AJKD
Perspective

Genome-Wide Association Studies in Nephrology: Using Known Associations for Data Checks

Matthias Wuttke, MD,[1] Franz Schaefer, MD,[2] Craig S. Wong, MD, MPH,[3] and Anna Köttgen, MD, MPH[1]

Prior to conducting genome-wide association studies (GWAS) of renal traits and diseases, systematic checks to ensure data integrity and analytical work flow should be conducted. Using positive controls (ie, known associations between a single-nucleotide polymorphism [SNP] and a corresponding trait) allows for identifying errors that are not apparent solely from global evaluation of summary statistics. Strong genetic control associations of chronic kidney disease (CKD), as derived from GWAS, are lacking in the non-African ancestry CKD population; thus, in this perspective, we provide examples of and considerations for using positive controls among patients with CKD. Using data from individuals with CKD who participated in the CRIC (Chronic Renal Insufficiency Cohort) Study or PediGFR (Pediatric Investigation for Genetic Factors Linked to Renal Progression) Consortium, we evaluated 2 kinds of positive control traits: traits unrelated to kidney function (bilirubin level and body height) and those related to kidney function (cystatin C and urate levels). For the former, the proportion of variance in the control trait that is explained by the control SNP is the main determinant of the strength of the observable association, irrespective of adjustment for kidney function. For the latter, adjustment for kidney function can be effective in uncovering known associations among patients with CKD. For instance, in 1,092 participants in the PediGFR Consortium, the *P* value for the association of cystatin C concentrations and rs911119 in the *CST3* gene decreased from 2.7×10^{-3} to 2.4×10^{-8} upon adjustment for serum creatinine–based estimated glomerular filtration rate. In this perspective, we give recommendations for the appropriate selection of control traits and SNPs that can be used for data checks prior to conducting GWAS among patients with CKD.
Am J Kidney Dis. 65(2):217-222.

INDEX WORDS: Genome-wide association study (GWAS); single-nucleotide polymorphism (SNP); genetic marker; positive control; data checking; systematic error; chronic kidney disease (CKD); renal trait; cystatin C.

Genome-wide association studies (GWAS) investigate millions of genetic markers per person to identify genomic regions in which genetic variation associates with a trait or disease. For each single-nucleotide polymorphism (SNP), summary statistics (eg, a *P* value) are calculated for its association with the phenotype. The resulting GWAS file contains millions of lines, which makes visual data and plausibility checks challenging. Although there are excellent tools for systematically checking genome-wide summary statistics, including allele frequencies and computed association statistics, for their global distribution,[1,2] other systematic errors can go unnoticed. For example, incorrect association results may arise from a mismatch between the genotypes and phenotypes of the individuals (ie, inadvertent scrambling of the data). Because most SNPs are not expected to show an association with the phenotype of interest, such errors would escape global checks of summary statistics. Hence, additional checks to reliably assess the integrity of data and analytical workflow are required.

These considerations highlight the importance of using a positive control, one or several genomic markers that are known to reproducibly associate with an available trait or phenotype. A SNP that is a suitable positive control should have an effect strong enough to be detected in as little as a few hundred samples. As outlined in Box 1, a practical approach is to survey the GWAS Catalog of the National Human Genome Research Institute (NHGRI) for phenotypes and diseases available in a given study and identify SNPs that previously have shown genome-wide significant associations ($\sim P < 1 \times 10^{-7}$ or $< 1 \times 10^{-8}$, depending on the study). The SNPs should have been replicated successfully and shown associations in samples of the same ancestry as the data to be analyzed. We suggest reviewing the cited publications from the GWAS Catalog to select the SNP(s) that explain the largest amount of the trait variance. If this

From the [1]Division of Nephrology, Medical Center- University of Freiburg, Freiburg; [2]Division of Pediatric Nephrology, University Medical Center Heidelberg, Heidelberg, Germany; and [3]Department of Pediatrics, University of New Mexico Children's Hospital, Albuquerque, NM.

Received May 21, 2014. Accepted in revised form September 7, 2014. Originally published online November 17, 2014.

Address correspondence to Anna Köttgen, MD, MPH, Department of Internal Medicine IV, Medical Center- University of Freiburg, Berliner Allee 29, 79110 Freiburg, Germany. E-mail: anna.koettgen@uniklinik-freiburg.de

0272-6386
http://dx.doi.org/10.1053/j.ajkd.2014.09.019

is not reported, summary statistics usually can help guide the selection of SNPs with the largest effect estimate and lowest *P* values in a given study. Last, it is advantageous to have control SNPs represented on a given genotyping chip so that imputation is not required.

Finding a good positive control is challenging for GWAS in the field of kidney disease. In individuals of African descent, variants in the *APOL1* gene have been shown to associate strongly with focal segmental glomerulosclerosis, hypertension-attributed end-stage renal disease, and chronic kidney disease (CKD) from a variety of causes.[3,4] Therefore, these markers might serve as positive controls. Because these variants are ancestry specific, data checks in samples that are not of African ancestry require the use of other positive controls.

Using quantitative control phenotypes such as biomarker concentrations generally is recommended due to the superior statistical power to detect associations with continuous phenotypes compared with binary phenotypes (Box 1). Further, many of them are widely available. However, reduced kidney function influences blood concentrations of many biomarkers by altering their production, metabolism, and/or elimination. As a result, the genetic influence on marker concentrations can become less apparent.

A straightforward solution would be to use a biomarker with extrarenal production and, at least partially, extrarenal elimination or a phenotypic trait unaffected by decreased kidney function. As an alternative, when evaluating the positive control association, it may be feasible to adjust for glomerular filtration rate (GFR) to reduce the effect of reduced kidney function on the biomarker blood concentration. In the following paragraphs, and summarized in Table 1, we present several examples and considerations.

One suitable marker with extrarenal production and largely extrarenal elimination is bilirubin. Serum bilirubin concentrations reflect the balance between its production and elimination. Although liver and kidney disease can coexist, CKD has not been described as having a major effect on hepatic function, and serum bilirubin values of patients with CKD usually are in the reference range. Polymorphisms in *UGT1A*, which encodes the bilirubin UDP (uridine diphosphate)-glucoronosyltransferase, were identified initially as associated with bilirubin concentrations in some cohorts forming the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Consortium.[5] In this study, the SNP rs6742078 explained 18% of the variance in bilirubin concentrations, with $P < 5 \times 10^{-324}$ and a 0.23-unit higher log(total serum bilirubin concentration in μmol/l) per T allele among 9,464 individuals studied. Effect sizes of this magnitude should be detectable easily at genome-wide significance even in samples of smaller size.

Row A of Table 1 shows the association result for rs6742078 with log(serum bilirubin concentration) in 1,527 participants of European American ancestry of the CRIC (Chronic Renal Insufficiency Cohort) Study.[6] Allele frequency, effect direction, and effect size were consistent with previously published results,

Box 1. Workflow to Select Phenotypes and Genetic Markers to Assess Known Associations for Quality Control Purposes

Step 1: Select 1, or preferably several, available control phenotypes. Preference should be given to phenotypes/traits that are continuous and measured (eg, biomarker concentrations) rather than to those that are self-reported. Ideally, the chosen biomarker is not generated in the kidney and does not exhibit net renal secretion or reabsorption.

Step 2: Look for previous GWAS of the corresponding phenotype/trait in the GWAS Catalog (www.genome.gov/gwastudies) and in PubMed. Ensure that the published association was found among individuals of the same ancestry as your study population, the GWAS was sufficiently powered (large sample size), and the findings were replicated.

Step 3: Among significantly associated markers (typically $P < 5 \times 10^{-8}$), select that which explains the largest amount of phenotype/trait variance. If this is not reported in the original publication, select for large effect size estimates and low *P* values instead. If several markers can be considered, prefer those with high minor allele frequencies and those that have been genotyped (rather than imputed) in your own study.

Step 4: In your study, to the extent possible, model the association between control trait and marker in the same way as was done in the original report, including trait transformation and units. Ensure that the modeled allele and strand match those in the published report of the association.

Step 5: Compare direction and effect size of your association to the published result. Also assess whether the *P* value meets statistical significance in your study, but (especially in smaller studies) do not expect the *P* value to be as low as those initially published, which often originate from very large meta-analyses.

Step 6: If blood concentrations of the chosen biomarker might be influenced by kidney function, rerun the association analyses adjusting for eGFR.

Step 7: If the positive control does not show the expected direction of association or the magnitude of effect differs substantially, attempt to evaluate at least a second control trait. A typical mistake that can cause the repeated absence of known associations (and is not identified in any other data checks such as quality control, exploratory data analysis, data cleaning of phenotype and genotype information, and repetition of association analyses using a different statistical program) is a mismatch of the order of individuals in the phenotype and in the genotype file. This mistake results in the random shuffling of genotypes and phenotypes, giving rise to null associations.

Abbreviations: eGFR, estimated glomerular filtration rate; GWAS, genome-wide association study.