The Loss of *GSTM1* is Associated with Kidney Failure and Heart Failure:

The Atherosclerosis Risk in Communities (ARIC) Study

Supplementary Methods, Figures, and Tables

Corresponding author:

Adrienne Tin (atin1@jhu.edu)

Supplementary Methods

Study Population and CHARGE-S Case-cohort Design

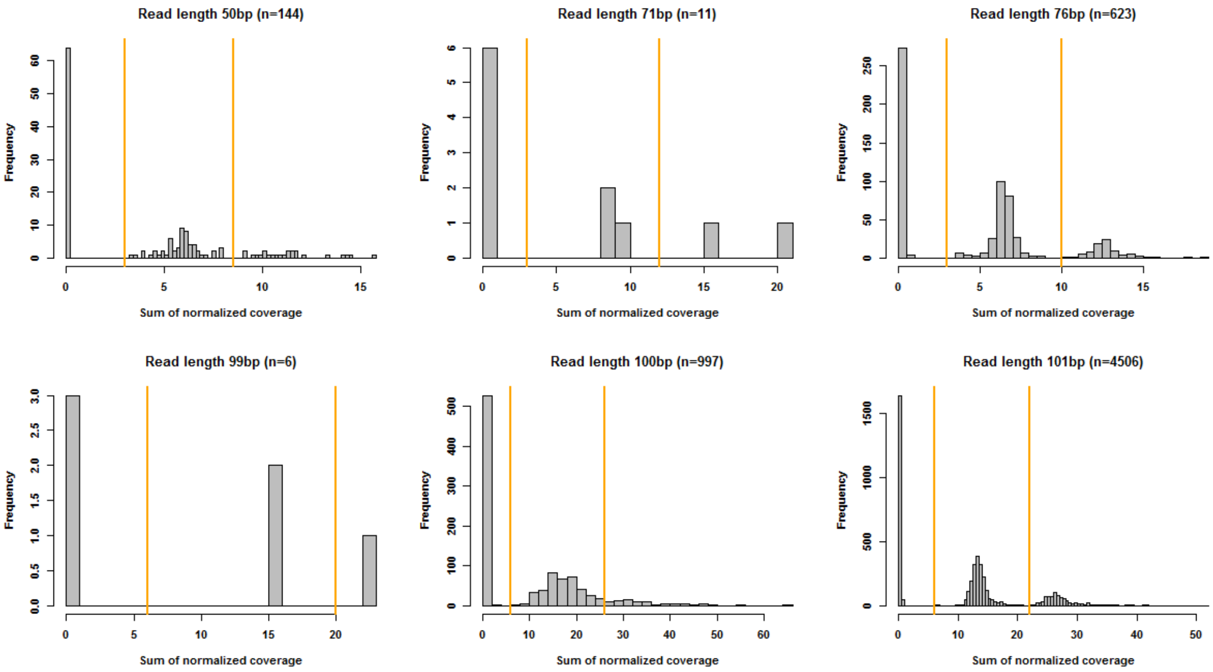
The present study included ARIC participants who were selected in the National Heart, Lung and Blood Institute (NHLBI) Exome Sequencing Project and CHARGE-S project (the database of Genotypes and Phenotypes [dbGaP] study accession: phs000668.v1.p1).[1] Both projects used a case-cohort design that included a common reference sample and cases from multiple phenotypes. Of the 6491 ARIC participants included in this accession number, whole exome sequencing reads from the Illumina platform were available for 6371 participants. We downloaded the exome sequencing reads of chromosome 1, which contains *GSTM1*, from dbGaP. The CODEX package was used for quality control.[2] Regions were excluded based on the following quality filters: coverage >4000 reads, mappability < 0.4, GC content < 0.20 or > 0.80, or mapping quality < 20. These quality filters excluded 7 samples. We further excluded samples with overall median coverage < 40 reads at chromosome 1 (n=73) and read lengths with less than 5 samples: read length of 67 base pair (bp, n=2), 68 bp (n=1), and 70 bp (n=1). In total, *GSTM1* copy number was available for 6,287 participants. After further excluding those with eGFR < 15 mL/min/1.73m^2 or missing eGFR (n=89), having missing values in clinical covariates (n=159) or genetic principal components generated using genotypes from the Illuminia HumanExome BeadChip (n=324), the overall analyzed sample included 5,715 participants.

Within the overall analyzed sample, 4,137 were in the common reference samples, and 1,578 were cases of the following phenotypes: high blood pressure (n=298), low blood pressure (n=277), early onset myocardial infarction cases (n=60) and controls (n=160), high and low levels of low density lipoprotein cholesterol (high: n=133, low, n=137), stoke (n=9), fibrinogen (n=76), kidney function (n=73), early menopause (n=77), electrocardiogram QT interval (n=77), waist circumference (n=68). Details of the case selection criteria are described at the dbGaP website.[1] To account for this case-cohort design, we included sampling weights in the Cox regression analysis of *GSTM1* deletion with incident kidney failure, heart failure, and chronic kidney disease (CKD). All cases were weighted as 1 and the participants in the common reference sample were weighted by the reciprocal of the probability of sampling.

Whole Exome Sequencing and Alignment

The exome was captured using NimbleGen SeqCap EZ VCRome (Roche, Basel, Switzerland). The enriched library was then sequenced by Illumina HiSeq platform at Human Genome Sequencing Center at Baylor College of Medicine. The Mercury pipeline[3] was used to process sequencing data. The raw short reads were aligned to the reference human genome (NCBI Genome Build 37, 2009) by Burrows-Wheeler Aligner.[4]. The mean read depth was 92x, and more than 92% of target regions were covered by at least 20 unique reads.

Supplementary Figure 1. Distribution of the sum of the normalized coverage at *GSTM1* by read length



The orange lines mark the thresholds for determining *GSTM1* copy numbers reported in **Supplementary Table 3**.

Supplementary Table 1. Characteristics of participants included and excluded from the present study

	European Americans			African Americans		
	Excluded	Included	P-value	Excluded	Included	P-value
N	8017	3461		2012	2254	
Age, mean (SD)	55.0 (5.8)	54.6 (5.6)	<0.001	54.5 (5.8)	53.65 (5.8)	<0.001
Male, n (%)	3817 (47.6)	1611 (46.5)	0.30	807 (40.1)	824 (36.6)	0.02
BMI, mean (SD)*	27.0 (4.93)	27.0 (4.7)	0.81	29.6 (6.1)	29.6 (6.2)	0.61
Smoking, n (%)*			0.02			0.27
Current smoker	1998 (24.9)	846 (24.4)		614 (30.7)	658 (29.2)	
Former smoker	2767 (34.5)	1289 (37.2)		453 (22.6)	555 (24.6)	
Never smoker	3244 (40.5)	1326 (38.3)		936 (46.7)	1041 (46.2)	
Diabetes, n (%)*	760 (9.5)	286 (8.3)	0.04	419 (22.2)	402 (17.8)	<0.001
Hypertension, n (%)*	2148 (27)	973 (28.1)	0.23	1180 (59.3)	1194 (53.0)	<0.001
Prevalent CHD, n (%)*	435 (5.6)	159 (4.6)	0.03	82 (4.3)	89 (3.9)	0.67
eGFR, mL/min/1.73m^2, mean (SD)*	99.5 (12.4)	99.1 (13.0)	0.11	109.8 (21.7)	111.9 (18.8)	<0.001
Incident kidney failure, n (%)*	212 (2.6)	99 (2.9)	0.56	150 (7.5)	157 (7.0)	0.54
Incident CKD, n (%)*	1158 (14.6)	635 (18.6)	<0.001	238 (13.1)	325 (14.7)	0.15
Incident heart failure, n (%)*	1351 (18.1)	547 (16.4)	0.04	453 (25.3)	516 (24.4)	0.52

The ARIC study enrolled 15,792 participants (11,478 European Americans, 4,266 African Americans, 48 other races).

*In European Americans, among those excluded, the number of participants with baseline data for BMI (8,007), smoking (8,009), diabetes (7,990), hypertension (7,958), prevalent CHD (7,763), eGFR (8,003), incident kidney failure (8,013), incident CKD (7,930), incident heart failure (7,478). Among those included, the number of participants with data for incident CKD (3,420), incident heart failure (3,329).

In African Americans, among those excluded, the number of participant with data for BMI (1,996), smoking (2,003), diabetes (1,891), hypertension (1,991), prevalent CHD (1,924), eGFR (1,876), incident kidney failure (1,999), incident CKD (1,823), incident heart failure (1,787). Among those included, the number of participants with data for incident CKD (2,215), incident heart failure (2,114).

Abbreviation. CHD, coronary heart disease; eGFR, estimated glomerular filtration rate; CKD, chronic kidney disease; SD, standard deviation.

Supplementary Table 2. Risk of incident CKD associated with *GSTM1* copy number (N=5,634)

	Hazard Ratio (95% Confidence Interval)			
	Genotypic Model (By *GSTM1* copy number)			
	0 copy	1 copy	2 copies	P for trend
European Americans				
Model 1	1.08 (0.95, 1.23)	1.14 (0.99, 1.30)	Reference	0.85
Model 2	1.08 (0.95, 1.24)	1.16 (1.01, 1.33)	Reference	0.97
African Americans				
Model 1	0.93 (0.78, 1.10)	0.92 (0.79, 1.06)	Reference	0.35
Model 2	0.86 (0.73, 1.02)	0.91 (0.79, 1.05)	Reference	0.09
Overall				
Model 1	1.00 (0.91, 1.11)	1.03 (0.94, 1.14)	Reference	0.78
Model 2	0.99 (0.90, 1.09)	1.05 (0.96, 1.16)	Reference	0.41

N (Events): European Americans, 3,420 (987); African Americans: 2,214 (660).
Model 1: age, sex, center, and 10 genetic principal components/
Model 2: Model 1 + prevalent diabetes, hypertension, coronary heart disease, smoking status, BMI, eGFR.
Race was included as a covariate in the overall analysis.

Supplementary Table 3. Exome Sequencing reads by read length

		Threshold for determining *GSTM1* copy number (unit: sum of normalized coverage at *GSTM1*)	
Read length (bp)	**N**	**0 copy**	**1 copy**
50	144	3	8
71	11	3	12
76	623	3	10
99	6	6	20
100	997	6	26
101	4506	6	22
Total	6287		

Total number of whole exome sequencing sample from dbGap: 6,371.
Number of samples after QC and removing those with the median of the overall normalized coverage < 40 at chromosome 1: 6,287.
The histograms of the sum of normalized coverage by read length are presented in **Supplementary Figure 1**.

Reference

1. Building on GWAS for NHLBI-Diseases : the U.S. CHARGE Consortium (CHARGE-S): ARIC (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000668.v1.p1).
2. Jiang, Y, Oldridge, DA, Diskin, SJ, Zhang, NR: CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res,* 43**:** e39, 2015.
3. Reid, JG, Carroll, A, Veeraraghavan, N, Dahdouli, M, Sundquist, A, English, A, Bainbridge, M, White, S, Salerno, W, Buhay, C, Yu, F, Muzny, D, Daly, R, Duyk, G, Gibbs, RA, Boerwinkle, E: Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics,* 15**:** 30, 2014.
4. Li, H, Durbin, R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**:** 1754-1760, 2009.