Supplementary Appendix

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

Supplement to: Hsu C, Yang W, Parikh RV, et al. Race, genetic ancestry, and estimating kidney function in CKD. N Engl J Med 2021;385:1750-60. DOI: 10.1056/NEJMoa2103753

Table of Contents

Investigators and collaborators

Investigators: Chi-yuan Hsu, MD, MSc, Wei Yang, PhD, Rishi V. Parikh, MPH, Amanda H. Anderson, PhD, Teresa K. Chen, MD, MHS, Debbie L. Cohen, MD, Jiang He, MD, PhD, Madhumita J. Mohanty, MD, James P. Lash, MD, Katherine T. Mills, PhD, Anthony N. Muiru, MD, Afshin Parsa, MD, MPH, Milda R. Saunders, MD, MPH, Tariq Shafi, MBBS MHS, Raymond R. Townsend, MD, Sushrut S. Waikar, MD, MPH, Jianqiao Wang, MS, Myles Wolf, MD, MMSc, Thida C. Tan, MPH, Harold I. Feldman, MD, MSCE, Alan S. Go, MD

Collaborators: the CRIC Study Investigators also include: Lawrence J. Appel, MD, MPH; Jing Chen, MD, MMSc, MSc; Robert G. Nelson, MD, PhD, MS; Mahboob Rahman, MD; Panduranga S. Rao, MD; Vallabh O Shah, PhD, MS; Mark L. Unruh, MD, MS

Table S1. Baseline CRIC study visit measurement details.

Variable	Measurement
Race	Race and Hispanic ethnicity were determined by self-report. Following the form of current GFR estimating equations,[1,2] participants were classified as Black or non-Black.
Demographic characteristics	Information on self-reported age and sex were collected.
Genetic ancestry estimation	Genotyping was conducted using the Illumina HumanOmni1-Quad v1.0 microarray.[3] A general admixture model[4] was derived using individuals from the 1000 Genomes Project[5] as the reference data. A cluster size of five was selected based on previous studies and verified by comparing the log-likelihood of candidate models to the CRIC data. The five clusters correspond to the five super-populations in the reference data that include African, American, European, East Asian and South Asian. Each participant has individual ancestry percent estimates of five populations, whose sum is 100% (Further details are presented below in the **Supplementary Methods**).
Body Composition	We considered height, weight, body mass index and body surface area,[6] as well as bioelectrical impedance analysis (BIA)-derived measures of phase angle and fat-free muscle mass.[7,8]
Protein intake	Dietary protein intake was assessed using the Diet History Questionnaire[9] and using 24-hour urine quantified urine urea nitrogen and protein along with body weight.[10,11]
Creatinine Production and extra-renal elimination	We measured 24-hour urine creatinine excretion which reflected the balance between creatinine production (both endogenous and exogenous) and extra-renal elimination.
Tubular secretion of creatinine	Tubular secretion of creatinine was quantified using the ratio and the absolute difference between 24-hour urine creatinine clearance and measured GFR.
Serum creatinine	SCr was measured using an enzyme-based assay on the Hitachi Vitros 950 AT (coefficient of variation [CV], 1.1%),[12] with calibration traceable to an IDMS reference measurement procedure.
Serum cystatin C	Serum cystatin C concentration was determined using a particle-enhanced immunonephelometric assay on the Siemens BN™ II System (CV 4.9%). An internal cystatin C standardization was implemented to correct for drift over time across different calibrator lots and reagent lots.[12]
Measured glomerular filtration rate	GFR was measured using urinary ^{125}I-iothalamate clearance based on a time-weighted average across up to four collection periods (after dropping the first period) and indexed to body surface area (ml/min/1.73 m^2)(median intra-test CV 9.7%).[12]

References for Table S1

1. Levey AS, Coresh J, Greene T, et al. Using standardized serum creatinine values in the Modification of Diet in Renal Disease study equation for estimating glomerular filtration rate. Ann Intern Med 2006;145(4):247-54. (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16908915).
2. Inker LA, Schmid CH, Tighiouart H, et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. N Engl J Med 2012;367(1):20-9. (Research Support, N.I.H., Extramural Validation Studies) (In eng). DOI: 10.1056/NEJMoa1114248.
3. Parsa A, Kanetsky PA, Xiao R, et al. Genome-wide association of CKD progression: the Chronic Renal Insufficiency Cohort Study. J Am Soc Nephrol 2017;28(3):923-934. DOI: 10.1681/asn.2015101152.
4. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 2011;12(1):246. DOI: 10.1186/1471-2105-12-246.
5. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526(7571):68-74. DOI: 10.1038/nature15393.
6. Mosteller RD. Simplified calculation of body-surface area. N Engl J Med 1987;317(17):1098. DOI: 10.1056/NEJM198710223171717.
7. Bansal N, Zelnick LR, Himmelfarb J, Chertow GM. Bioelectrical impedance analysis measures and clinical outcomes in CKD. Am J Kidney Dis 2018;72(5):662-672. (In eng). DOI: 10.1053/j.ajkd.2018.03.030.
8. Wilson FP, Xie DW, Anderson AH, et al. Urinary creatinine excretion, bioelectrical impedance analysis, and clinical outcomes in patients with CKD. Clin J Am Soc Nephro 2014;9(12):2095-2103. (In English). DOI: 10.2215/Cjn.03790414.

9. Scialla JJ, Appel LJ, Wolf M, et al. Plant protein intake is associated with fibroblast growth factor 23 and serum bicarbonate levels in patients with chronic kidney disease: the Chronic Renal Insufficiency Cohort study. J Ren Nutr 2012;22(4):379-388.e1. DOI: https://doi.org/10.1053/j.jrn.2012.01.026.
10. Khairallah P, Isakova T, Asplin J, et al. Acid load and phosphorus homeostasis in CKD. Am J Kidney Dis 2017;70(4):541-550. (In eng). DOI: 10.1053/j.ajkd.2017.04.022.
11. Maroni BJ, Steinman TI, Mitch WE. A method for estimating nitrogen intake of patients with chronic renal failure. Kidney Int 1985;27(1):58-65. (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3981873).
12. Anderson AH, Yang W, Hsu CY, et al. Estimating GFR among participants in the Chronic Renal Insufficiency Cohort (CRIC) Study. Am J Kidney Dis 2012;60(2):250-61. (Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't) (In eng). DOI: 10.1053/j.ajkd.2012.04.012.

Table S2. Baseline characteristics of development and validation datasets.

Variable, mean (SD)	Development Set (N=844)	Validation Set (N=404)	Standardized Difference
Age, years	55.7 (11.9)	56.1 (12.5)	0.03
Women, n (%)	367 (43.5)	172 (42.6)	0.01
Self-reported race, n (%)			0.08
American Indian/Alaska Native	8 (0.9)	0 (0.0)	
Asian	41 (4.9)	19 (4.7)	
Native Hawaiian or Other Pacific Islander	5 (0.6)	0 (0.0)	
Black/African American	296 (35.1)	151 (37.4)	
White	411 (48.7)	192 (47.5)	
Multiracial	16 (1.9)	5 (1.2)	
Unknown/Not reported	67 (7.9)	37 (9.2)	
Hispanic ethnicity, n (%)	113 (13.4)	59 (14.6)	0.02
Highest educational attainment, n (%)			0.06
6th grade or less	40 (4.7)	15 (3.7)	
7th to 12th grade no high school diploma	108 (12.8)	50 (12.4)	
High school graduate or equivalent	165 (19.5)	76 (18.8)	
Technical or vocational school degree	38 (4.5)	23 (5.7)	
Some college education but not completed degree	192 (22.7)	95 (23.5)	
College graduate	186 (22.0)	79 (19.6)	
Professional or graduate degree	115 (13.6)	66 (16.3)	
Iothalamate glomerular filtration rate (iGFR), ml/min/1.73m^2	48.3 (19.6)	48.2 (19.8)	<0.01
Serum creatinine, mg/dL	1.7 (0.5)	1.7 (0.6)	0.03
Urine creatinine, g/24hr	1.367 (0.602)	1.402 (0.564)	0.06
Creatinine clearance (CrCl), ml/min	61.2 (32.3)	62.7 (30.5)	0.05
Serum cystatin C, mg/L	1.45 (0.51)	1.45 (0.51)	0.01
Fat-free mass, kg	59.7 (15.4)	60.6 (15.0)	0.06
Missing, n (%)	12 (1.4)	10 (2.5)	
BIA phase angle, degrees	6.8 (2.9)	6.6 (1.6)	0.07
Missing, n (%)	8 (0.9)	9 (2.2)	
Body mass index, kg/m^2	31.0 (6.5)	31.7 (7.0)	0.11

Height, cm	169.2 (9.7)	168.9 (9.4)	0.03
Weight, kg	88.9 (20.1)	90.6 (21.2)	0.08
Body surface area, m^2	2.0 (0.3)	2.0 (0.3)	0.06
Self-reported dietary protein intake, g/day	72.4 (36.1)	73.2 (38.5)	0.02
Missing, n (%)	169 (20.0)	96 (23.8)	
Urine protein, g/24h	1.1 (2.2)	1.3 (2.7)	0.09
Missing, n (%)	1 (0.1)	2 (0.5)	
Urine Urea Nitrogen, g/24h	8.6 (4.2)	8.9 (4.3)	0.06
Missing, n (%)	6 (0.7)	0 (0.0)	

Table S3. Root mean squared errors and precision of estimated GFRs for all serum creatinine (SCr)-based and Cystatin C models reported in manuscript Tables 2 and 4.

Model Covariates	Root Mean Squared Error	Precision: IQR (95% CI) of the difference, mL/min/1.73m^2 (iGFR - eGFR)	
		Black	Non-Black
SCr, age, sex	11.39	12.04 (9.53, 15.47)	13.22 (11.06, 14.87)
SCr, age, sex, self-reported race	11.22	11.21 (9.00, 14.62)	12.97 (11.02, 14.78)
SCr, age, sex, % African ancestry	11.21	11.16 (8.86, 14.79)	13.09 (11.07, 14.62)
Cystatin C, age, sex	10.76	11.03 (9.09, 14.17)	11.60 (10.40, 14.74)
Cystatin C, age, sex, self-reported race	10.76	10.91 (9.08, 14.16)	11.70 (10.37, 14.75)
Cystatin C, age, sex, % African ancestry	10.76	10.79 (9.05, 14.22)	11.69 (10.38, 14.74)

Abbreviations: CI: confidence interval; IQR, interquartile range; iGFR, ^{125}I-iothalamate glomerular filtration rate; eGFR, estimated glomerular filtration rate
Models derived on a development subset of 844 (67%) participants and performance of estimated GFR reported on a validation set of 404 (33%) participants.
All 95% confidence intervals correspond to the 2.5th and 97.5th percentile values from 1000 bootstrapped samples of the validation set.
Root mean square error (RMSE) is calculated in the validation set on the same scale as the measured and estimated GFR.

Table S4. 10-fold cross validation metrics for manuscript Tables 2 and 4 using the full study sample (N=1248) instead of split-sample development and validation.

Model Covariates	Root Mean Squared Error	Median (IQR) Bias, mL/min/1.73m^2 (iGFR - eGFR)		P30		P10	
		Black	Non-Black	Black	Non-Black	Black	Non-Black
SCr, age, sex	11.20	3.29 (-2.32, 10.83)	-1.07 (-7.78, 5.25)	81	81	32	36
SCr, age, sex, self-reported race	10.91	0.43 (-5.14, 7.54)	0.78 (-5.79, 7.13)	84	83	36	36
SCr, age, sex, % African ancestry	10.90	0.64 (-5.26, 7.69)	0.83 (-5.83, 7.09)	83	84	36	37
Cystatin C, age, sex	10.61	0.17 (-5.34, 5.85)	0.09 (-5.30, 6.59)	85	85	38	40
Cystatin C, age, sex, self-reported race	10.61	0.25 (-5.13, 6.04)	-0.08 (-5.49, 6.49)	85	85	38	40
Cystatin C, age, sex, % African ancestry	10.61	0.27 (-5.14, 6.13)	-0.08 (-5.49, 6.55)	85	85	39	40

Abbreviations: SCr, serum creatinine; CI: confidence interval; iGFR, ^{125}I-iothalamate glomerular filtration rate; eGFR, estimated glomerular filtration rate; P30, percent of estimated GFR within 30% of iothalamate GFR; P10, percent of estimated GFR within 10% of iothalamate GFR
Models metrics reported using predictions from the combined validation folds from 10-fold cross validation, corresponding to the full study sample size (N=1248).

Table S5. Model performance metrics corresponding to manuscript Tables 2 and 4 using models with interaction terms between self-reported race or African ancestry and serum creatinine and cystatin C.

Model Covariates	Median (95% CI) difference, mL/min/1.73m^2 (iGFR - eGFR)		P30 (95% CI)		P10 (95% CI)		Interaction coefficient (non-transformed)
	Black	Non-Black	Black	Non-Black	Black	Non-Black	Estimate (95% CI)
SCr, age, sex, self-reported race, self-reported race*SCr	1.21 (-0.63, 2.59)	0.80 (-0.40, 2.40)	86 (80, 92)	82 (77, 87)	43 (35, 51)	37 (31, 43)	0.061 (-0.043, 0.167)
SCr, age, sex, % African ancestry, % African ancestry*SCr	1.18 (-0.07, 2.45)	0.90 (-0.37, 2.29)	86 (81, 92)	83 (77, 87)	43 (35, 51)	37 (31, 43)	0.008 (-0.005, 0.021)
Cystatin C, age, sex, self-reported race, self-reported race*cystatin C	0.98 (-1.15, 2.71)	-0.04 (-1.00, 1.16)	84 (78, 90)	82 (77, 87)	44 (36, 52)	38 (32, 44)	0.065 (-0.024, 0.155)
Cystatin C, age, sex, % African ancestry, % African ancestry*cystatin C	1.01 (-1.23, 2.61)	-0.02 (-0.96, 1.14)	84 (78, 90)	83 (78, 87)	43 (35, 51)	38 (32, 45)	0.009 (-0.002, 0.020)

Abbreviations: SCr, serum creatinine; CI: confidence interval; iGFR, ^{125}I-iothalamate glomerular filtration rate; eGFR, estimated glomerular filtration rate; P30, percent of estimated GFR within 30% of iothalamate GFR; P10, percent of estimated GFR within 10% of iothalamate GFR
Models derived on a development subset of 844 (67%) participants and performance of estimated GFR reported on a validation set of 404 (33%) participants.
All 95% confidence intervals correspond to the 2.5th and 97.5th percentile values from 1000 bootstrapped samples of the validation set.

Table S6. Associations between self-reported Black race and % African ancestry with non-GFR determinants of serum creatinine concentration.

Potential explanatory variable (non-GFR determinants of SCr) Models take the form of potential explanatory variable = [Race or African ancestry] + Age + Sex + iGFR	N	Self-reported Black race (vs. non-Black race)		African Ancestry (per 10% higher)	
		Difference for Black race	95% CI	Difference for higher % African ancestry	95% CI
Body composition and muscle mass					
Body mass index	1248	2.64 kg/m^2	1.89 to 3.39	0.335 kg/m^2	0.241 to 0.428
Body surface area	1248	0.12 m^2	0.09 to 0.15	0.014 m^2	0.011 to 0.018
Height	1248	1.68 cm	0.87 to 2.48	0.182 cm	0.080 to 0.282
Weight	1248	9.32 kg	7.07 to 11.54	1.150 kg	0.869 to 1.429
BIA phase angle, degrees (natural log-transformed)	1231	9.2%	6.5 to 11.9	1.2%	0.9 to 1.5
BIA-estimated fat-free mass	1226	4.46 kg	3.11 to 5.80	0.569 kg	0.400 to 0.737
Level of dietary protein intake					
Dietary protein intake from Diet History Questionnaire	983	-0.90 g/day	-5.54 to 3.76	-0.124 g/day	-0.699 to 0.454
Dietary protein intake from 24-hr urine	1239	-4.01 g/day	-7.12 to -1.03	-0.464 g/day	-0.854 to -0.091
Creatinine production and elimination (24-hr urine creatinine, g/day)†	1248	13.0%	7.9 to 18.3	1.6%	1.0 to 2.1
Tubular secretion of creatinine					
Ratio of creatinine clearance to iGFR (CrCl/iGFR)‡	1248	-0.03	-0.08 to 0.02	-0.004	-0.011 to 0.002
Absolute *difference* between creatinine clearance and iGFR (CrCl-iGFR)‡	1248	-1.04 mL/min/1.73m^2	-3.38 to 1.30	-0.161 mL/min/1.73m^2	-0.454 to 0.132

†In model for urine creatinine, both urine creatinine and iGFR are natural log-transformed.
‡Models for CrCl/iGFR and CrCl-iGFR are only adjusted for age and sex and not iGFR.
Abbreviations: BIA, bioelectric impedance analysis; BSA, body surface area; CrCl, creatinine clearance; iGFR, ^{125}I-iothalamate glomerular filtration rate.

Figure S1. Assembly of study sample from the Chronic Renal Insufficiency Cohort (CRIC) study.

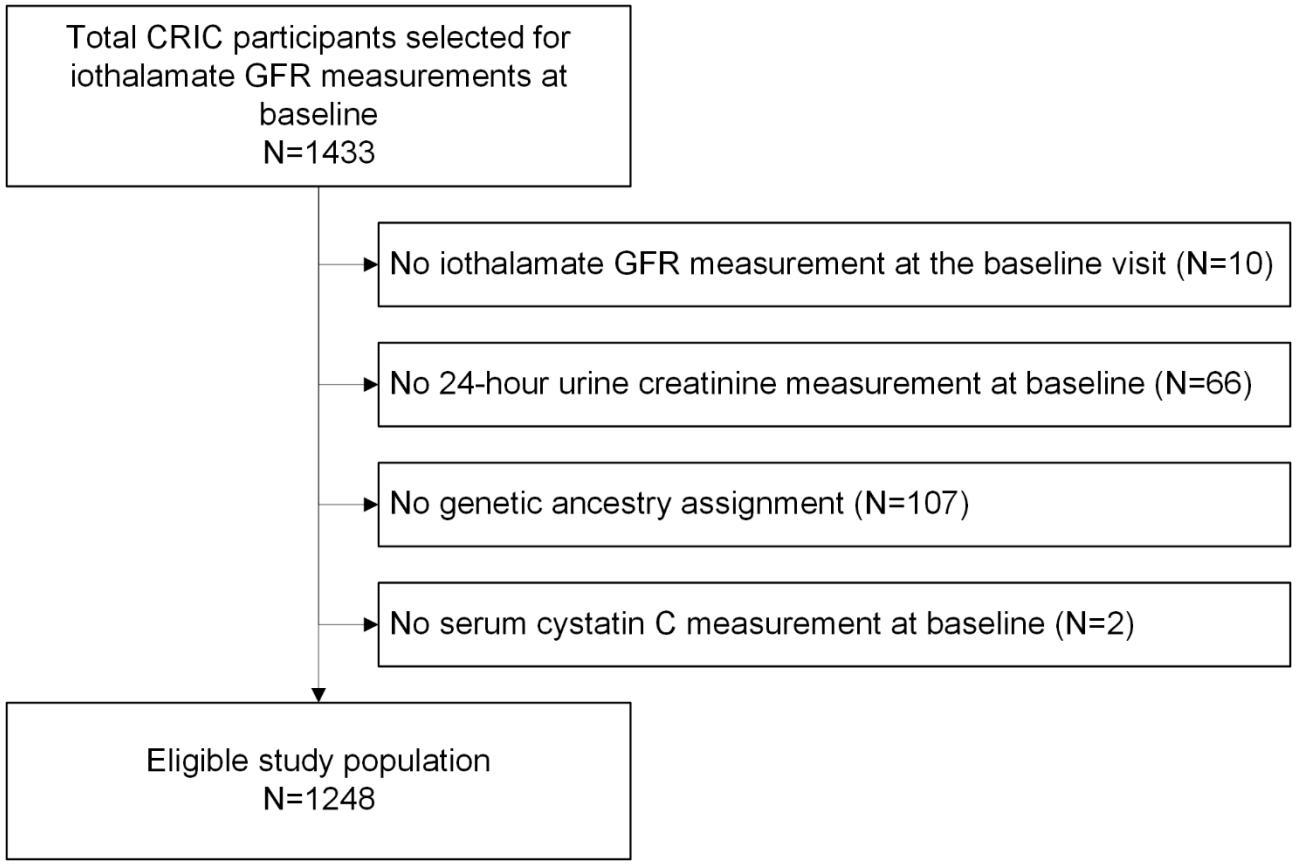
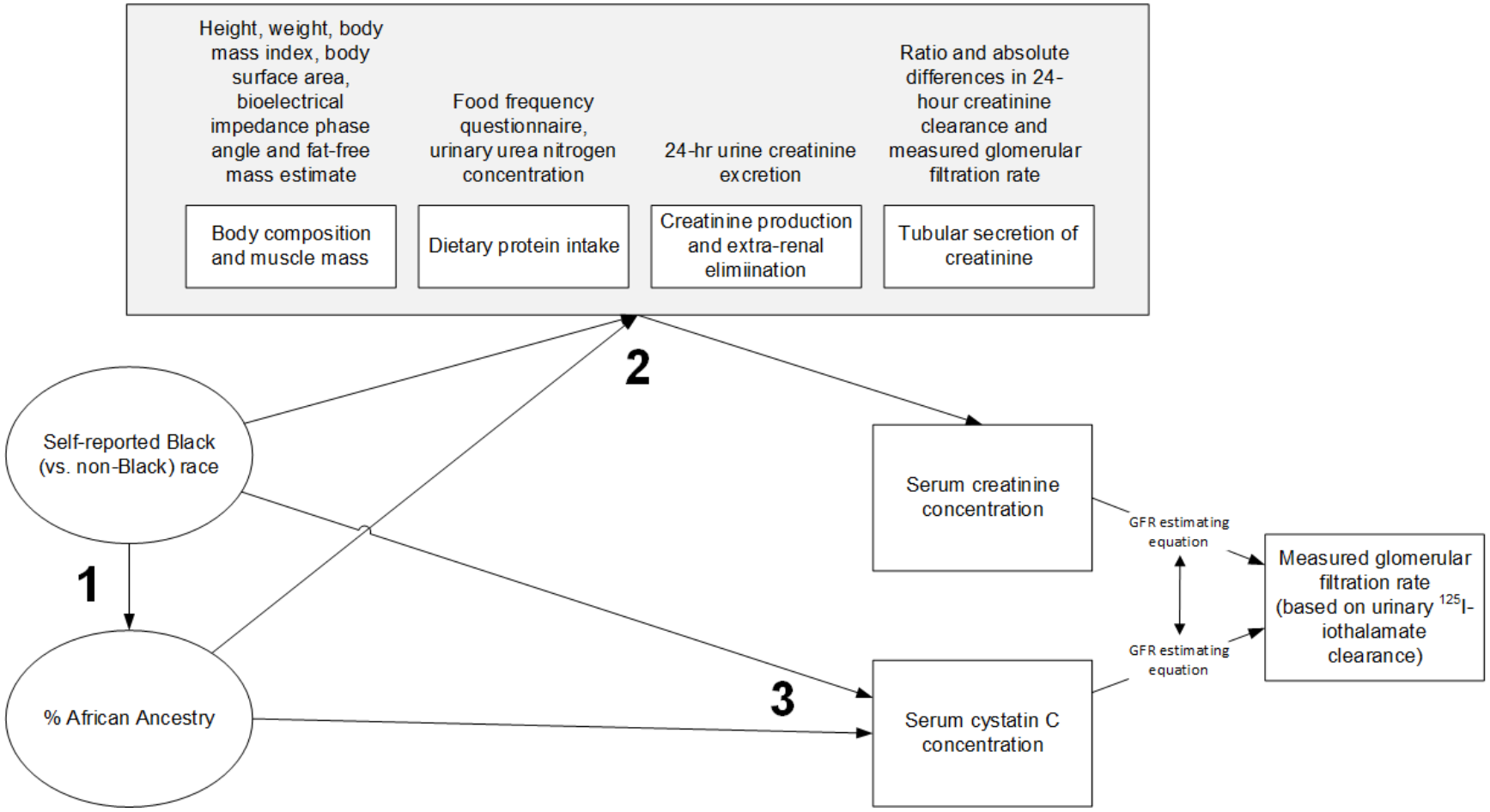


Figure S2. Conceptual approach to evaluating associations of self-reported race, genetic ancestry, serum creatinine, serum cystatin C and measured glomerular filtration rate in adults with chronic kidney disease.



1 Is it possible to estimate GFR just as well using genetically-defined ancestry instead of self-reported race in adults with mild-to-moderate chronic kidney disease?

2 Are genetic ancestry or self-reported Black race independently associated with components of creatinine production, secretion or excretion that contribute to variations in SCr levels independent of GFR? Can these variables be used to attenuate the race or ancestry coefficient in GFR estimating equations?

3 Are genetic ancestry or self-reported Black race necessary for GFR estimation when using serum cystatin C, and is a cystatin C-only equation without race or ancestry similar in accuracy to a serum creatinine-based equation that includes race or ancestry?

Figure S3. Distribution of % genetic ancestry by self-reported Black race (top figure) vs. non-Black race (bottom figure) in the study population.

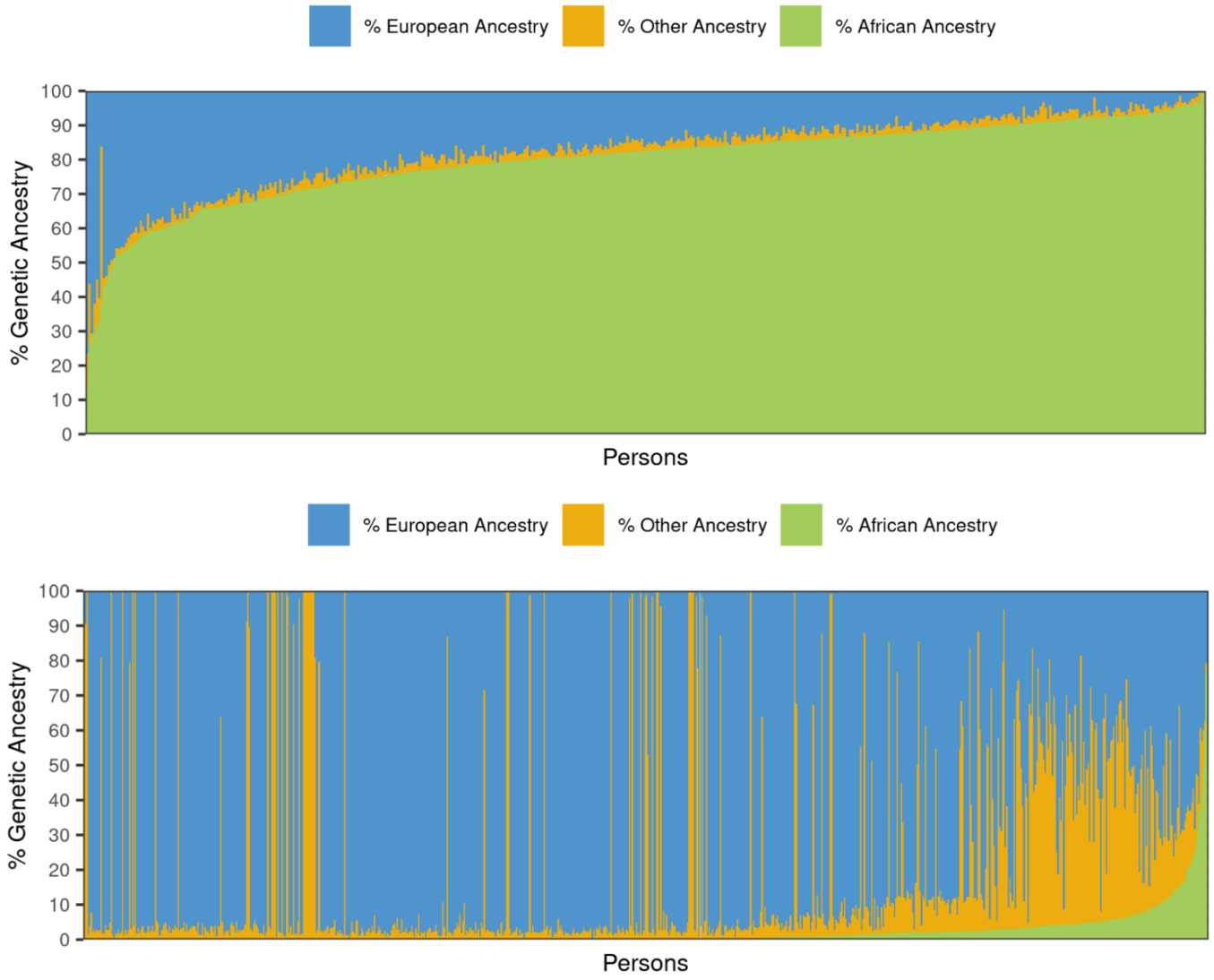
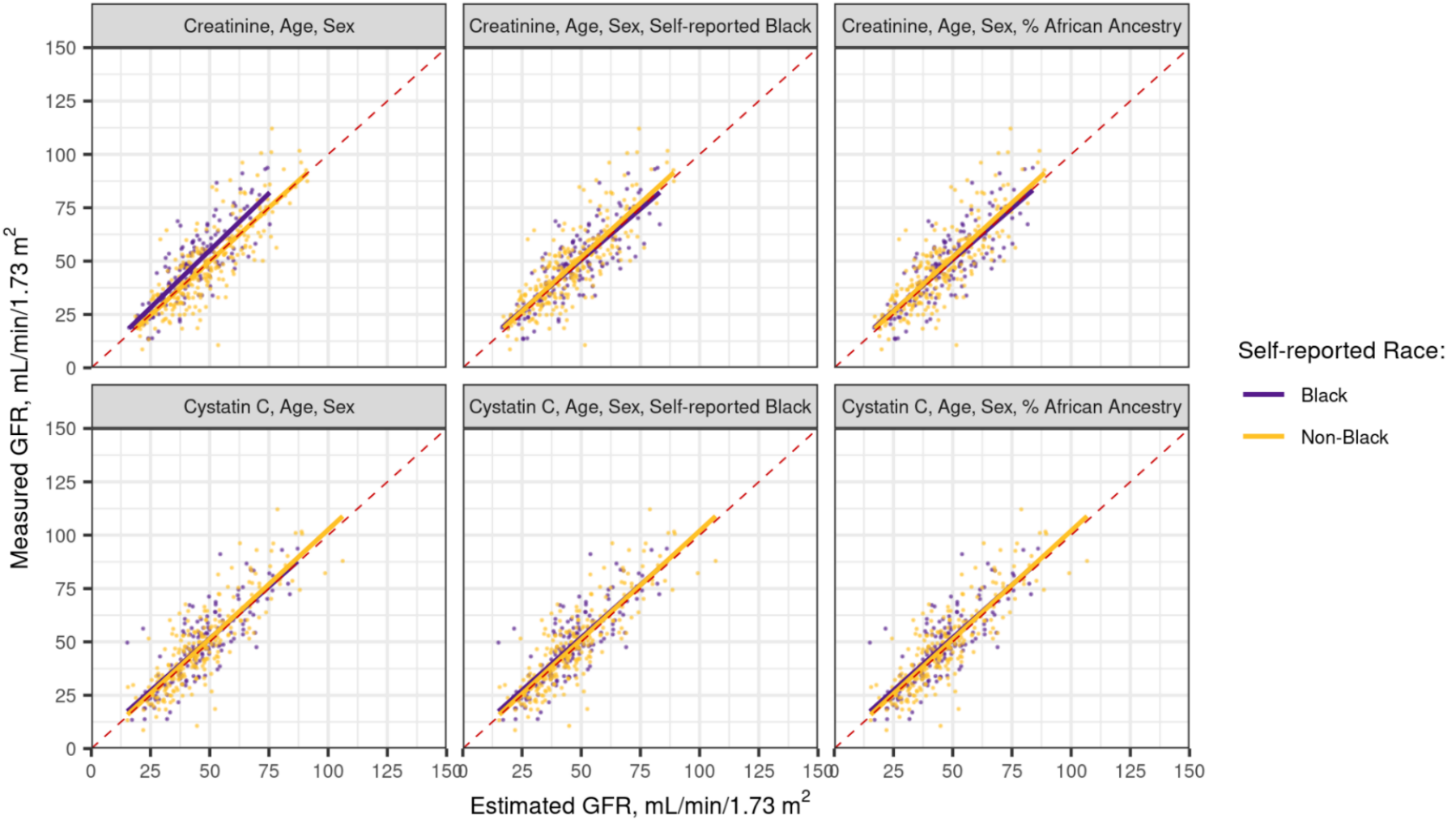


Figure S4. Measured GFR vs. estimated GFR in the validation set using different estimating equations.

Each dot represents one individual in the validation set. The solid lines are the linear fit between the measured GFR and estimated GFR stratified by self-reported race. A deviation from the dotted diagonal line indicates bias associated the GFR estimating equation.



Supplementary Methods. Methods for CRIC Ancestry estimation.

1 A general admixture model

To begin with, we introduce a general admixture model. Consider a genotype dataset $G = \{g_{ij}\}$ with genotypes at J single nucleotide polymorphisms (SNPs) from I unrelated individuals. These individuals are drawn from an admixed population with contributions from K postulated ancestral populations. Population k contributes a fraction q_{ik} of individual i's genome. The effect allele at the SNP j has frequency f_{kj} in population k for $k = 1, \ldots, K$ and g_{ij} is the observed effect allele counts of SNP j of individual i. Here we only consider the bi-allelic genetic variants so the g_{ij} will take values from $\{0, 1, 2\}$. Both the q_{ik} and the f_{kj} are unknown.

The observed dataset can be modeled by a mixture model with parameters $\{q_{ik}\}$ and $\{f_{kj}\}$. Under the assumption that the genotypes of individuals are formed by the random union of gametes, g_{ij} is modeled by the binomial distribution $Binom(2, \sum_k q_{ik} f_{kj})$ with

$$P(g_{ij} = c) = \binom{2}{c} \left[\sum_k q_{ik} f_{kj}\right]^c \left[\sum_k q_{ik}(1 - f_{kj})\right]^{2-c}, \quad c = 0, 1, 2. \tag{1}$$

Therefore for independent individuals and genetic variants in linkage equilibrium, the log-likelihood of the entire sample $\{g_{ij}\}$(up to an additive constant) is

$$L(Q, F) = \sum_i \sum_j \left\{ g_{ij} \ln[\sum_k q_{ik} f_{kj}] + (2 - g_{ij}) \ln[\sum_k q_{ik}^t (1 - f_{kj})] \right\} \tag{2}$$

$$\text{where } 0 \le f_{kj} \le 1, \quad \sum_{k=1}^{K} q_{ik} = 1. \tag{3}$$

The parameter matrices $Q = \{q_{ik}\}$ and $F = \{f_{kj}\}$ have dimensions $I \times K$ and $K \times J$. An efficient algorithm for estimating the parameters is implemented by the software `Admixture` [Alexander and Lange, 2011].

2 Projection Analysis

A number of large genome-wide datasets of human populations such as 1000 Genomes Project [Consortium et al., 2015] are now publicly available. Since these large datasets summarize worldwide human population structure, we use them as reference panels in combination with the study sample to estimate individual ancestry of study sample. This function is implemented by the projection command in `Admixture` [Shringarpure et al., 2016].

Specifically, we first do the admixture analysis on the 1000 Genomes Project data and estimate effect allele frequency $\hat{F} = \{\hat{f}_{kj}\}$ for each learned clusters $k = 1, \cdots, K$ based on (2). Since the population ancestry information of each learned cluster is known from the reference data, $\hat{F}$ can be viewed as the learned population structure. Then with $\hat{F} = \{\hat{f}_{kj}\}$ estimated from reference data and a set of CRIC study genotype data $\{g_{ij}^{(s)}\}$, we estimate the individual ancestry of CRIC data by maximizing following function,

$$L(Q^{(s)}; \hat{F}, \{g_{ij}^{(s)}\}) = \sum_i \sum_j \left\{ g_{ij}^{(s)} \ln[\sum_k q_{ik}^{(s)} \hat{f}_{kj}] + (2 - g_{ij}^{(s)}) \ln[\sum_k q_{ik}^{(s)} (1 - \hat{f}_{kj})] \right\} \quad (4)$$

where $Q^{(s)} = \{q_{ik}^{(s)}\}$ are the ancestry proportion parameters for each CRIC participants and $\sum_{k=1}^{K} q_{ik}^{(s)} = 1$. The optimized function (4) has almost the same form as (2) except that $\hat{F}$ is fixed here.

The projection approach has many advantages when a good reference dataset is available, as pointed by Shringarpure et al. [2016]. First, when a new dataset is strongly unbalanced in its distribution of populations, the accuracy of ancestry inference may be affected by the unbalance [Shringarpure and Xing, 2014], while the projection method avoids this problem. Besides, the meaning of each cluster in the study samples is suggested from the reference panel. Finally, it can be applied to estimate individual ancestry for a set of related individuals in the study samples without excluding related samples.

3 Implement Details

Pre-processing: For the CRIC study data, we first removed SNPs and individuals that did not meet standard quality-control criteria: SNPs with missing rates > 0.05, minor allele frequency < 0.01, with no founder genotypes observed were excluded as well as the individuals with missing rates > 0.1. In addition, we only kept the genotype data of bi-allelic variants on the chromosome 1-22. SNPs were further pruned for linkage disequilibrium (LD) with a window size of 50 SNPs, a step size of 10 SNPs, and a R^2 threshold of 0.1 using `PLINK 1.9`. Besides, we removed SNPs that did not intersect between the two datasets and

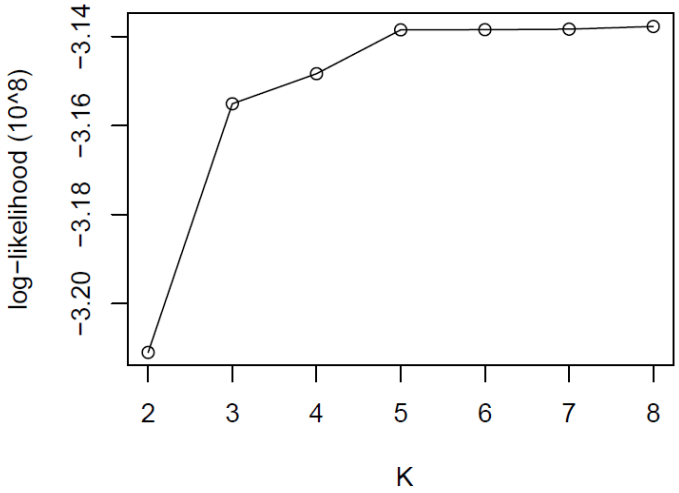
palindromic (A/T, G/C) SNPs. Finally there are 102317 variants, 3635 individuals left for the target data and 2503 individuals for the 1KGP reference data.

Then, the following steps were taken to estimate the individual ancestry of CRIC data:

- From the 1000 Genome Project data, get $\hat{F} = \{\hat{f}_{kj}\}$ given pre-determined $K = 5$.
- For the CRIC study data, estimate individual ancestry $\hat{Q}^{(s)}$ based on the estimated $\hat{F} = \{\hat{f}_{kj}\}$ and pre-determined K according to (4).
- The meaning of $\hat{Q}^{(s)}$ is acquired from the known ancestry information of reference data.

A brief discussion on the choice of K: based on previous studies, it is reasonable to choose K = 5 as our final input parameter, corresponding to the 5 super-population in reference data, African, American, European, East Asian and South Asian. Besides, this choice is also verified by comparing the log-likelihood value of fitted function (4) with different value of K. In Figure 1, it clearly shows that $K = 5$ is a sensible modeling choice for CRIC data.

Figure 1: Log-likelihood value of fitted model for CRIC data using K from 2 to 8. For each value of K, first we get estimated $\hat{F}$ from the reference data and then estimate individual ancestry $\hat{Q}^{(s)}$ of CRIC study samples based on the log-likelihood function (4). An elbow in the curve can be seen at K = 5 and for $K \geq 5$, the log-likelihood values are similar to each other. Therefore, $K = 5$ was decided to be an appropriate number of population for the CRIC data.



References for CRIC Ancestry estimation

1. Alexander DH, Lange K. Enhancements to the admixture algorithm for individual ancestry estimation. BMC bioinformatics 2011;12:246.
2. 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. Nature 2015;526:68-74.
3. Shringarpure S, Xing EP. Effects of sample selection bias on the accuracy of population structure and ancestry inference. G3: Genes, Genomes, Genetics 2014;4: 901-911.
4. Shringarpure SS, Bustamante CD, Lange K, Alexander DH. Efficient analysis of large datasets and sex bias with admixture. BMC bioinformatics 2016;17:1-6.