SUPPLEMENTAL MATERIAL

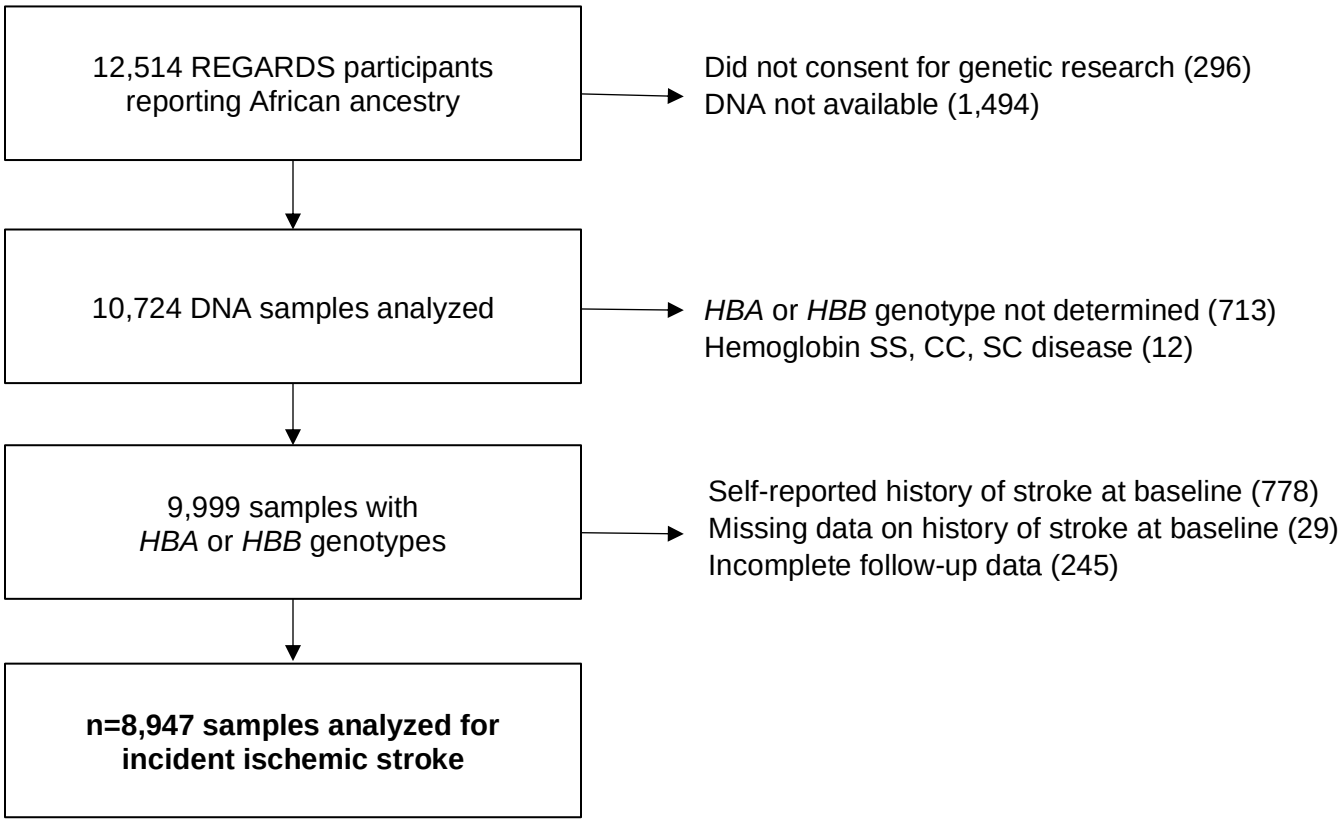
Alpha globin gene copy number and incident ischemic stroke risk among Black Americans

Table of Contents

I. Supplemental Figures and Tables

Supplemental Figure S1. Incident ischemic stroke and *HBA* copy number cohort study flow diagram

REGARDS = Reasons for Geographic and Racial Differences in Stroke (REGARDS) longitudinal cohort study; *HBA* = alpha globin gene; *HBB* = beta globin gene



Supplemental Table S1. Clinical and demographic characteristics by *HBA* copy number

		***HBA* copy number**			
	All Subjects	**2**	**3**	**4**	**≥ 5***
Subjects, N (%)	8947 (100)	368 (4)	2480 (28)	6014 (67)	85 (1)
Age, years	63.0 (57, 70)	64.0 (58, 70)	64.0 (57, 70)	63.0 (57, 70)	67.0 (61, 72)
Female sex, N (%)	5549 (62)	237 (64)	1532 (62)	3730 (62)	50 (59)
Smoking status, N (%)					
Never	4098 (46)	159 (44)	1160 (47)	2738 (46)	41 (48)
Past	3313 (37)	149 (41)	905 (37)	2227 (37)	32 (38)
Present	1493 (17)	57 (16)	402 (16)	1022 (11)	12 (14)
Region, N (%)					
Non-Belt	4430 (50)	162 (44)	1255 (51)	2959 (49)	54 (64)
Belt	2963 (33)	129 (35)	790 (32)	2026 (34)	18 (21)
Buckle	1554 (17)	77 (21)	435 (18)	1029 (17)	13 (15)
Medically insured, N (%)	8043 (90)	332 (90)	2241 (91)	5393 (90)	77 (91)
Education level, N (%)					
Less than high school	1640 (18)	63 (17)	464 (19)	1098 (18)	15 (18)
High school graduate	2462 (28)	102 (28)	6987 (28)	1624 (27)	27 (32)
Some college	2435 (27)	99 (27)	698 (28)	1615 (27)	23 (27)

College graduate or more	2404 (28)	104 (28)	607 (24)	1673 (28)	20 (24)
Income, N (%)					
≤ $20K	2306 (29)	98 (30)	660 (30)	1528 (29)	20 (26)
$20K - $34K	2341 (30)	100 (30)	674 (31)	1561 (30)	29 (38)
$35K - $74K	2392 (30)	102 (31)	674 (31)	1592 (30)	24 (31)
≥ $75K	864 (11)	30 (9)	198 (9)	632 (12)	4 (5)
Atrial fibrillation	636 (7)	23 (6)	184 (8)	422 (7)	7 (8)
Left ventricular hypertrophy	1265 (14)	51 (14)	358 (15)	845 (14)	11 (13)
Hypertension†, N (%)	7571 (86)	315 (87)	2114 (87)	5066 (86)	76 (92)
Diabetes mellitus, N (%)	2491 (28)	95 (26)	719 (29)	1655 (28)	22 (26)
Chronic Kidney disease, N (%)	2314 (27)	76 (21)	645 (27)	1566 (27)	27 (34)
Regular aspirin use	3268 (37)	130 (35)	917 (37)	2188 (36)	33 (39)
Lipid-lowering medication use	2640 (30)	108 (30)	729 (30)	1767 (30)	36 (43)
Framingham Stroke Risk Score	7.1 (3.8, 13.5)	6.9 (3.6, 13.9)	7.1 (3.9, 13.7)	7.1 (3.7, 13.4)	8.1 (5.1, 16.5)
ARIC Stroke Risk Score	6.2 (2.8, 14.1)	6.0 (2.7, 14.6)	6.3 (2.9, 14.6)	6.2 (2.8, 13.7)	6.0 (4.1, 18.5)
rs11248850					
G/G	4059 (59)	234 (81)	1311 (68)	2481 (54)	35 (53)
A/G	2492 (36)	53 (18)	577 (30)	1837 (40)	24 (36)

A/A	372 (5)	3 (1)	44 (2)	318 (7)	7 (11)
rs7203560					
A/A	6005 (87)	158 (54)	1388 (72)	4394 (95)	65 (98)
A/G	882 (13)	108 (37)	535 (28)	238 (5)	1 (2)
G/G	36 (<1%)	24 (8)	8 (<1%)	4 (<1%)	0 (0)
Hemoglobin, g/dL	13.1 (12.2, 14.0)	12.3 (11.5, 13.2)	12.9 (12.1, 13.8)	13.3 (12.4, 14.1)	13.1 (12.2, 14.1)
MCV, fL	88.0 (84.0, 92.0)	74.0 (72.0, 77.0)	84.0 (82.0, 87.0)	90.0 (87.0, 93.0)	88.0 (86.0, 92.0)
MCH, pg	29.8 (27.9, 30.9)	23.8 (22.9, 24.8)	27.9 (26.9, 28.9)	30.3 (29.2, 31.4)	29.8 (29.1, 30.9)
MCHC, g/dL	33.4 (32.9, 33.9)	32.1 (31.6, 32.5)	33.0 (32.6, 33.5)	33.7 (33.2, 34.1)	33.7 (33.2, 33.9)
RDW-CV, %	13.9 (13.3, 14.8)	15.0 (14.4, 15.9)	14.2 (13.5, 15.1)	13.8 (13.2, 14.6)	13.6 (13.2, 14.3)

HBA = alpha globin gene; P = p value; N = number; K = thousand; RBC = red blood cell; MCV = mean corpuscular volume; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; RDW-CV = red cell distribution width-coefficient of variation; ARIC = Atherosclerosis Risk in Communities Study; SNP = single nucleotide polymorphism; No. = number. Values are median (25^{th}, 75^{th} percentile) except where otherwise indicated.

*83 subjects had 5 *HBA* gene copies and 2 subjects had 6 *HBA* copies; †P values for tests of differences by *HBA* genotype generated from the chi-squared test for categorical variables and the Kruskal-Wallis non-parametric ANOVA test for continuous variables.

Missing data are as follows: medically insured (n=10, <0.01%); education (n=6, <0.01%); income (n=1,044 refused, 12%); hypertension (n=181, 2%); atrial fibrillation (n=229,3%); left ventricular hypertrophy (n=145, 2%); regular aspirin use (n=4, <0.01%); lipid-lowering medication use (n=87, 1%) kidney disease (n=359, 4%); diabetes mellitus (n=44, <0.01%); smoking status (n=43, <0.01%); Rs11248850, *HBA* regulatory SNP (n=2,023); rs7203560, *HBA* regulatory SNP (n=2,024); hemoglobin (n=2,853, 32%); MCV (n=2,858, 32%); MCH (n=2,853, 32%); MCHC (n=2,853, 32%); RDW-CV (n=2,863, 32%)' Framingham risk score (n=503, 6%); ARIC risk score (n=390, 4%). All other variables in Table 1 had no missing values.

Supplemental Table S2. **Pre-specified tests for interaction between *HBA* genotype and age, sex, and sickle cell trait on incident ischemic stroke in fully adjusted models**

AgeHBA***	0.61
SexHBA***	0.77
Sickle cell traitHBA***	0.74

HBA= alpha globin gene; HR= hazard ratio; CI= 95% confidence interval

*P values were generated with Cox proportional hazards multivariable regression models employing a linear effect (i.e., additive model for risk) of *HBA* allele count on the log of the hazard ratio. Each model was adjusted for HBA genotype, age per year, male sex, region, medically insurance status, education level, income, hypertension, atrial fibrillation, left ventricular hypertrophy, diabetes mellitus, and smoking status with age, sex, and sickle cell trait individually added with interaction terms. Multiple imputations were performed for missing data.

Supplemental Table S3. Pre-specified sensitivity analyses for the association of *HBA* genotype with incident ischemic stroke – fully adjusted models with the separate addition of covariates for each model.

	HR	95% CI	P value*
Hemoglobin			
HBA copy number	1.05	(0.89,1.24)	0.55
Hemoglobin	0.96	(0.86, 1.08)	0.52
Chronic kidney disease			
HBA copy number	1.03	(0.88,1.20)	0.75
Chronic kidney disease	1.81	(1.49, 2.20)	<0.001
Hemoglobin and Chronic kidney disease ‡			
HBA copy number	1.03	(0.88, 1.21)	0.71
Hemoglobin	0.99	(0.88, 1.10)	0.83
Chronic kidney disease	1.81	(1.49, 2.20)	<0.001
Regular aspirin use			
HBA copy number	1.04	(0.88, 1.21)	0.67
Regular aspirin use	1.26	(1.05, 1.51)	0.01
Regular statin use			
HBA copy number	1.04	(0.88, 1.21)	0.66

Regular statin use	1.05	(0.87, 1.28)	0.60
Principal components of ancestry			
HBA copy number	1.01	(0.86,1.19)	0.89
PC1	1.01	(0.91,1.12)	0.86
PC2	0.97	(0.88,1.07)	0.55
PC3	0.99	(0.90,1.09)	0.89
PC4	0.97	(0.89,1.07)	0.55
PC5	1.07	(0.98,1.18)	0.14
PC6	1.05	(0.96,1.16)	0.27
PC7	0.92	(0.84,1.01)	0.08
PC8	1.02	(0.92,1.12)	0.75
PC9	0.90	(0.82,0.99)	0.03
PC10	1.04	(0.95,1.15)	0.36
rs11248850, *HBA* regulatory SNP			
HBA copy number	1.04	(0.89,1.22)	0.63
Rs11248850	0.98	(0.83, 1.15)	0.81
rs7203560, *HBA* regulatory SNP			

HBA copy number	1.06	(0.89, 1.25)	0.52
rs7203560	1.09	(0.83, 1.44)	0.52

HBA = alpha globin gene; HR = hazard ratio; CI = confidence interval; PC = principal component; SNP = single nucleotide polymorphism

*P values were generated with Cox proportional hazards multivariable regression models employing a linear effect (i.e., additive model for risk) of *HBA* allele count on the log of the hazard ratio. Each model was adjusted for *HBA* genotype, age per year, sex, region, medically insurance status, education level, income, hypertension, atrial fibrillation, left ventricular hypertrophy, diabetes mellitus, and smoking status with listed covariates added in separate models. Multiple imputations were performed for missing data. †Principal components of ancestry model with n=7,032 participants and all other models with n=8,947. ‡This sensitivity analysis included both chronic kidney disease and hemoglobin added to the fully adjusted base model.

II. Supplemental Statistical Methods

Study Design

REGARDS is a longitudinal cohort study designed to determine the reasons for racial disparities in stroke and cognitive decline in Black and White Americans aged ≥ 45 years.[1] REGARDS enrolled 30,239 participants from the 48 continental United States from 2003 to 2007. All self-reported Black participants consenting to genetic research were included in this study (Figure 1). All participants provided oral and written informed consent. The REGARDS study was approved by the Institutional Review Boards of participating centers. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline. The analytic plan was prespecified and approved by the REGARDS Publications Committee.

Incident ischemic stroke was defined as non-hemorrhagic stroke with a focal neurological deficit lasting ≥ 24 hours confirmed by the medical record or a focal or non-focal neurological deficit with positive imaging confirmed with medical records in participants without a prior history of stroke. The covariates age, sex, race, health insurance (yes or no), highest education level obtained (less than high school, high school, some college, college or more), annual income (≤ $20K, $20-34K, $35-74K, ≥ $75K), and smoking status (categorized by never, past, or current smoker), history of hyperlipidemia, regular use of lipid lowering medication, and regular use of aspirin were self-reported. Self-reported use of lipid lowering medication was restricted to those also reporting hyperlipidemia. Region was defined as previously described in three geographic areas: stroke belt buckle, stroke belt, and stroke nonbelt.[2] Atrial fibrillation was defined by self-report of a physician diagnosis or ECG evidence. The presence of left ventricular hypertrophy (LVH) was defined by 12-lead ECG. Hypertension was defined as systolic blood pressure >=130, diastolic blood pressure >=80, or self-reported current medication use to control blood pressure. Chronic kidney disease was defined by the 2021 CKD-Epi creatinine-

cysteine equation and estimated glomerular filtration rate less than 60mL/min/1.73m^2, including those with end-stage kidney disease, and/or urine albumin to creatinine ratio $\geq$ 30 mg/g measured on urine collected during the baseline in-home examination. Fasting glucose levels $\geq$ 126 mg/dL, random glucose $\geq$ 200 mg/dL, or self-reported use of glucose-lowering medication was used to define diabetes mellitus. Hemoglobin, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and red-cell distribution width-coefficient of variation (RDW-CV) values were measured or calculated from blood collected during the in-home examination. The first 10 principal components of ancestry were calculated from Infinium Expanded Multi-Ethnic Genotyping Array data on 7,032 (79%) participants. *HBA* regulatory single nucleotide polymorphisms (SNPs) (rs11248850, rs11865131, rs7203560) were determined from these array data.[3,4] The SNPs rs11248850 and rs11865131 were in high linkage disequilibrium and were concordant in all but five participants. Therefore, we evaluated only rs11248850 and not rs11865131 in the models. The Framingham Stroke Risk Score (10-Year probability of stroke risk percentage)[5] and Atherosclerosis Risk in Communities Study (ARIC) Stroke Risk Score, 10-year probability of ischemic stroke risk (percentage) were calculated among those who self-reported at baseline never having had a stroke.[6]

Statistical methods

Multivariable Cox proportional hazards regression modeling was used to estimate the hazard ratio of *HBA* copy number on time to first ischemic stroke. Covariates included age, sex, region, insurance status, education level, income, hypertension, atrial fibrillation, left ventricular hypertrophy, smoking status, and diabetes mellitus. Pre-specified tests of interaction between *HBA* copy number and age, sex, and sickle cell trait were performed. Pre-specified sensitivity analyses were performed by adding each of the following to the model: hemoglobin, chronic kidney disease (CKD), both hemoglobin and CKD, regular aspirin use, regular statin use, the

first ten principal components of ancestry, and putative *HBA* regulatory single nucleotide polymorphisms rs11248850 and rs7203560.

For the ischemic stroke time-to-event analysis, time to event was defined as the number of years between the initial in-home interview date and date D where D was the minimum of D1 and D2 which were defined as follows. D1 was the last follow-up date provided by REGARDS as the last time the participant was contacted for status. The outcome associated with this date was either “No Event” or “Death”. D2 was the date when the individual had an ischemic stroke. Individuals with a D2 date prior to their initial in-home interview were excluded from consideration in this paper. Individuals with a D2 date after interview but prior to D1 time were recorded with an ischemic stroke event and D = D2. Individuals with D2 after D1 had stroke onset after their last REGARDS follow-up and they were recorded with “No Event” and censored at date D = D1. This decision to censor stroke events occurring after the end of REGARDS follow-up avoids bias associated with having extended follow-up only for one type of subgroup - those having a stroke event.

Missing data for the primary outcome, secondary outcomes, and explanatory variables were typically rare (< 0.5%) with some exceptions, e.g., hemoglobin values (Table S1). Multiple imputation methods were used in the multivariable analyses. The R package “mice” Version 3.14.0 was used to create and analyze the resulting imputations.[7]

For diagnostic modeling of the model of ischemic stroke incidence using Cox proportional hazard techniques, Schoenfeld residuals were examined over follow-up time to detect violations of proportional hazards assumptions for the covariates in the analysis of ischemic outcomes. Examination of Schoenfeld residuals of the non-imputed and imputed data sets showed no suggestion of violation of proportional hazards for the regression covariates ($p = 0.98$).

REFERENCES:

1. Howard VJ, Cushman M, Pulley L, Gomez CR, Go RC, Prineas RJ, Graham A, Moy CS, Howard G. The reasons for geographic and racial differences in stroke study: objectives and design. *Neuroepidemiology*. 2005;25.

2. Howard G, Howard VJ. Twenty Years of Progress Toward Understanding the Stroke Belt. *Stroke*. 2020;51:742–750.

3. Raffield LM, Ulirsch JC, Naik RP, Lessard S, Handsaker RE, Jain D, Kang HM, Pankratz N, Auer PL, Bao EL, Smith JD, Lange LA, Lange EM, Li Y, Thornton TA, Young BA, Abecasis GR, Laurie CC, Nickerson DA, McCarroll SA, Correa A, Wilson JG, Consortium NT-O for PM (TOPMed), Hemostasis H&, Diabetes, Groups and SVTopmW, Lettre G, Sankaran VG, Reiner AP. Common α-globin variants modify hematologic and other clinical phenotypes in sickle cell trait and disease. *PLOS Genet*. 2018;14:e1007293.

4. Milton JN, Rooks H, Drasar E, McCabe EL, Baldwin CT, Melista E, Gordeuk VR, Nouraie M, Kato GR, Kato GJ, Minniti C, Taylor J, Campbell A, Luchtman-Jones L, Rana S, Castro O, Zhang Y, Thein SL, Sebastiani P, Gladwin MT, Walk-PHAAST Investigators, Steinberg MH. Genetic determinants of haemolysis in sickle cell anaemia. *Br J Haematol*. 2013;161:270–278.

5. McClure LA, Kleindorfer DO, Kissela BM, Cushman M, Soliman EZ, Howard G. Assessing the Performance of the Framingham Stroke Risk Score in the Reasons for Geographic and Racial Differences in Stroke Cohort. *Stroke*. 2014;45:1716–1720.

6. Chambless LE, Heiss G, Shahar E, Earp MJ, Toole J. Prediction of ischemic stroke risk in the Atherosclerosis Risk in Communities Study. *Am J Epidemiol*. 2004;160:259–269.

7. Van Buuren S, Goorthius-Oudshoom K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45:1–67.