. Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **45**, D995–D1002 (2017).
72. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
73. Benjamini, Y. & Yekutieli, D. Quantitative trait Loci analysis using the false discovery rate. *Genetics* **171**, 783–790 (2005).
74. Sinkala, M. <https://doi.org/10.5281/zenodo.7516618>. 09-01-2022 edn (2022).

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 53163. The funding for this project was provided by H3ABioNet, supported by the National Institutes of Health Common Fund under grant number U24HG006941. The content of this publication is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

M.S., S.E., and N.M. conceptualized the study. M.S., N.M., S.E., and J.C. designed the

methodology, and M.M. M.S., M.M., J.C., and S.E. performed the formal analysis of the data. M.S., N.M., and S.E. drafted the manuscript. Editing and reviewing the manuscript were carried out by M.S., N.M., S.E., J.C., and M.M. Data visualisations were produced by M.S. and S.E.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-04443-8>.

Correspondence and requests for materials should be addressed to Musalula Sinkala.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: George Inglis.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023