Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Kumar R, Seibold MA, Aldrich MC, et al. Genetic ancestry in lung-function predictions. N Engl J Med 2010. DOI: 10.1056/NEJMoa0907897.

Online Appendix:

Title: Self identified Race versus Genetic Ancestry for Lung Function Predictions among African Americans

Author Block: Rajesh Kumar MD[1*]; Max A. Seibold PhD[2*]; Melinda C. Aldrich, PhD, MPH[2]*; L. Keoki Williams MD, MPH[3]*; Alex P. Reiner MD, MSc[4]; Laura Colangelo MS[5]; Joshua Galanter MD[2]; Chris Gignoux MS[2]; Donglei Hu PhD[2]; Saunak Sen PhD[2]; Shweta Choudhry PhD[2]; Edward L. Peterson PhD[6]; Jose Rodriguez-Santana MD[7]; William Rodriguez-Cintron MD[8]; Michael A. Nalls, PhD[9]; Tennille S. Leak, PhD[10]; Ellen O'Meara, PhD[11] Bernd Meibohm, PhD[121]; Stephen B. Kritchevsky, PhD[13]; Rongling Li, MD, PhD, MPH[14]; Tamara B. Harris, MD, MS[15]; Deborah A. Nickerson, PhD[16]; Myriam Fornage PhD[17]; Paul Enright, MD[18]; Elad Ziv, MD[2]; Lewis J Smith MD[19]; Kiang Liu PhD[5]; Esteban González Burchard MD, MPH[2].

*** These authors are co-first authors and contributed equally to the manuscript.**

Affiliations:

1. Children's Memorial Hospital, Department of Pediatrics, Northwestern University Feinberg School of Medicine. 2. Department of Medicine, University of California San Francisco. 3. Center for Health Services Research, Department of Internal Medicine, Henry Ford Health System. 4. Department of Epidemiology, University of Washington School of Community Health and Public Medicine. 5. Department of Preventive Medicine, Northwestern University Feinberg School of Medicine. 6. Department of Biostatistics and Research Epidemiology, Henry Ford Health System 7.Pediatric Pulmonary Program of San Juan, San Juan, Puerto Rico. 8. San Juan VAMC, University of Puerto Rico School of Medicine, San Juan, Puerto Rico. 9. Laboratory of Neurogenetics, Intramural Research Program, National Institute on Aging. 10. Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh. 11. University of Washington, Collaborative Health Studies Coordinating Center. 12. College of Pharmacy, University of Tennessee Health Science Center. 13. Sticht Center on Aging, Wake Forest School of Medicine. 14. Department of Preventive Medicine, College of Medicine, University of Tennessee Health Science Center. 15. Laboratory of Epidemiology, Demography and Biometry, Intramural Research Program, National Institute on Aging. 16. Department of Genome Sciences, University of Washington. 17. Institute of Molecular Medicine and Division of Epidemiology School of Public Health, University of Texas Health Sciences Center at Houston. 18. College of Public Health, University of Arizona. 19. Division of Pulmonary and Critical Care; Department of Medicine, Northwestern University Feinberg School of Medicine.

This research was funded by contracts N01-HC-48047 through 48050 and N01-HC-95095 from the National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD. The National Institutes of Health also provided support for the following investigators: Dr. Burchard (HL078885, HL088133, ES015794, Robert Wood Johnson Foundation Amos Medical Faculty

Development Program, Flight Attendant Medical Research Institute (FAMRI)). Dr. Kumar (K23HL093023-01), Dr. Reiner (R01 HL71862, R01 HL71017, R01 AG032136), Dr. Williams (American Asthma Foundation, R01 AI61774, R01 AI79139, R01 HL79055), and Dr. Choudhry (Tobacco-Related Disease Research Program New Investigator Award, 15KT-0008). CHS was supported by contract numbers N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, grant number U01 HL080295 from the National Heart, Lung, and Blood Institute, with additional contribution from the National Institute of Neurological Disorders and Stroke. A full list of principal CHS investigators and institutions can be found at http://www.chs-nhlbi.org/pi.htm. Support for CHS was also provided by U19 AG023122 from the National Institute on Aging Longevity Consortium (PI Dr. Cummings). Health ABC is supported in part by the Intramural Research Program of the NIH, National Institute on Aging and by contracts N01-AG-6-2101; N01-AG-6-2103; N01-AG-6-2106. Genotyping in Health ABC was supported by NCRR grant U54-RR020278.

Corresponding Author:
Rajesh Kumar MD
Division of Allergy and Immunology
Children's Memorial Hospital
2300 Children's Plaza, box #60
Chicago Il
60614
Ph: 773 327 3614
Fax: 773 327 3790

Table of Contents:

Table 1

A. Models with and without the inclusion of ancestry in CARDIA males (n=309)

	With ancestry			Without ancestry		
Variable	**FEV_1**	**FVC**	**FEV_1/FVC**	**FEV_1**	**FVC**	**FEV_1/FV**
Intercept	3885.614[d]	4763.862[d]	81.782[d]	3906.125[d]	4794.381[d]	81.688[d]
Age (years)	3.748	20.570[b]	-0.290c	3.724	20.535[b]	-0.290[c]
African ancestry	-8.144[c]	-12.118[d]	0.037	--	--	--
Pack-years of smoking	-9.125	-7.822	0.000	-8.386	-6.722	-0.003
BMI (kg/m^2)	9.111	20.168[b]	-0.145	9.245	20.367[b]	-0.146
Height (cm)	38.531[d]	53.702[d]	-0.118[b]	39.232[d]	54.744[d]	-0.121[b]
Height^2 (cm^2)	-0.373	0.057	-0.008	-0.382	0.044	-0.008
Chicago	-91.636	-206.255[b]	1.613	-110.068	-233.680[b]	1.697
Minnesota	56.150	-22.967	1.464	36.484	-52.227	1.554
Oakland	-44.112	-193.851	2.289[b]	-85.879	-255.995[c]	2.480[c]

[a] p<0.10, [b] p<0.05, [c] p<0.01, [d] p<0.001

FEV_1 denotes forced expiratory volume at one second; FVC, forced vital capacity; and BMI, body mass index.

B. Models with and without the inclusion of ancestry in CARDIA females (n=468)

	With ancestry			Without ancestry		
Variable	**FEV_1**	**FVC**	**FEV_1/FVC**	**FEV_1**	**FVC**	**FEV_1/FV**
Intercept	2891.948[d]	3368.380[d]	86.098[d]	2905.905[d]	3382.010[d]	86.177[d]
Age (years)	-9.551[b]	0.229	-0.313[d]	-9.789[b]	-0.003	-0.315[d]
African ancestry	-5.538[c]	-5.408[c]	-0.032	--	--	--
Pack-years of smoking	-14.950[c]	-8.719	-0.222[c]	-14.409[c]	-8.191	-0.219[c]
BMI (kg/m^2)	-0.688	-0.122	-0.029	-1.272	-0.692	-0.033
Height (cm)	33.624[d]	42.579[d]	-0.071[b]	33.875[d]	42.825[d]	-0.070[b]
Height^2 (cm^2)	0.004	0.157	-0.005[a]	-0.016	0.137	-0.005[b]
Chicago	4.121	60.641	-1.674[b]	8.863	65.273	-1.647[b]
Minnesota	-56.821	-57.747	-0.175	-88.684[b]	-88.862[a]	-0.357
Oakland	-58.769	-91.078[a]	0.567	-90.364[b]	-121.932[b]	0.387

[a] p<0.10, [b] p<0.05, [c] p<0.01, [d] p<0.001

FEV_1 denotes forced expiratory volume at one second; FVC, forced vital capacity; and BMI, body mass index.

Table 2

Associations between lung function and ancestry comparing sitting and standing height in the equations among individuals in the HABC study

	Estimates with Standing Height			Estimates with Sitting Height		
Variable	**FEV_1**	**FVC**	**FEV_1/FVC**	**FEV_1**	**FVC**	**FEV_1/FVC**
Intercept	2477.127	3184.098	76.885	2700.944[d]	3574.137[d]	75.859[d]
Age (years)	-17.271[d]	-19.863[d]	-0.127	-16.229[c]	-18.558[c]	-0.128
Sex	-348.660[d]	-442.678[d]	-0.288	-416.428[d]	-548.847[d]	-0.118
African	-3.991[d]	-5.497[d]	0.012	-3.280[c]	-4.488[d]	0.011
Pack-years of smoking	-4.800[d]	-3.451[d]	-0.062[d]	-4.542[d]	-2.962[d]	-0.064[d]
BMI (kg/m^2)	-1.358	-10.250[c]	0.207[d]	-4.876	-15.751[d]	0.218[d]
Height (cm)	21.119[d]	33.963[d]	-0.070[b]	-	-	-
Height^2 (cm^2)	-0.152	0.092	-0.004	-	-	-
Sitting height	-	-	-	30.171[d]	49.801[d]	-0.117[a]
Sitting height^2 (cm)	-	-	-	-0.143	0.367	-0.008
Site1	65.862[b]	121.306[d]	-0.276	100.171[c]	151.115[d]	0.350

[a] $p<0.10$, [b] $p<0.05$, [c] $p<0.01$, [d] $p<0.001$

FEV_1 denotes forced expiratory volume at one second; FVC, forced vital capacity; and BMI, body mass index.

Table 3

Comparison of *standard race-based models* with *ancestry-based models* to fit measures of forced expiratory volume in one second in African American participants from three population-based cohorts*

Variable	Study Cohort											
	CARDIA				HABC				CHS			
	Female subjects ≥18 years (n=466)		Male subjects ≥20 years (n=271)		Female subjects ≥18 years (n=432)		Male subjects ≥20 years (n=366)		Female subjects ≥18 years (n=367)		Male subjects ≥20 years (n=212)	
	Standard race-based model	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*
Intercept	-0.06597	-0.12437	0.42053	0.50933	0.921	0.047	1.93[b]	2.06[b]	0.499	0.737	1.281	1.216
Age (years)	0.02572	0.03378	-0.00263	-0.00401	-0.007	0.014	-0.023[b]	-0.022[b]	0.011	0.005	-0.022[c]	-0.021[c]
Age^2 (years^2)	-0.000805	-0.000964			-0.00008	-0.0002			-0.00018	-0.00014		
Height^2 (cm^2)	0.000104[d]	0.000102[d]	0.000111[d]	0.000109[d]	0.00007[d]	0.00007[d]	0.00007[d]	0.00006[d]	0.000047[d]	0.00005[d]	0.00008[d]	0.00008[d]
African ancestry		-0.00633[d]		-.00849[c]		-0.004[c]		-0.005[b]		-0.0026[c]		-0.0021
MSE†	0.11867	0.11519	0.26232	0.25532	0.1248633	0.1254647	0.2639493	0.2592643	0.1460157	0.1435879	0.3095253	0.308947
Model R^2‡	0.336	0.357	0.209	0.233	0.163	0.179	0.117	0.121	0.121	0.138	0.159	1.165
Adjusted model R^2§	0.332	0.351	0.203	0.224	0.157	0.170	0.113	0.113	0.113	0.128	0.151	0.153

CARDIA denotes the Coronary Artery Risk Development in Young Adults study; HABC, the Health Aging and Body Composition study; CHS, the Cardiovascular Health Study; and MSE, mean square error.

* *Standard race-based models* used the age, race, and sex specific covariates found by Hankinson *et al.* to best predict forced expiratory volume in one second (FEV_1).[4] *Ancestry-based models* used the same age, race, and sex specific covariates, but also included a term for individual African ancestry.

†Mean square error is a measure of the difference between the estimator variables and the variable being estimated (i.e., FEV_1). The model with the smallest MSE is considered to have the best fit.

‡Model R^2 refers to the proportion of the variance in FEV_1 explained by the variables included in the model. Higher values indicate that the variables explain a larger proportion of the variance.

§Adjusted model R^2 is a modification of the R^2 statistic which accounts for the number of explanatory variables included in the model.

[a] $p=0.05$, [b] $p<0.05$, [c] $p<0.01$, [d] $p<0.001$

Table 4

Comparison of *standard race-based models* with *ancestry-based models* to fit measures of forced vital capacity in African American participants from three population-based cohorts*

Variable	Study Cohort											
	CARDIA				HABC				CHS			
	Adult Females (n=466)		Adult Males (n=271)		Adult Females (n=435)		Adult Males (n=378)		Adult Females (n=367)		Adult Males (n=212)	
	Standard race-based model	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*	*Standard race-based model*	*Ancestry-based model*
Intercept	-0.022	-0.086	-0.442	-0.301	6.998	5.859	1.638[a]	1.745[a]	0.819	1.161	1.825[a]	1.725
Age (years)	-0.012	-0.003	0.014	0.011	-0.174	-0.147	-0.025[b]	-0.025[b]	0.003	-0.005	-0.028[b]	-0.027[c]
Age^2 (years^2)	0.0002	0.00002			0.001	0.0009			-0.0001	-0.00009		
Height^2 (cm^2)	0.0001[d]	0.0001[d]	0.0002[d]	0.0001[d]	0.00009[d]	0.000096[d]	0.0001[d]	0.0001[d]	0.00007[d]	0.00007[d]	0.0001[d]	0.0001[d]
African ancestry		-0.007[d]		-0.013[d]		-0.005[d]		-0.008[d]		-0.004[c]		-0.003
MSE†	0.16053	0.15651	0.36464	0.34587	0.1649010	0.1636526	0.3107951	0.3043288	0.2570186	0.2519036	0.5290671	0.5266405
Model R^2‡	0.368	0.385	0.267	0.308	0.201	0.225	0.214	0.232	0.128	0.147	0.166	0.174
Adjusted model R^2§	0.364	0.380	0.262	0.300	0.196	0.217	0.210	0.225	0.120	0.138	0.158	0.162

[a] $p<0.10$, [b] $p<0.05$, [c] $p<0.01$, [d] $p<0.001$

CARDIA denotes the Coronary Artery Risk Development in Young Adults study; HABC, the Health Aging and Body Composition study; CHS, the Cardiovascular Health Study; and MSE, mean square error.

* *Standard race-based models* used the age, race, and sex specific covariates found by Hankinson *et al.* to best predict forced vital capacity (FVC).[4] *Ancestry-based models* used the same age, race, and sex specific covariates, but also included a term for individual African ancestry.

†Mean square error is a measure of the difference between the estimator variables and the variable being estimated (i.e., FVC). The model with the smallest MSE is considered to have the best fit.

‡Model R^2 refers to the proportion of the variance in FEV_1 explained by the variables included in the model. Higher values indicate that the variables explain a larger proportion of the variance.

§Adjusted model R^2 is a modification of the R^2 statistic which accounts for the number of explanatory variables included in the model.

[a] $p=0.05$, [b] $p<0.05$, [c] $p<0.01$, [d] $p<0.001$

Online text supplement:

Primary study population

The Coronary Artery Risk Development in Young Adults (CARDIA) study is a prospective cohort study initiated in 1985 with 5,115 African American and European American participants aged 18 to 30 years and recruited from 4 clinical sites. Detailed study design, recruitment procedures, and baseline findings have been previously reported.[1, 2] For the analysis of the CARDIA population we included all African American subjects who had Affymetryx 6.0 genotyping carried out, had satisfactory lung function performed, and did not have asthma at the baseline study visit (N=777).

Demographic and clinical measures

Epidemiologic data, including demographic characteristics and smoking exposure at baseline (as lifetime pack years), were collected using a standardized questionnaire. Race/ethnicity was self-reported and subjects chose from the following racial/ethnic categories: Hispanic, non-Hispanic white, non-Hispanic black, American Indian or Alaska Native, Asian or Pacific Islander, other, or not known. At each examination, weight and height were measured using a balance-beam scale and a vertical ruler. Body mass index (BMI) was calculated as weight in kilograms divided by height in meters squared. Spirometry was measured in accordance with American Thoracic Society (ATS) criteria for accuracy and precision.[3] Standard quality control and testing procedures were followed at all examinations.[4, 5]

Additional study populations

The association between baseline lung function and African ancestry was replicated in two independent cohorts: the Health Aging and Body Composition (HABC) study and the Cardiovascular Health Study (CHS).

The HABC study included African American and white participants who were well-functioning Medicare recipients residing near Pittsburgh, Pennsylvania or Memphis, Tennessee between April 1997 and June 1998.[6] Demographic and clinical data were collected using structured questionnaires, medical examinations and spirometry. We included all 813 participants who self identified as African American, performed satisfactory spirometry according to ATS criteria, and did not have a diagnosis of pulmonary disease. Individuals were genotyped for 1,332 AIMs, and the software program STRUCTURE was used to estimate African and European ancestry.[7]

The CHS included African American adults 65 years and older recruited from the following U.S. sites: Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Pittsburgh, Pennsylvania.[8] At study entry, CHS participants underwent a medical examination which included spirometry.[9] Individual African ancestry was estimated by a maximum likelihood method using 24 AIMs.[10] The sample for the current study included 579 African American men and women who had lung function measured at baseline and did not have a prior diagnosis of asthma or COPD.

We used two additional study populations, the Study of African Americans, Asthma, Genes & Environments (SAGE) and the Study of Asthma Phenotypes and Pharmacogenomic Interactions

by Race-Ethnicity (SAPPHIRE), to assess the effect of the differing pulmonary prediction models on classifying asthma severity. Both cohorts comprised African American individuals with asthma. The SAGE cohort included individuals age 8-40 years recruited from the San Francisco Bay Area, whereas the SAPPHIRE cohort included individuals age 12-56 years from southeast Michigan. In both studies, spirometry was performed at the time of enrollment.

Details of Marker selection in CARDIA:

Alexander *et al*. developed a block relaxation algorithm to maximize the likelihood of the admixture; the program which uses this method is called ADMIXTURE.[11] To implement, we first filtered genome wide array data (Affymetrix 6.0 platform) from CARDIA using the program PLINK[12] to remove genotypes with >5% missing values and Hardy-Weinberg equilibrium p-values $<10^{-5}$. We then removed SNPs in high linkage disequilibrium by pruning one SNP from a pair with an $r^2 \geq 0.8$. Fifty SNPs were analyzed at a time and the frame was advanced by 5 SNPs after each analysis. This protocol resulted in final sample of 631,243 autosomal SNPs from which to estimate admixture. We used the HapMap CEU and YRI populations as proxies for European and sub-Saharan African ancestral populations, respectively. ADMIXTURE estimated the individual ancestral proportions for 80 individuals at a time.

REFERENCES

1. Friedman GD, Cutter GR, Donahue RP, et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol 1988;41(11):1105-16.
2. Cutter GR, Burke GL, Dyer AR, et al. Cardiovascular risk factors in young adults. The CARDIA baseline monograph. Control Clin Trials 1991;12(1 Suppl):1S-77S.
3. Standardization of Spirometry, 1994 Update. American Thoracic Society. Am J Respir Crit Care Med 1995;152(3):1107-36.
4. Standardization of spirometry--1987 update. Statement of the American Thoracic Society. Am Rev Respir Dis 1987;136(5):1285-98.
5. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. Eur Respir J 2005;26(2):319-38.
6. Visser M, Newman AB, Nevitt MC, et al. Reexamining the sarcopenia hypothesis. Muscle mass versus muscle strength. Health, Aging, and Body Composition Study Research Group. Ann N Y Acad Sci 2000;904:456-61.
7. Nalls MA, Wilson JG, Patterson NJ, et al. Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. Am J Hum Genet 2008;82(1):81-7.
8. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1991;1(3):263-76.
9. Griffith KA, Sherrill DL, Siegel EM, Manolio TA, Bonekat HW, Enright PL. Predictors of loss of lung function in the elderly: the Cardiovascular Health Study. Am J Respir Crit Care Med 2001;163(1):61-8.
10. Reiner AP, Ziv E, Lind DL, et al. Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study. Am J Hum Genet 2005;76(3):463-77.
11. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19(9):1655-64.
12. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(3):559-75.