1.a. Full Title: Association between candidate genetic variants and incident chronic kidney disease: the Atherosclerosis Risk in Communities (ARIC) Study

b. Abbreviated Title (Length 26 characters):
Genetic risk for incident CKD

2. Writing Group:
Writing group members: Anna Kottgen, Wen Hong Linda Kao, Eric Boerwinkle, Alanna Morrison, Lance Bare, Josef Coresh

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _AK_ [please confirm with your initials electronically or in writing]

First author: Anna Kottgen
Address: 2024 E. Monument St.
Welch Center for Prevention, Epidemiology, and Clinical Research
Baltimore, MD 21287

Phone: (410) 245-6897  Fax: (410) 955-4076
E-mail: akottgen@jhsph.edu

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):
Josef Coresh
Address: 2024 E. Monument St.
Welch Center for Prevention, Epidemiology, and Clinical Research
Baltimore, MD 21287

Phone: (410) 955-0495  Fax: (410) 955-4076
E-mail: coresh@jhu.edu

3. Timeline: Analyses to start immediately; analyses and manuscript preparation are projected to take place over the next year.
4. Rationale:

Chronic kidney disease (CKD) has been recognized as a public health problem which affects an estimated 19 million adults in the US\textsuperscript{1,2}. Progression of CKD may lead to end-stage renal disease (ESRD). The yearly mortality rates for individuals treated with dialysis for ESRD exceed 20\% \textsuperscript{3}. Therefore, early identification of individuals at increased risk for CKD and effective intervention is essential. Previous studies have shown that sub-groups of individuals susceptible to kidney disease exist \textsuperscript{4}. Additionally, multiple studies have confirmed that kidney disease is heritable \textsuperscript{5}. Apart from hypertension, diabetes mellitus, and other cardiovascular risk factors, genetic causes contribute to the complex disease CKD.

CKD is associated with numerous co-morbid conditions and an elevated risk of cardiovascular disease (CVD) and mortality \textsuperscript{6}. This association could be causal in either direction, or it could originate from a common underlying mechanism causing both kidney and cardiovascular disease. In fact, the pathogenesis of CKD appears to be similar to the atherosclerotic process, and both share many risk factors such as hypertension. Therefore, CKD and CVD may also share common susceptibility genes, and evidence from the ARIC Study supports this hypothesis \textsuperscript{7,8}.

Traditional methods for disease gene identification, such as linkage analysis, have been very effective in mapping rare disorders for which single mutations are sufficient to cause disease. However, they have not been successful in identification of genes for complex conditions, such as CKD. Association studies in large study populations provide greater power for identifying variants responsible for such traits \textsuperscript{9}. The large representative sample, bi-racial population, information about and measurement of cardiovascular and renal risk factors and function, and the extended follow-up of the ARIC Study make it possible to prospectively study genetic risk factors that may confer susceptibility to both CVD and CKD. In addition, studying large prospective cohorts like the ARIC cohort allows for quantification both of the relative risk as well as the attributable risk associated with CKD susceptibility genes.

5. Main Hypothesis/Study Questions:

**Main hypothesis:** Some of the genetic variants that have been associated with incident coronary heart disease in previous studies will be associated with incident CKD in the ARIC Study, possibly reflecting a common underlying causal pathway.

**Study questions:**
1. Will the association of these genetic variants and incident CKD be consistently present for different definitions of incident CKD, as well as cross-sectionally with measures of kidney function (estimated glomerular filtration rate) and kidney damage (albuminuria)?
2. How will such associations change upon stratification by prevalent CHD at baseline, or stratification or adjustment for variables that could represent a common underlying mechanism to both CVD and CKD, such as hypertension?
3. Will the association of such genetic variants and CKD be present in both African American and Caucasian individuals? Furthermore, can differences in frequencies of
such genetic variants between these two populations account for part of the differences in disease risks?

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Study design: Prospective follow-up of all ARIC participants meeting the inclusion criteria from baseline (visit 1, 1987-1989) through January 1, 2003.

The single nucleotide polymorphisms (SNPs) to be examined have been genotyped as part of a panel proposed by Celera diagnostics in collaboration with Dr. Boerwinkle as described in ARIC Ancillary Study 2004.11. They have been used to obtain a genetic risk score for CHD prediction in ARIC (see MP #1095). In brief, these SNPs were selected from two sources: a literature review of CHD association studies (24 SNPs in 21 a priori candidate genes), and ongoing large-scale genomic association studies of CHD (92 SNPs in 87 genes). (need to find out which ones were the 101 SNPs in the Celera AS folder – how many candidates vs. those from large-scale genomic association studies). In addition, would be good to know p-values of the associations for each of those SNPs in the original case-control studies in order to be able to formulate the prior better (see statistical significance section). The 92 SNPs from the genomic association studies demonstrated association with CHD in at least one of three case-control studies in white participants. One study from the Cleveland Clinic Foundation compared cases with severe coronary artery disease with controls free of disease. The other two association studies compared cases with severe myocardial infarction to controls with no history of myocardial infarction.

Inclusions/exclusions: Participants who did not consent to genotyping will be excluded from analysis (use of DNA data distributed by the Coordinating Center, confirmation by using the variable “res_dna” in datafile “ictder02” (n = 45 individuals did not consent to DNA use for the type of study outlined here)). Moreover, individuals reporting race other than African American or white will be excluded (n = 48), as will be individuals missing variables needed to calculate estimated glomerular filtration rate (eGFR) as a measure of kidney function at visit 1 (n = 150, all due to missing serum creatinine (variable chma09)). Depending on the definition of incident CKD used (see below), individuals with severe hypercreatinemia (n = 40, serum creatinine \( \geq 2.0 \) mg/dl for men, \( \geq 1.8 \) mg/dl for women) or those with eGFR < 60 at visit 1 (n = 462) will also be excluded from analyses.

The number of individuals excluded due to missing or unknown genotype will depend on the SNP under investigation.

Outcome: The primary outcome will be incident CKD (n = 1,616) as defined by a decrease in eGFR from \( \geq 60 \) ml/min/1.73m\(^2\) at baseline to <60 ml/min/1.73m\(^2\) at the second or fourth follow-up visit, or a hospitalization discharge or death coded for chronic
renal disease (*International Classification of Diseases, Ninth Revision* [ICD-9] codes 581-583 or 585-588), hypertensive renal disease (*ICD-9* code 403), hypertensive heart and renal disease (*ICD-9* code 404), unspecified disorder of kidney and ureter (*ICD-9* code 593.9), diabetes with renal manifestations (*ICD-9* code 250.4), kidney transplantation, renal dialysis, or adjustment/fitting of catheter (*ICD-9* codes V42.0, V45.1, or V56), hemodialysis (*ICD-9* code 39.95) or peritoneal dialysis (*ICD-9* code 54.98), without acute renal failure (*ICD-9* codes 584, 586, 788.9, and 958.5) as the primary or secondary hospitalization code, all from ARIC surveillance datasets (c02occ1, c02celb1, c02dtha1).

In secondary analyses, CKD defined as defined by a rise in serum creatinine of at least 0.4 mg/dl above baseline or CKD hospitalization or death (defined as above) as an outcome will be explored (n = 1,201). Additionally, albuminuria at visit 4 as well as eGFR at visit 1 and 4 will be investigated as secondary outcomes cross-sectionally.

eGFR as a measure of kidney function will be calculated using the abbreviated Modification of Diet in Renal Disease (MDRD) Study formula: eGFR (ml/min/1.73m² = 175 * (serum creatinine)^-1.154 * age^-0.203 * (0.742 if female) * (1.21 if black)¹⁰.

Other variables of interest: Variables needed to calculate eGFR: serum creatinine from visit 1, 2, and 4 (chma09, chmb08, lipd6a), age (v1age01), race (racegrp), and gender (gender). The variable “ACRv2” will be used to assess the albumin-to-creatinine ratio (ACR) at visit 4. The SNPs to be explored are contained in the datafile “Celera AS”, and are labeled “cvXXX” where X represents a number. Further covariates include risk factors for CHD: blood pressure (variables sbp21a and dbp21a), use of anti-hypertensive medication (hyptmd04), HDL- and LDL-cholesterol (hld02 and ldl01), diabetes mellitus (diabts03), smoking (cursmk01), and body mass index (bmi01) at visit 1. Variables for both prevalent CHD at visit 1 (prvchd05) as well as incident CHD over the course of follow-up (variable “in_02sp” in the datafile “inc_by02”) will also be needed for analyses. Other covariates may be selected, depending on the specific SNP of interest and the hypothesized function of the corresponding gene.

Data analysis:

**Data checks:** Hardy-Weinberg equilibrium (HWE) will be checked by race for each SNP by using the chi-square goodness-of-fit test as well as using the Fisher exact test¹¹. Differences in the distribution of genotypes between individuals included in the study and those excluded from our analyses will be conducted. Moreover, differences in the distribution of incident CKD cases between those missing genotype data and those with available genotype data will be assessed using chi-square tests.

**Exploratory and cross-sectional data analyses:** The distribution of baseline characteristics in the study population by genotype as well as by outcome will be computed using t-tests, chi-square tests and ANOVA as applicable. All association analyses will first be examined within each self-reported race groups. For cross-sectional analyses of the association of genotypes with eGFR and albuminuria, mean eGFR and albumin-to-creatinine ratio (ACR) will be estimated and compared for the three genotypes at each SNP using ANOVA. Multiple linear regression will be used to examine the association of genotypes and continous eGFR and ACR. Models will be
constructed to account for effects of potential confounders similar to the construction of Cox proportional hazards models (see below).

**Survival analyses:** For CKD cases, follow-up time will be counted from visit 1 until the visit date at which the creatinine rise / eGFR decline occurred, or the date of CKD hospitalization discharge or death, or the earlier of the two dates for participants meeting both definitions. Non-cases will be censored at the earlier of the date of last contact (or date of non-CKD death) or December 31, 2002. Incidence rates of CKD will be calculated using person-time methods. Kaplan-Meier estimates of mortality will be computed, and log-rank tests will be used to compare survival curves among the genotypes.

In regression analyses, an additive genetic model will be assumed unless indicated otherwise by results of the analysis, or unless the allele frequency of a given candidate variant is low, in which case a dominant model combining the risk of heterozygotes and homozygotes will be used. Genotype will be coded as 0 (zero copies of the risk increasing allele), 1 (one copy of the risk increasing allele), or 2 (two copies of the risk increasing allele). Stratified analyses will be conducted to examine the potential for interaction and effect modification between the covariate (e.g., hypertension) and the association of the SNP and incident CKD.

Each SNP will be tested for association with incident CKD in crude Cox proportional hazard analyses, by race if applicable. Those with a p-value of < 0.1 in these analyses will be considered for further analyses. Cox proportional hazard regression will then be used to estimate the effect size (RH of incident CKD) and 95% confidence interval for each SNP. Subsequent multivariate models will include basic variables (age, sex, race if no interaction by race is observed), co-morbidities thought to act as potential confounders (prevalent CHD at baseline, hypertension, diabetes mellitus), and relevant potential intermediate variables depending on the putative function of the gene in which the SNP is located.

In additional analyses, individuals with prevalent CHD will be excluded from analyses, and those with incident CHD will be censored at the time of their CHD event. To examine the impact of informative censoring in these analyses, sensitivity analyses for this scenario will be conducted assuming the most extreme scenarios of informative censoring (i.e. all incident CHD-censored individuals would have developed HF or all of them would have survived the entire follow-up without developing HF).

**Determination of statistical significance:**
We recognize the limitations of screening a large panel of genetic variants as risk factors for CKD. Determining statistical significance in the face of multiple testing is controversial. Instead, rather than using a Bonferroni correction and controlling the overall type-1 error rate at a level of 0.05, we propose to critically evaluate significant findings at the α-level of 0.05 after multi-variate adjustment for each SNP by applying the false positive report probability (FPRP) as proposed by Wacholder et al. The FPRP is the probability of no true association between a genetic variant and disease given a statistically significant finding. It is based on 1) the prior probability \( \pi \) of a true association of the tested genetic variant with the disease, 2) the observed p-value, and 3) the statistical power to detect the effect size of the alternative hypothesis at the given \( \alpha \) level, based on sample size, allele frequency, and the specified effect size for the presumed association under the alternative hypothesis.
The power in ARIC is high for moderate associations. A simplified analysis of CKD cases vs. non-cases (ignoring follow-up time) shows that assuming \( \alpha = 0.05 \), an OR of 1.3 (1.6) and allele frequencies of 0.2, 0.1, 0.05 and 0.01 results in calculated power of 98.5% (100%), 88.3% (100%), 64.4% (98.9%), and 19.0% (53.7%).

A FPRP level of “noteworthiness” will be chosen at 0.5, i.e. the probability of a true association between the genetic variant and incident CKD given a statistically significant finding will be more likely than not for values below 0.5. This FPRP will be chosen lower at 0.2 for SNPs that were selected as candidate SNPs for CKD based on results from prior studies. The prior probability \( \pi \) of a true association of the tested SNP with CKD will be classified into one of 3 categories applying the following criteria:

1) \( \pi = 0.1 \) (high probability): non-synonymous candidate SNP for CKD with functional evidence of this SNP or the gene in which it is located from animal models, monogenetic diseases, or replicated in \( \geq 3 \) independent human studies.

2) \( \pi = 0.01 \) (moderate probability): non-synonymous SNP or SNP in predicted splice sites or transcription factor binding sites in genes within a pathway thought to underlay the disease of interest or associated with diseases with possibly related etiologies in prior independent studies 12.

3) \( \pi = 0.001 \) (low probability): randomly selected synonymous or non-synonymous SNP.

The SNPs under investigation in this proposal will be assigned a prior of 0.01 (moderate probability), with individual SNPs most likely having a somewhat higher or lower prior. In addition, a sensitivity analysis of the effect across a wide range of priors will be performed in order to inform others who might assume a prior probability different from ours.

In addition, it will be attempted to replicate positive findings in other cohorts comparable to the ARIC study with respect to study population and phenotype and genotype assessment, such as the Framingham Heart Study. This replication approach has been successfully applied, demonstrating the feasibility and the potential of genome-wide SNP association studies for identification of novel genes for complex traits 13. The publications committee will be notified if the paper will include a replication dataset from Framingham or another study.

**Limitations:**

A limitation to the proposed research project is the problem of multiple comparisons that we will address as outlined above. Another problem is the possible impact of undetected population substructures on the study findings. However, the impact of population stratification on false-positive findings has been reported as low 11. In addition, we will be getting data from ancestry informative makers from CIDR early next year. This information can then be incorporated into our analyses.

If there are positive findings, they will need to be scrutinized in terms of biological mechanism and tested in further studies.
7.a. Will the data be used for non-CVD analysis in this manuscript? ____ Yes  
    _X__ No

    b. If Yes, is the author aware that the file ICTDER02 must be used to exclude 
    persons with a value RES_OTH = “CVD Research” for non-DNA analysis, and 
    for DNA analysis RES_DNA = “CVD Research” would be used?  ____ 
    Yes  ____ No
    (This file ICTDER02 has been distributed to ARIC PIs, and contains 
    the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? _X_ Yes     ____ No

8.b. If yes, is the author aware that either DNA data distributed by the 
    Coordinating Center must be used, or the file ICTDER02 must be used to 
    exclude those with value RES_DNA = “No use/storage DNA”?  
    _X_ Yes  ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC 
    Study manuscript proposals and has found no overlap between this proposal and 
    previously approved manuscript proposals either published or still in active status. 
    ARIC Investigators have access to the publications lists under the Study Members Area 
    of the web site at:  http://www.cscc.unc.edu/ARIC/search.php

    ____X__ Yes  _______ No

10. What are the most related manuscript proposals in ARIC (authors are 
    encouraged to contact lead authors of these proposals for comments on the new 
    proposal or collaboration)?

    MS 1095: Coronary heart disease risk prediction in the Atherosclerosis Risk in 
    Communities (ARIC) Study using a genetic risk score
    MP 1142: Genetic risk of Coronary Heart Disease in the Atherosclerosis Risk in 
    Communities (ARIC) study: Application of a Genetic Risk Score

11. a. Is this manuscript proposal associated with any ARIC ancillary studies or use 
    any ancillary study data?  ____X__ Yes  ____ No

11.b. If yes, is the proposal
    ____  A. primarily the result of an ancillary study (list number*__ genetic data 
    from ARIC ancillary study 2004.11, albuminuria data from ARIC ancillary study 
    2002.02)
    ____  B. primarily based on ARIC data with ancillary data playing a minor role 
    (usually control variables; list number(s)* _____)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12. Manuscript preparation is expected to be completed in one to three years.  If a
manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

References


