SUPPLEMENTAL MATERIAL

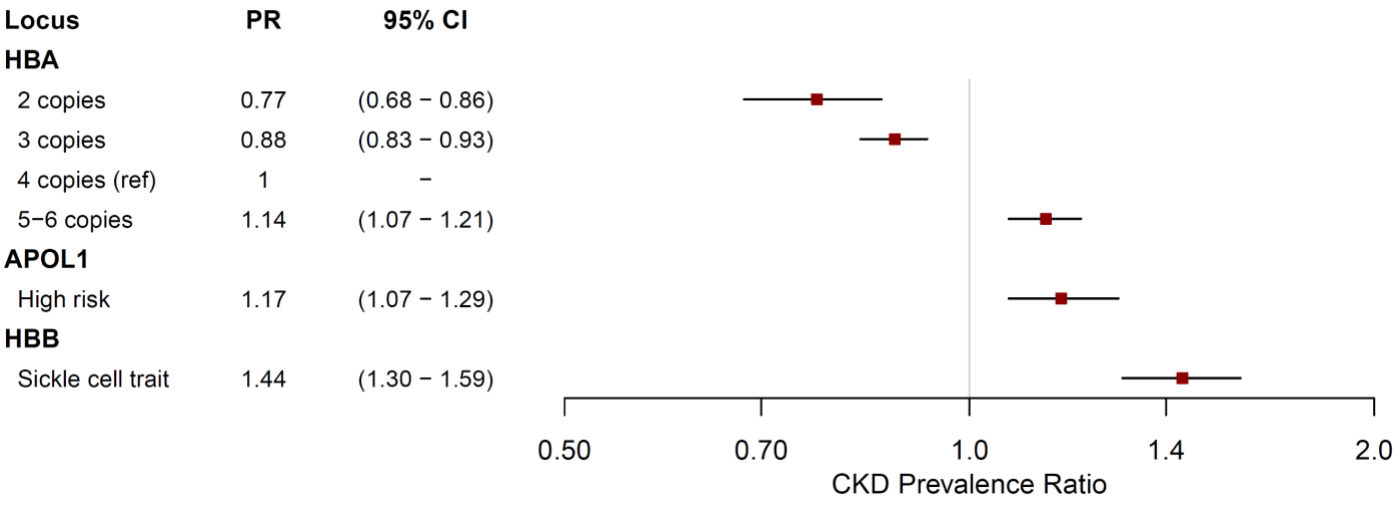
Alpha globin gene copy number is associated with prevalent chronic kidney disease and incident end-stage kidney disease among Black Americans

A. Parker Ruhl, MD, MHS[1,2], Neal Jeffries, PhD[3], Yu Yang, MD[4], Rakhi P. Naik, MD, MHS[5], Amit Patki, MS[6], Lydia H. Pecker, MD[5], Bryan T. Mott, PhD[7], Neil A. Zakai, MD, MSc[8,9], Cheryl A. Winkler, PhD[10], Jeffrey B. Kopp, MD[11], Leslie A. Lange, PhD[12], Marguerite R. Irvin, PhD, MS[13], Orlando M. Gutierrez, MD, MMSc[13,14], Mary Cushman, MD, MSc[8,9], Hans C. Ackerman, MD, DPhil, MSc[1]

[1]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD; [2]Pulmonary Branch, National Heart, Lung, and Blood Institute, Bethesda, MD; [3]Office of Biostatistics Research, National Heart, Lung, and Blood Institute, Bethesda, MD; [4]Division of Blood Diseases and Resources, National Heart, Lung, and Blood Institute, Rockville, MD; [5]Division of Hematology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland; [6]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama; [7]University of Alabama at Birmingham School of Medicine, Birmingham, Alabama; [8]Department of Medicine, Larner College of Medicine at the University of Vermont, Burlington, Vermont; [9]Department of Pathology & Laboratory Medicine, Larner College of Medicine at the University of Vermont, Burlington, Vermont; [10]Basic Research Program, National Cancer Institute, Frederick National Laboratory for Cancer Research, Frederick, Maryland; [11]Kidney Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland; [12]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado, Denver, CO; [13]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, Alabama; [14]Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama

I. Supplement Figure and Tables

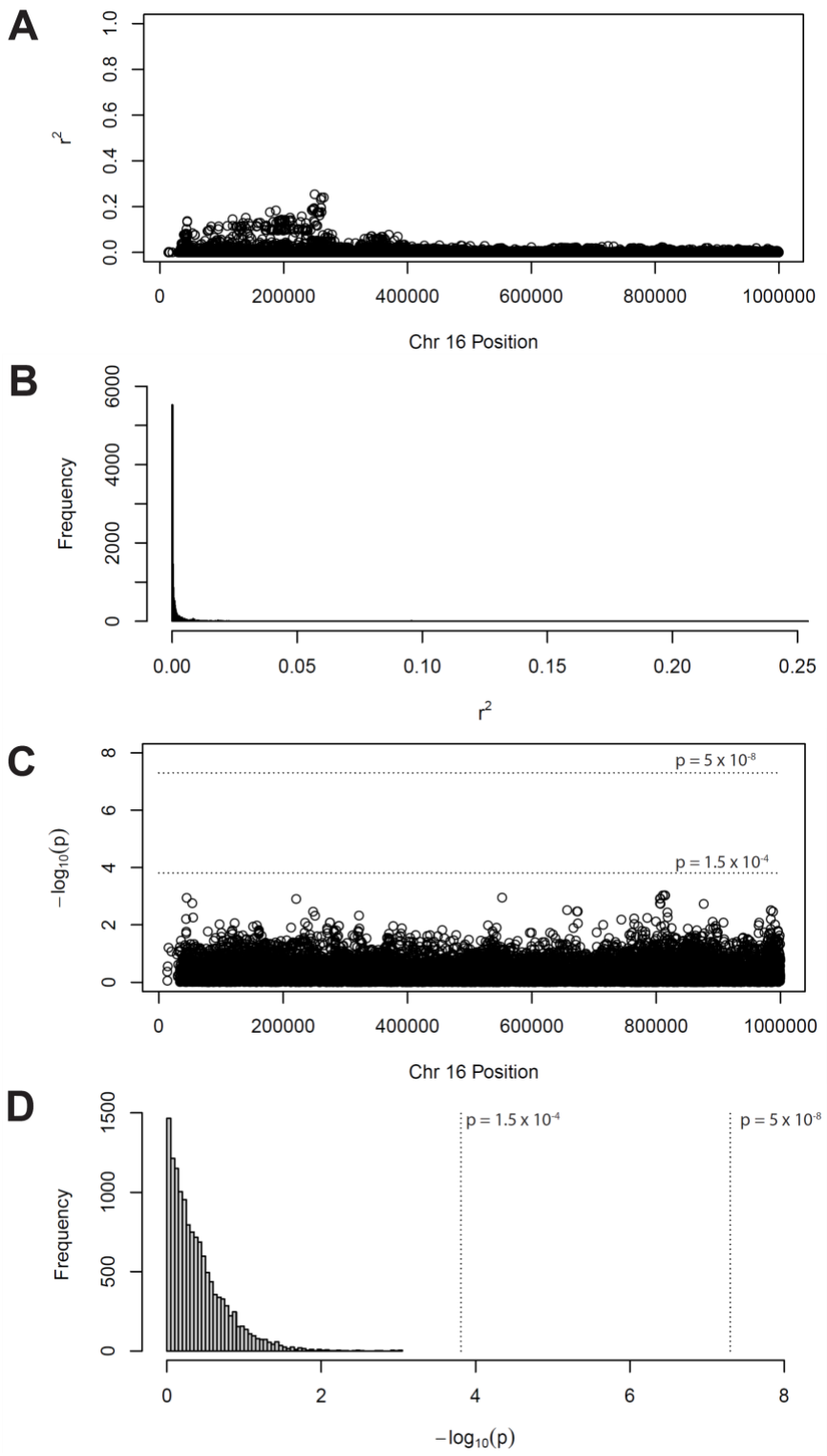
Supplement Figure 1. Comparison of prevalence ratios for *HBA*, *APOL1*, and *HBB* genetic risk factors for CKD



HBA= alpha globin gene; *APOL1*= apolipoprotein-1; *HBB*= hemoglobin beta-sickle cell trait; CKD= chronic kidney disease; PR= prevalence ratio

Supplemental Figure 1 Legend. Prevalence ratio of chronic kidney disease by alpha globin *HBA* copy number, *APOL1*, and sickle cell trait *HBB-SCT* –adjusted analysis. Prevalence ratios of genetic risk factors for chronic kidney disease represented on a log scale. Depicted prevalence ratios for each *HBA* copy number were calculated based on the reported modified Poisson multivariable regression model adjusting for 13 risk factors for chronic kidney disease.

Supplement Figure 2. Linkage disequilibrium and association analysis of sequence variants in the first 1 Mb of chromosome 16 flanking *HBA1* and *HBA2*.



Supplement Figure 2. A, Pairwise linkage disequilibrium (r^2) between the -3.7 kb structural variant and SNPs genotyped on the Infinium Expanded Multi-Ethnic Genotyping Array in 8,841 study participants. The deletion is found at position 173619-177403. **B**, Histogram showing the frequency distribution of r^2 values. **C**, Association between each SNP and CKD prevalence using a model that includes the age, sex, and the first 4 principal components of ancestry. Horizontal lines indicate the p-value significance thresholds for genome-wide significance (5×10^{-8}) and for a 1 Mb region (1.5×10^{-4}). **D**, Histogram showing the frequency distribution of association tests for all SNPs in the 1 Mb region of chromosome 16.

Supplement Table 1. Association of *HBA* copy number with CKD prevalence, incident reduced eGFR, and incident ESKD – fully adjusted models including ten principal components of ancestry.

	Chronic kidney disease prevalence*		Incident reduced eGFR†		Incident end-stage kidney disease‡	
	Modified Poisson (n=7,635)		Modified Poisson (n=2,970)		Cox proportional hazards (n=7,633)	
	PR	**CI**	**RR**	**CI**	**HR**	**CI**
***HBA* copy number, per gene copy**	1.15	(1.08 - 1.23)	1.07	(0.93 - 1.23)	1.33	(1.09 - 1.63)
PCA‖						
PCA 1	1.02	(0.98 - 1.06)	0.98	(0.90 - 1.05)	0.93	(0.82 - 1.05)
PCA 2	0.98	(0.95 - 1.02)	1.02	(0.95 - 1.10)	1.12	(1.00 - 1.26)
PCA 3	1.01	(0.97 - 1.04)	0.95	(0.88 - 1.03)	0.94	(0.84 - 1.06)
PCA 4	1.01	(0.98 - 1.05)	1.00	(0.93 - 1.08)	0.99	(0.88 - 1.08)
PCA 5	1.02	(0.99 - 1.06)	0.98	(0.90 - 1.06)	0.96	(0.86 - 1.08)
PCA 6	1.01	(0.97 - 1.05)	1.00	(0.92 - 1.07)	0.87	(0.78 - 0.98)
PCA 7	0.96	(0.93 - 1.00)	1.05	(0.97 - 1.13)	0.97	(0.86 - 1.09)
PCA 8	1.01	(0.97 - 1.05)	1.07	(0.99 - 1.15)	1.02	(0.91 - 1.15)
PCA 9	0.98	(0.94 - 1.01)	0.98	(0.91 - 1.06)	0.94	(0.84 - 1.05)
PCA 10	1.01	(0.98 - 1.05)	0.94	(0.91 - 1.06)	0.94	(0.84 - 1.06)

HBA= alpha globin gene; CKD= chronic kidney disease; eGFR= estimated glomerular filtration rate; ESKD= end-stage kidney disease; PR= prevalence ratio; CI= 95% confidence interval; RR= relative risk; HR = hazard ratio; PCA= principal components of ancestry.

*CKD prevalence was defined by eGFR < 60mL/min/1.73m^2 or urine albumin to creatinine ratio ≥ 30mg/g. †Incident reduced eGFR was defined by an eGFR < 60mL/min at the follow-up in-

home visit and greater than 40% decline in eGFR from baseline, among those who had eGFR ≥ 60 mL/min at baseline. ‡Incident end-stage kidney disease (ESKD) was identified by linkage to the United States Renal Data System data through December 31, 2018. §P values were calculated using either modified Poisson or Cox proportional hazards multivariable regression models, as noted, employing a linear effect of *HBA* allele count; ‖8,841 participants had GWAS data available for PCA analysis. The CKD prevalence model had n=7,635 subjects with data available for the PCA analysis, the incident reduced eGFR model had n=2,970 subjects, and the incident ESKD model had n=7,633 subjects. Multiple imputations were performed for other missing data. The following variables were included in the model but are not displayed in this table: Sickle cell trait, *APOL1* high-risk status, hemoglobin, age, sex, body mass index, hypertension, diabetes mellitus, smoking status, medically insured, region, education level, income.

Supplement Table 2. Pre-specified tests for interaction between *HBA* copy number and SCT on the outcomes of CKD prevalence, incident reduced eGFR, and incident ESKD in fully adjusted models.

	Chronic kidney disease prevalence*			Incident reduced eGFR†			Incident end-stage kidney disease‡		
	Modified Poisson (n=9,908)			Modified Poisson (n=3,733)			Cox proportional hazards (n=9,905)		
	PR	**CI**	***P value*§**	**RR**	**CI**	***P value*§**	**HR**	**CI**	***P value*§**
SCT*_HBA_	1.13	(0.93 - 1.38)	0.22	0.97	(0.61 - 1.56)	0.90	1.62	(0.80 – 3.28)	0.18

SCT= sickle cell trait; *HBA*= alpha globin gene; CKD= chronic kidney disease; eGFR= estimated glomerular filtration rate; ESKD= end-stage kidney disease; PR= prevalence ratio; CI= 95% confidence interval; RR= relative risk; HR= hazard ratio

*Chronic kidney disease was defined by estimated glomerular filtration rate [GFR] <60mL/min/1.73m^2 or urine albumin to creatinine ratio $\geq$30mg/g. †Incident reduced eGFR was defined by an eGFR < 60mL/min at the follow-up in-home visit and greater than 40% decline in eGFR from baseline, among those who had eGFR ≥ 60 mL/min at baseline. ‡Incident end-stage kidney disease (ESKD) was identified by linkage to the United States Renal Data System data through December 31, 2018. §P values were calculated using either modified Poisson or Cox proportional hazards multivariable regression models, as noted, employing a monotonic effect of *HBA* allele count. Multiple imputations were performed for missing data. The following variables were included in the model but are not displayed in this table: *HBA* copy number, sickle cell trait, *APOL1* high-risk status, hemoglobin, age, sex, body mass index, hypertension, diabetes mellitus, smoking status, medically insured, region, education level, income.

Supplement Table 3. Pre-specified tests for interaction between each of Age, Sex, Hypertension, or *APOL1* and *HBA* on the outcome of CKD prevalence in fully adjusted models.

Separate fully adjusted models with interaction terms individually added	Chronic kidney disease prevalence[†] Modified Poisson (n=9,908)		
	PR	**CI**	***P value***[‡]
Age**HBA***	1.00	(0.99 - 1.01)	0.80
Female Sex**HBA***	0.94	(0.84 - 1.06)	0.29
Hypertension**HBA***	0.97	(0.79 - 1.18)	0.74
APOL1* high-risk genotype*HBA***	0.95	(0.82 - 1.11)	0.54

APOL1= apolipoprotein-L1; *HBA*= alpha globin; CKD= chronic kidney disease; PR= prevalence ratio; CI= 95% confidence interval.

[†]CKD prevalence was defined by estimated glomerular filtration rate <60mL/min and/or urine albumin creatinine ratio $\geq$ 30 mg/g. [‡]Each interaction term was added separately to the main CKD prevalence model. P values were calculated using modified Poisson multivariable regression models employing a linear effect of *HBA* allele count on the log of the relative risk with all clinical and demographic covariates utilized in the main CKD prevalence model. The following variables were included in the model but not displayed in this table: *HBA* copy number, sickle cell trait, *APOL1* high-risk status, hemoglobin, age, sex, hypertension, diabetes mellitus, body mass index, smoking status, medically insured, region, education level, income.

Supplement Table 4. Comparison of participants with and without second visit data

	Second visit data available?*	
	Yes	**No**
Participants, N (%)	3,733 (43%)	4,977 (57%)
***HBA* copy number**		
Two *HBA* copies	152 (4%)	202 (4%)
Three *HBA* copies	1024 (27%)	1406 (28%)
Four *HBA* copies	2522 (68%)	3318 (67%)
Five or Six *HBA* copies	35 (1%)	51 (1%)
Age, years	62 (57, 68)	63 (57, 71)
Female sex, N (%)	2362 (63%)	2981 (60%)
Body mass index, kg/m^2	30.0 (26.5, 34.4)	29.6 (25.9, 34.3)
Smoking status, N (%)		
Never	1784 (48%)	2107 (43%)
Past	1392 (37%)	1818 (37%)
Current	541 (15%)	1025 (21%)
Region, N (%)		
Non-Belt	1756 (47%)	2543 (51%)
Belt	1287 (34%)	1623 (33%)
Buckle	690 (18%)	811 (16%)
Medically insured, N (%)	3367 (90%)	4420 (89%)
Education level, N (%)		
Less than high school	474 (13%)	1119 (23%)
High school graduate	1018 (27%)	1401 (28%)
Some college	1034 (28%)	1348 (27%)
College graduate or more	1206 (32%)	1100 (22%)
Income, N (%)		
≤ $20K	795 (24%)	1466 (34%)
$20K - $34K	930 (27%)	1319 (31%)
$35K - $74K	1161 (34%)	1149 (27%)
≥ $75K	498 (15%)	364 (8%)

Hemoglobin, per g/dL	13.2 (12.4, 14.0)	13.1 (12.3, 14.1)
RBC count, million cells/cmm	4.5 (4.2, 4.8)	4.5 (4.2, 4.8)
MCV, fL	88 (84, 92)	88 (84, 92)
MCH, pg	29.6 (27.9, 30.9)	29.6 (28.0, 31.0)
MCHC, g/dL	33.5 (32.9, 33.9)	33.4 (32.9, 34.0)
RDW-CV, %	13.9 (13.3, 14.7)	14.0 (13.3, 14.8)
Hypertension, N (%)	2558 (72%)	3668 (77%)
Diabetes mellitus, N (%)	870 (23%)	1467 (30%)
Sickle cell trait, N (%)	269 (7%)	356 (7%)
***APOL1* high-risk, N (%)**	465 (13%)	625 (13%)

HBA= alpha globin gene; 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; eGFR= estimated glomerular filtration rate; *APOL1*= apolipoprotein-L1. Values are median (25^{th}, 75^{th} percentile) except where otherwise indicated.

*Participants with missing eGFR at baseline (n=65) and baseline eGFR < 60 were excluded from this analysis (n=1,133).

Supplement Table 5. Post-hoc sensitivity analysis of the association of *HBA* copy number with CKD prevalence when hemoglobin is omitted from the model.

	Chronic kidney disease prevalence* (*n*=9,908)	
	Modified Poisson	
	PR	**CI**
***HBA* copy number, per gene copy**	1.10	(1.03 - 1.16)
Sickle cell trait	1.47	(1.32 - 1.62)
***APOL1* high-risk**	1.17	(1.07 - 1.29)
Age, per year	1.03	(1.03 - 1.04)
Female sex	0.89	(0.83 - 0.96)
Body mass index†	1.09	(1.05 - 1.12)
Hypertension	1.88	(1.68 - 2.10)
Diabetes mellitus	1.73	(1.61 - 1.85)
Smoking status		
Never (ref)	-	-
Past	0.97	(0.90 - 1.04)
Present	1.23	(1.11 - 1.35)
Medically Insured	0.92	(0.81 - 1.04)
Region		
Non-Belt (ref)	-	-
Belt	0.98	(0.91 - 1.06)
Buckle	1.02	(0.93 - 1.12)

Education level		
< HS Grad (ref)	-	-
HS Grad	0.93	(0.84 - 1.01)
Some College	0.97	(0.88 - 1.06)
≥ College Grad	0.90	(0.84 - 1.02)
Income		
< $20K (ref)	-	-
$20K - $34K	0.98	(0.90 - 1.06)
$35K - $74K	0.92	(0.84 - 1.02)
≥ $75K	0.73	(0.62 - 0.88)

HBA= alpha globin gene; CKD= chronic kidney disease; PR= prevalence ratio; CI= 95% confidence interval; APOL1= apolipoprotein-L1; (ref) indicates reference category; HS= high school; K= thousand.

*CKD prevalence was defined by a urine albumin to creatinine ratio ≥ 30 mg/g or an estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73m2 at baseline. Analysis performed using modified Poisson multivariable regression model employing a linear effect of *HBA* allele count on the log of the prevalence ratio. †Body mass index scaled by standard deviation. All variables shown in table were included in the multivariable model. Multiple imputations were performed for missing data. The number of subjects available for this analysis was 9,908.

Supplement Table 6. Post-hoc sensitivity analysis of the association of categorical *HBA* copy number with CKD prevalence

	Chronic kidney disease* (*n*=9,908) Modified Poisson		
	PR	**CI**	**P value†**
***HBA* copy number**			
2 *HBA* copies	0.74	(0.61 - 0.90)	0.003
3 *HBA* copies	0.90	(0.83 - 0.97)	0.004
4 *HBA* copies (ref)	-	-	
≥ 5 *HBA* copies	1.21	(0.95 - 1.56)	0.13
Sickle cell trait	1.43	(1.29 - 1.59)	< 0.0001
***APOL1* high-risk**	1.17	(1.07 - 1.29)	0.0007
Hemoglobin, per 1 g/dL	0.90	(0.87 - 0.92)	< 0.0001
Age, per year	1.03	(1.03 - 1.04)	< 0.0001
Female sex	0.78	(0.72 - 0.84)	< 0.0001
Body mass index‡	1.08	(1.05 - 1.12)	< 0.0001
Hypertension	1.86	(1.66 - 2.07)	< 0.0001
Diabetes mellitus	1.65	(1.54 - 1.76)	< 0.0001
Smoking status			
Never (ref)	-	-	
Past	0.98	(0.91 - 1.05)	0.52

Present	1.29	(1.17 - 1.43)	< 0.0001
Medically Insured	0.91	(0.80 - 1.04)	0.15
Region			
Non-Belt (ref)	-	-	
Belt	0.97	(0.90 - 1.05)	0.46
Buckle	1.00	(0.91 - 1.10)	0.97
Education level			
< HS Grad (ref)	-	-	
HS Grad	0.94	(0.86 - 1.03)	0.16
Some College	0.98	(0.90 - 1.08)	0.74
≥ College Grad	0.92	(0.83 - 1.02)	0.11
Income			
< $20K (ref)	-	-	
$20K - $34K	0.98	(0.90 - 1.07)	0.65
$35K - $74K	0.93	(0.84 - 1.02)	0.13
≥ $75K	0.75	(0.63 - 0.90)	0.002

HBA= alpha globin gene; PR= prevalence ratio; CI= 95% confidence interval; *APOL1*= apolipoprotein-L1; (ref) indicates reference category; HS= high school; K= thousand.

*CKD prevalence was defined by a urine albumin to creatinine ratio ≥ 30 mg/g or an estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73m2 at baseline. †Analysis was performed using modified Poisson multivariable regression model employing a categorical effect of *HBA*

allele count on the log of the prevalence ratio. ‡Body mass index was scaled by standard deviation. All shown variables in the table were included in multivariable model. Multiple imputations were performed for missing data. The number of participants available for this analysis was 9,908.

CKD	**Quantitative Analysis of *HBA* copy number**		**Categorical analysis of *HBA* copy number**	
	CKD PR	95% CI	CKD PR	95% CI
HBA 2	0.77	0.68 - 0.86	0.74	0.61 - 0.90
HBA 3	0.88	0.83 - 0.93	0.90	0.83 - 0.97
HBA 4	(reference)	-	(reference)	-
HBA 5-6	1.14	1.07 - 1.21	1.21	0.95 - 1.56

Supplement Table 7. Post-hoc sensitivity analysis of the association of categorical *HBA* copy number with ESKD incidence

	Incident end-stage kidney disease* (*n* = 9,905) Cox proportional hazards		
	HR	**CI**	**P value**[†]
***HBA* copy number**			
2 *HBA* copies	0.58	(0.32 - 1.03)	0.06
3 *HBA* copies	0.71	(0.55 - 0.92)	0.009
4 *HBA* copies (ref)	-	-	
≥ 5 *HBA* copies	0.62	(0.15 - 2.51)	0.50
Sickle cell trait	1.84	(1.33 - 2.55)	0.0002
***APOL1* high risk**	1.72	(1.31 - 2.27)	0.0001
Hemoglobin, per 1 g/dL	0.61	(0.55 - 0.67)	< 0.0001
Age, per year	1.00	(0.99 - 1.02)	0.62
Female sex	0.41	(0.31 - 0.54)	< 0.0001
Body mass index[‡]	1.05	(0.93 - 1.18)	0.45
Hypertension	4.01	(2.49 – 6.44)	< 0.0001
Diabetes mellitus	3.41	(2.70 - 4.31)	< 0.0001
Smoking status			
Never (ref)	-	-	
Past	1.19	(0.93 - 1.52)	0.17
Present	1.94	(1.40 - 2.68)	< 0.0001
Medically Insured	0.91	(0.62 - 1.36)	0.65

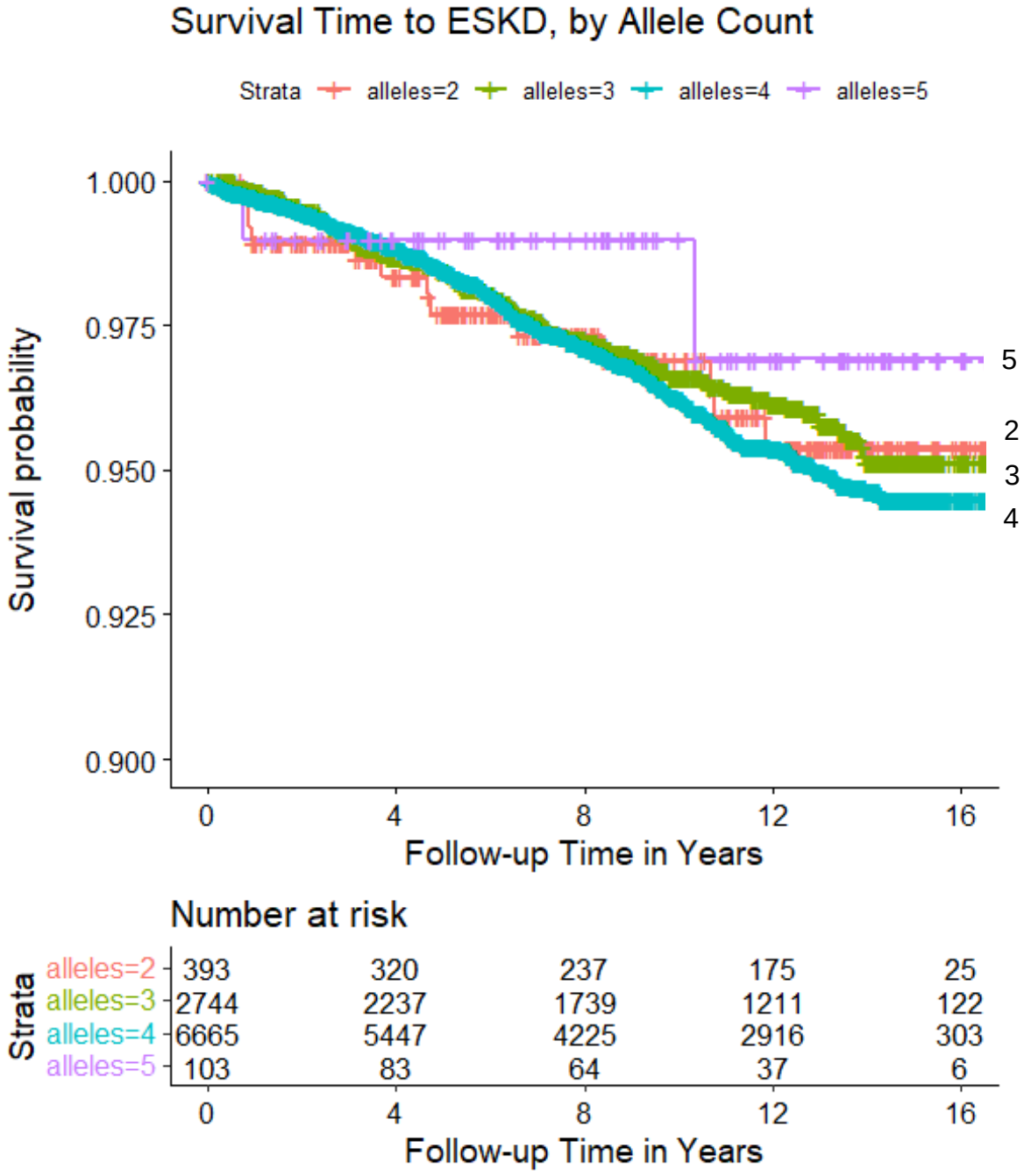
Region			
Non-Belt (ref)	-	-	
Belt	0.85	(0.66 - 1.09)	0.19
Buckle	0.97	(0.72 - 1.30)	0.82
Education level			
< HS Grad (ref)	-	-	
HS Grad	1.24	(0.90 - 1.70)	0.18
Some College	1.04	(0.74 - 1.47)	0.82
≥ College Grad	1.30	(0.89 - 1.89)	0.17
Income			
< $20K (ref)	-	-	
$20K - $34K	1.03	(0.77 - 1.38)	0.82
$35K - $74K	0.70	(0.49 - 1.01)	0.05
≥ $75K	0.42	(0.22 - 0.80)	0.008

HR=hazard ratio; CI= 95% confidence interval; *HBA*= alpha globin gene; *APOL1*= apolipoprotein-L1; (ref) indicates reference category; HS= high school; K= thousand.

*Incident end-stage kidney disease was identified by linkage to the United States Renal Data System through December 31, 2018. †Analysis was performed using Cox proportional hazards multivariable regression employing a categorial effect of *HBA* allele count on the log of the hazard ratio. ‡Body mass index was scaled by standard deviation. All variables in the table were included in the multivariable model. Multiple imputations were performed for missing data. The number of participants available for this analysis was n=9,905.

ESKD	**Quantitative Analysis of *HBA* copy number**		**Categorical analysis of *HBA* copy number**	
	ESKD HR	95% CI	ESKD HR	95% CI
HBA 2	0.57	0.39 - 0.84	0.58	0.32 - 1.03
HBA 3	0.76	0.62 - 0.92	0.71	0.55 - 0.92
HBA 4	(reference)	-	(reference)	-
HBA 5-6	1.32	1.09 - 1.61	0.62	0.15 - 2.15

Supplement Figure 3. Unadjusted survival curves of ESKD onset by genotype



Kaplan-Meier curves showing time to ESKD events. Survival probability indicates the probability of being ESKD-free. The time to event is defined as the number of years between the initial in-home interview date and date D where D is the minimum of last follow-up time provided by REGARDS or the date when the individual had onset of ESKD as indicated by USRDS. The decision to censor ESKD events occurring after the end of REGARDS follow-up avoids bias associated with having extended follow-up only for those having an ESKD event.

II. Estimation of population preventable fraction of *HBA* copy number on kidney disease

The observed population had the following allele frequencies:

HBA copy number	2	3	4	5 or 6
Prevalence, %	4.0	27.7	67.3	1.0

The population preventable fraction measures the degree to which the elimination of a beneficial risk factors in a population increases the corresponding population prevalence or risk of disease. The population attributable fraction measures the degree to which elimination of a detrimental risk factor decreases the population prevalence or risk.

Chronic Kidney Disease Prevalence

To estimate the fraction of disease prevalence that is prevented by reductions in *HBA* copy number, we first dichotomized the *HBA* copy number into two classes: (2, 3) versus (4, 5, 6) and asked to what extent is chronic kidney disease (CKD) prevalence decreased by the (2, 3) allele class.

We used a fully-adjusted Poisson regression model to estimate the risk associated with the (4, 5, 6) class versus the (2, 3) class:

Variables	**PR**	**CI**	**P value**
HBA copy number (4, 5, 6)	1.14	(1.06 - 1.23)	0.0003
Sickle cell trait	1.44	(1.30 - 1.59)	< 0.0001
APOL1 high risk	1.17	(1.07 - 1.29)	0.0008
Hemoglobin, per 1g/dL	0.90	(0.87 - 0.93)	< 0.0001
Age, per year	1.03	(1.03 - 1.04)	< 0.0001
Female sex	0.78	(0.72 - 0.85)	< 0.0001
Body mass index	1.09	(1.05 - 1.12)	< 0.0001
Hypertension	1.86	(1.66 - 2.07)	< 0.0001
Diabetes mellitus	1.65	(1.54 - 1.77)	< 0.0001
Smoking, past	0.98	(0.91 - 1.05)	0.50

Smoking, present	1.29	(1.17 - 1.43)	< 0.0001
Medically insured	0.91	(0.81 - 1.03)	0.15
Region belt	0.97	(0.90 - 1.05)	0.43
Region buckle	1.00	(0.91 - 1.09)	0.98
Education level high school graduate	0.94	(0.86 - 1.03)	0.16
Education level some college	0.99	(0.90 - 1.09)	0.74
Education level college graduate or more	0.92	(0.82 - 1.02)	0.10
Income $20k-$34k	0.98	(0.90 - 1.07)	0.63
Income $35k-$74k	0.92	(0.82 - 1.02)	0.13
Income $75k and above	0.75	(0.63 - 0.90)	0.001

Using methods based on Greenland and Drescher adapted to modified Poisson regression and multiple imputations we then estimated the population preventable fraction (PPF) as the increased prevalence of CKD that would exist in an alternative REGARDS population that has only the (4, 5, 6) allele class:

$$\mathrm{PPF} = \frac{\mathrm{Risk}_{\text{Alternative Population}} - \mathrm{Risk}_{\text{Population Observed}}}{\mathrm{Risk}_{\text{Population Observed}}}$$

The prevalence of CKD in the alternative REGARDS population would increase by 4.3% (95% CI 2.0 - 6.5) if the protective *HBA* allele classes (2,3) were absent from the population.

For comparison, a similar calculation of population attributable fraction (PAF) was performed for sickle cell trait (PAF was computed instead of the population preventable fraction because sickle cell trait is harmful, rather than beneficial). The prevalence of CKD in an alternative REGARDS population that does not carry sickle cell trait would decrease by 3.1 % (95% CI 2.1 - 4.2).

In conclusion, a reduction in *HBA* copy number explains a non-zero fraction of CKD risk that is similar in size to that attributable to sickle cell trait. While the population preventable fraction point estimate for *HBA* deletions is greater than the population attributable fraction for sickle cell trait (4.5% vs 3.1%), the 95% confidence intervals for these estimates overlap.

Incident End-Stage Chronic Kidney Disease

To estimate the fraction of disease risk that is attributable to deletions in *HBA* copy number, we first dichotomized the *HBA* copy number into two classes: (2, 3) versus (4, 5, 6).

We used a fully-adjusted Cox proportional hazards model to estimate the risk associated with the (4, 5, 6) class versus the (2, 3) class:

Variables	HR	CI	P value
HBA copy number (4, 5, 6)	1.44	(1.13 - 1.83)	0.003
Sickle cell trait	1.86	(1.35 – 2.58)	0.0002
APOL1 high risk	1.72	(1.31 - 2.27)	0.0001
Hemoglobin, per 1g/dL	0.61	(0.55 - 0.67)	< 0.0001
Age, per year	1.00	(0.99 - 1.02)	0.63
Female sex	0.41	(0.32 - 0.54)	< 0.0001
Body mass index	1.05	(0.93 - 1.18)	0.45
Hypertension	4.00	(2.49 – 6.42)	< 0.0001
Diabetes mellitus	3.42	(2.71 - 4.32)	< 0.0001
Smoking, past	1.19	(0.93 - 1.18)	0.45
Smoking, present	1.93	(1.40 – 2.67)	< 0.0001
Medically insured	0.91	(0.62 - 1.36)	0.66
Region belt	0.84	(0.66 - 1.08)	0.19
Region buckle	0.96	(0.72 - 1.29)	0.80
Education level high school graduate	1.24	(0.90 -1.71)	0.18
Education level some college	1.04	(0.74 - 1.47)	0.82
Education level college graduate or more	1.30	(0.89 – 1.89)	0.17
Income $20k-$34k	1.03	(0.77 - 1.38)	0.83

Income $35k-$74k	0.70	(0.49 – 1.01)	0.05
Income $75k and above	0.42	(0.22 - 0.80)	0.0008

Finally, applying methods of Zetterqvist et al. in a setting with multiple imputations we estimated the population preventable fraction (PPF) for the 2,3 class as the increased risk of incident end-stage kidney disease (ESKD) at follow-up time *t* that would exist in an alternative REGARDS population that had only the (4, 5, 6) allele class:

$$\text{PPF} = \frac{\text{Prob(ESKD by time t)}_{\text{Alternative Population}} - \text{Prob(ESKD by time t)}_{\text{Population Observed}}}{\text{Prob(ESKD by time t)}_{\text{Population Observed}}}$$

The risk of incident ESKD in the alternative REGARDS population would increase by 11.5%, (95% CI 4.0 – 18.9) if the protective *HBA* alleles (2,3) were absent from the population. This was estimated at the median follow-up time of 10.1 years, and calculations at the 25th and 75th percentile of follow-up time gave similar results.

For comparison, a similar calculation was performed for sickle cell trait. The risk of incident ESKD at 10.1 years of follow-up in an alternative REGARDS population that does not carry sickle cell trait would decrease by 5.6% (95% CI 2.1 - 9.1).

In conclusion, a reduction in alpha globin gene copy number explains a non-zero fraction of ESKD risk that is similar in size to that attributable to sickle cell trait. While the PPF point estimate for *HBA* deletions is greater than the PAF for sickle cell trait (11.5% vs 5.6%), the 95% confidence intervals for these estimates overlap.

III. Additional Methods

a. *HBA* Genotyping Methods

Two-dimensional clusters of droplet counts for target and reference genes were manually gated using Quantasoft (Bio-Rad) per the manufacturer's protocols. Droplet counts, copy number variant (CNV) values, and 95% CIs for CNV were extracted, visualized, and genotype was assigned using custom scripts in the R computing environment without user intervention. A subset of samples was validated against an independent approach employing multiple ligation-dependent probe amplification (MLPA) performed at the Mayo Clinic Laboratory, with 100% concordance. Inter-day variation of our assay was determined by performing the assay on two different days on 672 samples; quantitative copy number varied by less than 1% between days. Reference samples of known genotype were run as positive controls and reaction wells with water instead of DNA were run as negative controls each day.

b. APOL1 Genotyping Methods

APOL1 high-risk genotype was defined as the presence of two renal risk alleles compared to a reference of less than two high-risk alleles (see Supplemental Material) (David, et. al, Kidney Int Rep 2019, Kopp et. al, J Am Soc Nephrol 2011). Two single nucleotide polymorphisms (SNPs) in APOL1 (rs73885319 and rs60910145) and a six base-pair insertion/deletion polymorphism (rs71785313) were genotyped in REGARDS participants using TaqMan custom assays: rs73885319 (Assay ID-AH20SD1), rs60910145 (Assay ID-AHWR1JA), and RS71785313 (Assay ID-AH1RT7T) (ThermoFisher Scientific (Waltham, MA) (PMID: 21997394) (PMID: 30596185).

c. End-stage kidney disease Time-to-event Analysis Methods

The time to event is defined as the number of years between the initial in-home interview date and date D where D is the minimum of D1 and D2 which are defined as follows. D1 is the last follow-up date provided by REGARDS as the last time the participant was contacted for status. The outcome associated with this date is either "No Event" or "Death". D2 is the date when the individual had onset of ESKD as indicated by USRDS. 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 are recorded with event "ESKD" and D = D2. Individuals with D2 after D1 had ESKD onset after their last REGARDS follow-up and they are recorded with "No Event" and censored at date D = D1. This decision to censor ESKD 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 an ESKD event.

d. Multiple Imputation Procedure

Multiple imputation methods were used in the multivariable analyses. Data on the degree of missingness for the outcome of CKD prevalence are described in the footnote to Table 1 in the manuscript. Missing patterns for incident reduced eGFR and incident end-stage kidney disease were similar to those detailed for the CKD prevalence analysis. The R package "mice" Version 3.6.0 was used to create and analyze the resulting imputations (Zetterqvist J, et. al, European Journal of Epidemiology 2016). Each analysis presented is based upon 20 imputations (each developed using 30 Markov Chain based iterations) and the final model coefficients and their standard errors were derived using Rubin's method for pooling results across imputations (Greenland S, et. al, Biometrics 1993). The variables used in the

imputation procedure were those used in the corresponding regression model and sometimes augmented with additional variables (e.g. systolic and diastolic blood pressure for determining hypertension, and albumin creatinine ratio and eGFR for CKD prevalence). The imputations evolution over 30 iterations was examined visually for convergence and mixing. Further, the distributions of complete and imputed values were visually examined for aberrations.

e. Diagnostic Modeling Description

Our modeling was prespecified in our analytic plan, as described. We performed diagnostic investigation of the Poisson models (for CKD prevalence and incidence reduced eGFR) and Cox models (for ESKD incidence) in order to guide sensitivity analyses. For the Poisson models involving CKD prevalence and CKD incidence the R function “glm” and R package “sandwich” were used. Residuals were examined for evidence of poor fitting as evidenced by correlation between residuals and predictors of fitted values. Testing of Pearson residuals indicated that age and hemoglobin were perhaps inadequately modeled as having linear relationships on the log of the risk for CKD prevalence. Consequently, we extended our main model to include quadratic terms for age and hemoglobin. These additional terms had significant p-values but did not change the results for allele count or sickle cell trait in any meaningful way (point estimates of the risk ratio changed from 1.14 to 1.13 and 1.44 to 1.43 and both p-values remained < 0.0001). Conducting a similar analysis for incident reduced eGFR suggested the effect of age on log of risk was not adequately captured through a linear relationship and a quadratic term was added. The estimated effect for allele count was not qualitatively changed and remained non-significant (risk ratio estimate of 1.02, $p = 0.73$). For both CKD prevalence and

reduced eGFR incidence Pearson residuals were not significantly correlated with allele count.

For the model of ESKD 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 ESKD outcomes. There was some evidence of proportional hazards violation for the baseline age covariate ($p < 0.05$ for a chi-square test of non-proportional hazards in the majority of imputed datasets), however including a time by age interaction to address non-proportionality did not qualitatively alter the estimated effects for the other covariates – in particular the p-values for *HBA*, sickle cell trait, and hemoglobin remained significant and of similar magnitude as without the interaction.

f. Assessment of the Missing at Random Assumption

Missingness was generally rare with the exception of hemoglobin (32%), self-reported income (12%) and hypertension (4%). Hemoglobin is missing primarily because it was not initially collected for approximately the first 8000 of the REGARDS 30239 participants (all races combined). Given the administrative nature of the missing data an assumption of hemoglobin missing at random (i.e., the probability of missing depends on observed information rather than the underlying missing hemoglobin value) multiple imputation was used.

Income data reported as missing reflect refusal to provide information. These self-reported incomes might not be missing at random as the refusals might more likely coincide with higher or lower than average incomes. As a sensitivity analysis we first imputed the annual income category (either "less than $20k", "$20k-$34k", "$35k-$74k", or "$75k and above") using the multiple imputation algorithm and then moved the imputed category values one level higher if they were not already in the

highest category. For example, if a person had an original imputed value of $20k-$34k then in this sensitivity analysis they would now have a value of $35k-$74k. This corresponds to people refusing to answer having higher incomes than predicted. While the education and income coefficients change marginally in the resulting analysis, the remaining coefficients and p-values are essentially unchanged from those presented in the incident ESKD analysis. The results when lowering (instead of raising) the imputed income category are qualitatively similar. These results suggest that using a missing at random assumption for income is reasonable.

A similar analysis was performed with respect to missing values for baseline hypertension (388 of 9,905) individuals in the analysis of incident ESKD). When all 388 were imputed as non-hypertensive the results changed very little for all covariates except hypertension (and this changed only modestly). The HR estimate and p-value for HBA copy number remain the same. When all 388 were imputed as hypertensive the results again changed hardly at all (HR estimate and p-value for HBA copy number remain the same), again suggesting that using a missing at random assumption for hypertension does not likely lead to misleading estimates for any of the covariates.

These sensitivity analyses were conducted for the time to ESKD analysis as the primary results had a larger p-value (0.005) for HBA copy number than the CKD prevalence p-value (<0.0001). Given the strength of the association for the analysis of CKD prevalence we again expect the use of multiple imputation with a missing at random assumption does not generate misleading results because of violations of this assumption. Similarly, we do not expect the non-significant findings for HBA copy number on incident CKD to be qualitatively changed by plausible modifications to the missing at random assumption.

IV. References

David VA, Binns-Roemer EA, Winkler CA. Taqman Assay for Genotyping CKD-Associated APOL1 SNP rs60910145: A Cautionary Note. Kidney Int Rep 2019;4(1):184–5.

Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics*, 1993, 49, p. 865-872.

Kopp JB, Nelson GW, Sampath K, et al. APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. J Am Soc Nephrol 2011;22(11):2129–37.

Zetterqvist J, Pawitan Y, Sjolander A, Dahlqwist E, Sjölander A. Model-based estimation of the attributable fraction for cross-sectional, case–control and cohort studies using the R package AF. *European Journal of Epidemiology*, 2016, Vol.31(6), p.575-582