

## ORIGINAL ARTICLE

# Using multiple measures for quantitative trait association analyses: application to estimated glomerular filtration rate

Adrienne Tin<sup>1</sup>, Elizabeth Colantuoni<sup>1</sup>, Eric Boerwinkle<sup>2</sup>, Anna Kottgen<sup>3,4</sup>, Nora Franceschini<sup>5</sup>, Brad C Astor<sup>6</sup>, Josef Coresh<sup>1</sup> and Wen Hong Linda Kao<sup>1</sup>

Studies of multiple measures of a quantitative trait can have greater precision and thus statistical power compared with single-measure studies, but this has rarely been studied in the relation to quantitative trait measurement error models in genetic association studies. Using estimated glomerular filtration rate (eGFR), a quantitative measure of kidney function, as an example we constructed measurement error models of a quantitative trait with systematic and random error components. We then examined the effects on precision of the parameter estimate between genetic loci and eGFR, resulting from varying the correlation and contribution of the error components. We also compared the empirical results from three genome-wide association studies (GWAS) of kidney function in 9049 European Americans: a single measure model, a three-measure model of the same biomarker of kidney function and a six-measure model of different biomarkers of kidney function. Simulations showed that given the same amount of overall errors, inclusion of measures with less correlated systematic errors led to greater gain in precision. The empirical GWAS results confirmed that both the three- and six-measure models detected more eGFR-associated genomic loci with stronger statistical association than the single-measure model despite some heterogeneity among the measures. Multiple measures of a quantitative trait can increase the statistical power of a study without additional participant recruitment. However, careful attention must be paid to the correlation of systematic errors and inconsistent associations when different biomarkers or methods are used to measure the quantitative trait.

*Journal of Human Genetics* (2013) 58, 461–466; doi:10.1038/jhg.2013.23; published online 28 March 2013

**Keywords:** biomarkers; genome-wide association study; glomerular filtration rate; kidney function; repeated measures

## INTRODUCTION

Genome-wide association studies (GWAS) have discovered many robust, albeit modest, genetic associations for quantitative traits.<sup>1</sup> This is facilitated through large consortia that are comprised of many individual studies to attain appropriate statistical power. While increasing sample size can enhance statistical power by increasing the precision of the genetic parameter estimate, another way to increase the precision is to increase the number of phenotype measures per individual. In epidemiological studies, multiple measures of a phenotype within an individual can represent either: (1) repeated measures from a single assessment method (for example, weight over multiple years) or (2) multiple measures assessed using different methods (for example weight, percent body fat, waist-to-hip circumference to represent obesity). The impact of multiple measures on the precision of the genetic parameter, and hence the required sample size, will depend on the systematic and random

errors of the phenotypic measurement and the correlations between the measures within each individual. Formulas and procedures exist for calculating sample size for detecting associations with correlated data.<sup>2,3</sup> However, the impact of different types of measurement errors and the application of multiple-measure models in GWAS remains to be evaluated.

We therefore performed a simulation study, along with examination of an empirical data set, to assess how different types of measurement errors affect the precision of the genetic parameter estimate in genetic association analyses of multiple measures. Both the simulation and the empirical studies were based on estimated glomerular filtration rate (eGFR), a quantitative marker of kidney function. GFR is often estimated using serum creatinine (eGFR<sub>scr</sub>) because the direct measure of GFR is often impractical in both clinical and research settings.<sup>4</sup> Other biomarkers, such as serum cystatin C (CysC),  $\beta$ -trace protein (BTP) and  $\beta$ -2 microglobulin (B2M), have

<sup>1</sup>Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; <sup>2</sup>Human Genetics Center, University of Texas School of Public Health, Houston, TX, USA; <sup>3</sup>Department of Internal Medicine, Renal Division, University Medical Center Freiburg, Freiburg, Germany; <sup>4</sup>Renal Division, Freiburg University Hospital, Freiburg, Germany; <sup>5</sup>Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA and <sup>6</sup>School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA

Correspondence: Dr A Tin, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA.

E-mail: atin@jhsph.edu

Received 20 November 2012; revised 27 February 2013; accepted 3 March 2013; published online 28 March 2013

also been used as kidney function biomarkers.<sup>5–8</sup> In the estimation of GFR, there will be systematic errors that are the properties of each biomarker, and there will also be random errors due to day-to-day physiological change and laboratory measurement errors.

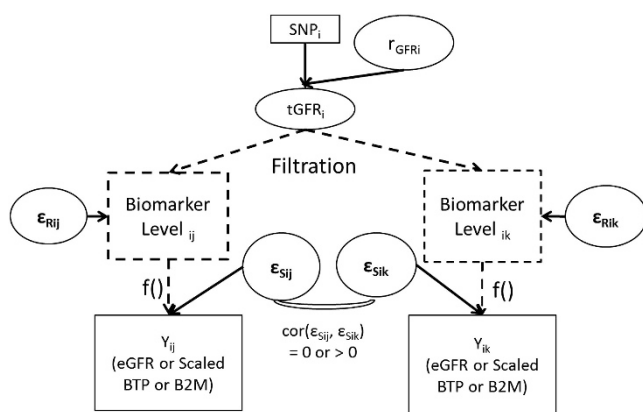
In the simulation study, we examined the impact of systematic and random errors, along with increasing the number of phenotypic measures, on the precision of the genetic parameter estimate. In the empirical data analysis, we aimed to answer the following questions:<sup>1</sup> to what extent do the longitudinal measures of eGFR based on SCr increase the precision of the genetic parameter estimate; and<sup>2</sup> does the addition of measures based on non-creatinine biomarkers provide further gain in the precision and reduce bias in detecting genetic associations?

## MATERIALS AND METHODS

### Simulation study

**GFR measurement error model.** In our GFR measurement error models, the observed outcome,  $Y_{ij}$  (representing eGFR or any other biomarker-based kidney function index), was determined by three latent components: (1) an individual's average stable true GFR (tGFR); (2) systematic errors ( $\epsilon_S$ ), including biomarker-specific interindividual variability, unrelated to the tGFR and not accounted for in the GFR estimating equation; and (3) random errors ( $\epsilon_R$ ), such as laboratory measurement errors and within-individual day-to-day physiological variations in GFR or biomarker levels. For the  $j$ th and  $k$ th observations of individual  $i$ , the correlation between the outcomes,  $(Y_{ij}, Y_{ik})$ , was determined by tGFR and systematic errors, which may or may not be correlated within an individual depending on the method of measurement. In our specification of the measurement error model, the three latent components of the outcome (tGFR, systematic errors and random errors) were assumed to be independent with standard normal distribution, Normal(0, 1). Figure 1 presents a GFR measurement error model of two measures. Details of the specification of the measurement error model are described in the Supplementary Methods section.

**Evaluation of data sets with complete data.** Using four models (summarized in Table 1), we investigated the impact of varying (1) the overall measurement errors, and (2) the correlation between the systematic errors,  $(\epsilon_{Sij}, \epsilon_{Sik})$ , on the



**Figure 1** GFR measurement error model. The SNP effect is assumed to act solely through tGFR.  $Y_{ij}$  is a function of tGFR, systematic errors ( $\epsilon_S$ ) and random errors ( $\epsilon_R$ ). The elements with dotted lines show the process of estimating GFR based on biomarker levels, and are not part of the measurement error model. **Notations and assumptions:** tGFR, true GFR, average stable GFR, standardized to  $N(0, 1)$ ;  $i$ , individual;  $j$  or  $k$ , observation within an individual;  $r_{GFRi}$ , tGFR level not explained by the SNP;  $Y_{ij}$ , outcome in a regression model, calculated from the biomarker level using an equation,  $f()$ ;  $\epsilon_S$ , systematic errors in  $Y_{ij}$ ;  $\epsilon_R$ , random errors in  $Y_{ij}$ ;  $\epsilon_S, \epsilon_R \sim N(0, 1)$ .

gain in precision of the genetic parameter estimate in multiple-measure models. We assumed that the causal single-nucleotide polymorphism (SNP) explained 0.5% of tGFR variance without measurement errors. The contribution of tGFR to the variance of the outcome,  $Y_{ij}$  was set to 0.7 in models 1 and 2, and reduced to 0.5 in models 3 and 4. Therefore, the percentage of the variance of  $Y_{ij}$  explained by the SNP was 0.35% for models 1 and 2 and 0.25% for models 3 and 4, similar to the modest effect size of the index SNPs in eGFR GWAS results.<sup>9</sup> The setting of 0.7 for the contribution of tGFR in models 1 and 2 was based on unpublished data from the Modification of Diet in Renal Disease (MDRD) Study and the African-American Study of Kidney Disease (AASK). In these two studies, the correlation between eGFR and tGFR, estimated from urinary clearance of <sup>125</sup>I-iothalamate (gold standard), was approximately 0.9 in patients with chronic kidney disease. This implies that the contribution of tGFR to the variance of eGFR was approximately 0.81 ( $= 0.9^2$ ). In the general population, the contribution of tGFR to eGFR variance would be lower.<sup>10</sup>

In models 1 and 2, the contribution of random errors to the variance of the outcome was set to 0.1 based on the estimates of within-individual variation in SCr<sup>11</sup> and measured GFR.<sup>11,12</sup> Therefore, the contribution of systematic errors to the variance of the outcome was set to 0.2 ( $= 1 - 0.7 - 0.1$ ). The contributions of systematic errors in models 3 and 4 were kept at the same level. The systematic errors,  $(\epsilon_{Sij}, \epsilon_{Sik})$ , where  $j \neq k$ , were assumed to be uncorrelated in models 1 and 3, and have a covariance of 0.5 in models 2 and 4.

From the GFR measurement error model parameters described above, we estimated the observed covariance of the outcomes and the residual variance of a generalized least-square (GLS) regression model (Table 1). GLS is a common method for modeling multiple correlated continuous outcomes and can estimate the association between a predictor and the outcomes taking into account correlated residuals. In this paper, we reserve the term “errors” to refer to the latent error components in the measurement error model and use “residuals” to refer to the portion of the outcome unexplained by predictor(s) in a GLS regression model.

We expressed the gain in precision of the genetic parameter estimate in a multiple-measure model over a single-measure model in terms of the estimated “change in equivalent sample size”. The formula for the estimate of the gain in precision is provided in Supplementary Methods section. This measure is relevant in situations where an investigator might be deciding between increasing power through additional recruitment of study participants (thus increasing  $Y_i$ ) or adding another outcome measure in the existing population (adding  $Y_{ik}$ ). Equivalent sample size was defined as the sample size in a single-measure ordinary least-square regression that would provide the same power as a multiple-measure model using GLS given the same effect size and  $\alpha$ -level.

**Evaluation of data sets with data missing completely at random.** Based on the above measurement error models, we simulated data sets with sample sizes of either 3000 or 6000 to assess the impact of randomly missing data on the gain of precision in the genetic parameter estimate in multiple-measure models. The mechanism of missing completely at random (MCAR)<sup>13</sup> was deemed to be appropriate for this study because our focus was the gain in precision of the parameter estimate instead of evaluating biases in parameter estimate. Each data set had three repeated outcome measures,  $Y_{i1}$ ,  $Y_{i2}$  and  $Y_{i3}$ , with a variance of 1 and a constant modest SNP effect size ( $\beta = 0.075 \times \text{s.d. of } Y_{ij}$ ). If the SNP has an allele frequency of 35%, it would explain about 0.25% of the variance of  $Y_{ij}$  similar to the SNP effects in models 3 and 4 in Table 1. Missing data rates in scenario 1 were 10% for the second measure and 25% for the third measure. The rates increased to 20 and 40% in scenario 2. With each missing data scenario, we simulated data sets with correlations between  $Y_{ij}$  and  $Y_{ik}$  ranging from 0.5 to 0.8. As the SNP effect was assumed to be constant across the outcome measures, the change in the correlations between  $Y_{ij}$  and  $Y_{ik}$  was assumed to be solely due to the change in the correlation of systematic errors ( $\epsilon_S$ ). Residuals were generated with a distribution of Normal(0, 1) and then transformed to have the desired correlation by multiplying the Cholesky decomposition of a variance-covariance matrix. Supplementary Table 1 presents the simulation parameters. In all, 10 000 iterations were performed

**Table 1** GFR measurement error models

Model	Overall			GFR measurement error model parameters (unobserved)						Estimates based on GFR measurement error model parameters				
	measurement errors	Systematic errors correlated	% tGFR variance explained by SNP	$w_1^2$	$w_2^2$	$w_3^2$	cov ( $\epsilon_{Sij}, \epsilon_{Sjk}$ )		cov ( $\epsilon_{Rij}, \epsilon_{Rjk}$ )		% of Var( $Y_{ij}$ ) explained by SNP	$\tau^2$	$\sigma^2$	cov( $Y_{ij}, Y_{ik}$ )
1	Less	No	0.5	0.7	0.2	0.1	0		0		0.35	0.69	0.30	0.70
2	Less	Yes	0.5	0.7	0.2	0.1	0.5		0		0.35	0.79	0.20	0.80
3	More	No	0.5	0.5	0.2	0.3	0		0		0.25	0.49	0.50	0.50
4	More	Yes	0.5	0.5	0.2	0.3	0.5		0		0.25	0.59	0.40	0.60

Abbreviations: SNP, single-nucleotide polymorphism; tGFR, true glomerular filtration rate.

The equivalent sample sizes of models 3 and 4 were indexed to the sample size of model 1 accounting for the assumed smaller effect of the SNP due to the smaller contribution of tGFR. Supplementary Figure 2 presents these results with error bars indicating 95% confidence intervals assuming a sample size of 10 000.

Notation and formula used:

$w_1^2$ ,  $w_2^2$  and  $w_3^2$  are the contributions of tGFR, systematic errors and random errors to the variance of the outcome,  $Y_{ij}$ .

% of Var( $Y_{ij}$ ) explained by SNP = % of tGFR variance explained by SNP  $\times w_1^2$ .

$\tau^2 = w_1^2 \times$  % of Var( $Y_{ij}$ ) explained by SNP +  $w_2^2 \times \text{cov}(\epsilon_{Sij}, \epsilon_{Sjk})$ .

$\sigma^2 = w_2^2 \times (1 - \text{cov}(\epsilon_{Sij}, \epsilon_{Sjk})) + \text{cov}(\epsilon_{Rij}, \epsilon_{Rjk})$ .

cov( $Y_{ij}, Y_{ik}$ ) =  $w_1^2 + w_2^2 \times \text{cov}(\epsilon_{Sij}, \epsilon_{Sjk})$ .

for data sets with a sample size of 3000, and 6000 iterations were performed for data sets with a sample size of 6000.

With each data set, we performed three analyses: (1) an ordinary least-square regression using the first measure ( $Y_{i1}$ ) as outcome; (2) a GLS regression using  $Y_{i1}$  and  $Y_{i2}$  as outcomes; and (3) a GLS regression using  $Y_{i1}$ ,  $Y_{i2}$  and  $Y_{i3}$  as outcomes. Changes in equivalent sample size were calculated using Equation (3) in Supplementary Methods section based on the variance of the SNP parameter estimates generated in the simulations. To obtain the 95% confidence interval of the variance of the SNP parameter estimate, we sampled the standard errors (s.e.) of the SNP parameter estimates in the single- and multiple-measure models separately, and then calculated the square of the ratio of the SEs. After repeating this procedure 1000 times, we obtained the 0.025 and 0.975 percentile for the 95% confidence interval. SAS 9.2 PROC GLM was used for ordinary least-square regression, and PROC MIXED with the repeated statement was used for GLS regression. A template of the SAS macro for running the association analysis was included in Supplementary Materials.

### Empirical data

**Study population.** The ARIC study is a prospective observational cohort study of 15 792 middle-aged adults (baseline age between 45 and 64 years) in four US communities. Details of the study design were reported previously.<sup>14</sup> Since the known genomic risk loci for reduced eGFR were detected in populations of European ancestry, only the ARIC European American cohort ( $n = 9049$ ) was included in this analysis.

**Phenotype and genotype in empirical data set.** In the ARIC study, the following measures of kidney function were available: three repeated measures of SCr at visits 1, 2 and 4 and measures of serum CysC, BTP and B2M at visit 4 (Supplementary Figure 1). The Supplementary Methods section reports the measurement methods of these biomarkers and the calculation of the outcome measures: eGFRscr and eGFR based on CysC, scaled BTP and scaled B2M. Over two million imputed SNPs were evaluated in the analysis. Details on genotyping and quality control are reported in the Supplementary Methods section.

**GWAS statistical analysis.** Three genome-wide scans were performed: (1) a single-measure model using eGFRscr at visit 1 as outcome; (2) a three-measure model using eGFRscr at visits 1, 2 and 4 as outcomes; and (3) a six-measure model using the three repeated measures of eGFRscr and the measures of eGFR based on CysC, scaled BTP and scaled B2M as outcomes. Covariates included age, gender, study center and the first 10 principal components with significant association with the outcome ( $P < 0.05$ ). The three- and six-measure models additionally included visit as a categorical covariate.

The single-measure model was analyzed using ProbABEL.<sup>15</sup> The multiple-measure models were analyzed using SAS 9.2 PROC MIXED with the repeated statement and a prespecified variance-covariance matrix to optimize performance. The Supplementary Methods section reports the generation of this variance-covariance matrix.

We calculated the genomic control factor ( $\lambda_{GC}$ ) for the results of each genome-wide scan to assess possible test statistic inflation and corrected the  $P$ -values when  $\lambda_{GC} > 1$ .<sup>16</sup> The model comparisons were based on the genomic control-corrected  $P$ -value (PValGC).

In addition, for the index SNPs of 16 known eGFR loci,<sup>9</sup> we performed separate regression analyses using standardized outcome measures to obtain standardized SNP parameter estimates and standard errors of the single-measure model and the three- and six-measure models.

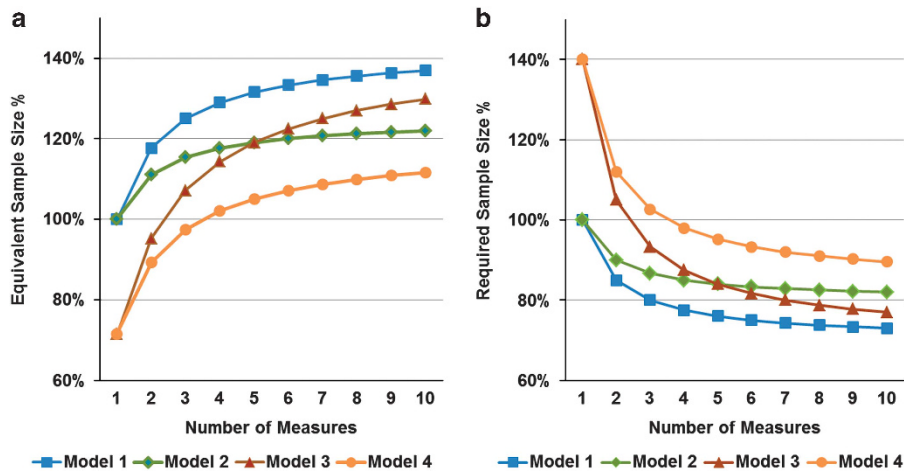
**Comparisons of the single- and multiple-measure models in the empirical data set.** The assumption of constant effect size across measures did not hold in the empirical data because the association between a SNP and a biomarker could change over time, and the association between an SNP and different biomarkers could vary. Therefore, we did not use the change in equivalent sample size as a metric for comparison in the empirical study. Instead, we compared the effect estimates of the index SNPs of the 16 known eGFR loci from the single-, three- and six-measure models and the change in standard error due to multiple measures. Next, we compared the GWAS results of the three models with respect to the number of loci with  $\text{PValGC} < 5 \times 10^{-8}$ . Only SNPs with minor allele frequency  $> 5\%$  were included.

## RESULTS

### Effect of multiple measures on change in equivalent sample size assuming complete data

Figure 2a shows the relationship between the change in equivalent sample size and the number of outcome measures in the four models described in Table 1. Figure 2b shows the reciprocal of equivalent sample size as required sample size. Equivalent sample size may be more intuitive when an investigator only has the option of obtaining new measures given a fixed sample size, whereas the change in required sample size may be useful when an investigator have the option of varying both the numbers of measures or participants. For these results, we varied the following parameters: (1) the number of outcome measures; (2) the total measurement errors; and (3) the correlation between systematic errors.

Under the assumptions of no missing data and constant effect size across measures, adding additional outcome measures always led to a gain in estimated equivalent sample size. Assuming an uncorrelated



**Figure 2** (a) Change in equivalent sample size (%) in the four measurement error models. The equivalent sample sizes of models 3 and 4 were indexed to the sample size of model 1 accounting for the smaller  $\beta$  value of the genetic parameter estimate due to the smaller contribution of tGFR. (b) Change in required sample size (= 1/equivalent sample size)% from the four measurement error models.

residual variance,  $\sigma^2$ , of 0.3 as in model 1, an increase of up to 10 measures led to 37% gain in equivalent sample size; however, this gain leveled off around five or six measures.

The relative gain in equivalent sample size with each additional measure was determined by the uncorrelated residual variance,  $\sigma^2$ , as shown in Equation (2) in Supplementary Methods section. For a fixed total measurement error, as in models 1 and 2, the model with less correlated systematic errors,  $\varepsilon_S$ , had relatively higher uncorrelated residual variance,  $\sigma^2$ , and resulted in more gain in equivalent sample size. For example, in model 1, the addition of a second measure led to an 18% gain in equivalent sample size but only 11% in model 2. Model 3 outperformed model 4 for the same reason.

Next, we also compared the impact of varying both the total measurement error and the correlation between systematic errors. Comparing models 2 and 3, model 2 had lower overall measurement errors but higher correlated residuals due to correlated systematic errors. The higher correlated residuals in model 2 led to a smaller gain in precision than model 3 with each additional measure. After the fifth measure, model 3 exceeded model 2 in estimated equivalent sample size. Supplementary Table 2 presents the changes in equivalent sample size with 95% confidence interval for sample sizes of 3000 and 10000. Even though the estimate of the expected gain in equivalent sample size with additional measures is independent of the sample size when assuming complete data, the 95% confidence intervals of the estimate are narrower with larger sample size.

#### Effect of multiple measures on change in equivalent sample size assuming data MCAR

Supplementary Table 3 shows the change in expected equivalent sample size when the data were MCAR. Similar to the results based on complete data, the gain in equivalent sample size was higher when the residuals were less correlated. As expected, higher missing data rate resulted in less gain in equivalent sample size with each additional outcome measure. When the outcome measures had a correlation of 0.5, the gains in equivalent sample size for adding a second measures with 0%, 10% and 20% missing data were 33%, 30% and 27%, respectively. Even with a missing data rate as high as 40%, there was still gain in equivalent sample size.

#### Application to kidney function measures in ARIC

Supplementary Table 4 reports the sample sizes, means, standard deviations and correlations of the outcome measures of kidney function in the empirical study. The correlations between eGFRscr across the three visits ranged from 0.63 to 0.69. The correlations between eGFRscr and measures of kidney function based on other biomarkers were lower. The lowest correlation was 0.34 between eGFRscr at visit 1 and scaled BTP at visit 4, and the highest correlation was 0.72 between eGFR based on CysC and scaled B2M both at visit 4.

#### Comparison of the results for 16 known eGFR-associated SNPs

We tested for the associations between kidney function and the index SNPs (with the SNP with the lowest  $P$ -value) at 16 known eGFR-associated loci using (1) a single-measure model with eGFRscr at visit 1 as the outcome; (2) a three-measure model including eGFRscr from visits 1, 2 and 4; (3) a six-measure model including both the repeated and the multiple measures of kidney function derived from different biomarkers; and (4) a single-measure model with eGFRscr at visit 4 as the outcome and the comparison of this results with a four-measure model including all four measures of kidney function at visit 4.

Overall, for most of the 16 index SNPs, the multiple-measure models resulted in lower association  $P$ -values due to the gain in precision of the beta estimates of the SNP effect (Supplementary Figure 3 and Supplementary Table 5). The standard error reduction of the multiple-measure models over the single-measure model was 12% for the three-measure model and 21% for the six-measure model. Compared with the single-measure model, the three-measure model had 15 index SNPs with lower  $P$ -values; five of them were at least one order of magnitude lower. Again, compared with the single-measure model, the six-measure model has 12 index SNPs with lower  $P$ -value; seven of them were at least one order of magnitude lower. The six-measure model resulted in larger  $P$ -value than the single-measure model at four loci due to weaker associations of the index SNPs with the non-creatinine biomarkers at *TFDP2* and *ANXA9* and opposite effect directions of the index SNPs with scaled BTP at *PIP5K1B* and *DAB2*. Supplementary Figure 4 shows the standardized  $\beta$ -estimates of the 16 index SNPs when regressed separately against the outcomes calculated from the four biomarkers at visit 4 of the ARIC study.

For 8 of the 16 index SNPs, the  $\beta$ -estimates against eGFRscr were larger than those against the non-creatinine-based outcomes. For the index SNPs of three loci (*ATXN2*, *PIP5K1B* and *DAB2*), the  $\beta$ -estimates against scaled BTP were in opposite directions from the estimates against the other three outcomes. Supplementary Table 6 presents the 95% confidence intervals of the estimates and *P*-values. Compared with the single-measure model of eGFRscr at visit 4 versus the four-measure model at visit 4, all the  $\beta$ -estimates were in the same direction. The standard error reduction from the four model was approximately 18%. However, only 9 of the 16 SNPs had lower *P*-values in the four-measure model (Supplementary Table 7).

### GWAS results

To determine whether the additional outcome measures would result in the identification of additional kidney loci in genome-wide scans, we performed three GWAS analyses. With respect to loci that reached genome-wide significance ( $P_{\text{ValGC}} < 5 \times 10^{-8}$ ), the single-measure model identified one locus (*NAT8*), the three-measure model identified three loci (*NAT8*, *SHROOM3* and *SPATA5L1*) and the six-measure model identified two loci (*NAT8* and *SHROOM3*; Supplementary Table 8). All loci had previously been discovered and replicated in a much larger sample.<sup>9</sup> One of the significant loci from the three-measure model of eGFRscr, *SPATA5L1*, was not significant in the six-measure model. This locus has been suspected to be a genetic locus related to creatinine production rather than kidney function, as evidenced by the lack of association with non-creatinine kidney function biomarkers.<sup>9</sup> One of the genes in this locus, *GATM*, encodes the rate-limiting enzyme in creatinine biosynthesis.<sup>17</sup> This six-measure model result suggests that the non-creatinine-based biomarkers reduced biases due to the correlated systematic errors of the creatinine-based outcomes.

### DISCUSSION

Using both simulated and empirical data sets, we showed that increasing the number of outcome measures per individual led to gains in equivalent sample size, and thus a gain in power, in genetic association analyses when the genetic effects were similar across measures. In addition, less correlated systematic errors led to greater gains in equivalent sample size. The marginal gain decreased with each additional measure and leveled off around the addition of the fifth or sixth measure. Lui and Cumberland<sup>18</sup> made similar observations in the situation of two-group complete balanced data.<sup>18</sup> The gain in equivalent sample size was relatively robust to data MCAR as the gain persisted even when the missing data rate was as high as 40%.

The results from our simulation study were corroborated by the results from the empirical study of multiple measures of kidney function in the ARIC study. We showed that inclusion of eGFRscr from three separate study visits (the three-measure model) was more powerful than the single-measure model using eGFRscr at visit 1. However, the addition of other biomarkers of kidney function, including eGFR based on CysC, scaled BTP and scaled B2M, did not make the six-measure model more powerful than the three-measure model despite additional gain in the precision of the SNP parameter estimates due to the heterogeneity of SNP associations with the different biomarkers. While longitudinal repeated measures of a trait can be used to estimate change over time, our study focused on the use of repeated measures to detect associations that are similar across multiple measures. We estimated the mean effect over time and not change over time.

When studying multiple measures of an outcome, one can consider either repeated measures using the same method or multiple measures of an underlying trait using different methods, such as the use of different biomarkers for kidney function in this work. The contrast between these two scenarios was represented by the results of the three-measure model and the six-measure model from our empirical study. For repeated measures of the same outcome, correlated systematic errors between the multiple measures of the outcome may limit the gain in precision of the SNP parameter estimate. When using multiple measures based on different methods to represent one underlying trait, some measures may contain additional measurement errors, which reduce the statistical power of the study. In the kidney function empirical study, the inclusion of additional non-creatinine-based outcomes did not identify more loci with lower *P*-values. The non-creatinine-based outcomes may have more measurement errors due to the lack of population-based equations for calculating eGFR based on these biomarkers. Therefore, regardless of the number of measures, well-measured phenotypes that minimize measurement errors are important for detecting associations, a topic that has been studied extensively.<sup>19</sup>

A few studies have used repeated measures in the setting of genome-wide association studies and have found mixed results with respect to the gain in efficiency. Rasmussen-Torvik *et al.*<sup>20</sup> compared the results from using the average of four repeated measures of fasting glucose over 12 years ( $N = 5782$ ) to the results from four separate GWAS of fasting glucose from each study visit ( $N$  ranged from 8372 at visit 1 to 6421 at visit 4) and found that, despite a smaller sample size, the results from the analysis of the average fasting glucose values were stronger. The *P*-values of the index SNP at five candidate regions were lower by three to eight orders of magnitude, mostly due to reduction in standard errors. This suggests that the average of a trait can reduce intraindividual variations and lead to stronger statistical associations. On the other hand, Malhotra *et al.*<sup>21</sup> conducted GWAS of body mass index (BMI) that used up to 17 repeated measures and the maximum BMI (from 1965 to 2004) in 1120 Pima Indians.<sup>21</sup> No genome-wide significant loci were identified. Of the 20 top SNPs reported from the repeated-measure analysis, nine had *P*-values that were lower than their corresponding *P*-values from the analysis using the maximum BMI, and the differences in *P*-values were less than two orders of magnitude. The gain in efficiency from the repeated measures was not apparent, which is possible if maximum BMI captures an individual's overall disposition toward obesity better than repeated measures of BMI over a very long period of time where BMI might fluctuate greatly.

One limitation of our work is that we only used GLS for analyses of multiple outcome measures and did not evaluate other methods. Ferreira and Purcell<sup>22</sup> proposed the use of canonical correlation analysis for analyzing correlated outcomes in GWA studies. Coin *et al.*<sup>23</sup> proposed using multiple phenotypes as predictors and a genetic variant as outcome in a regression model. Both of these methods require the use of complete data. In addition, as it was shown in the GFR measurement error model, the correlation between measures can come from two sources: the true measure of the trait of interest and correlated systematic errors, which does not help the detection of genetic associations of the trait. Therefore, the application of methods for combining multiple measures requires some assumptions and understanding of the measurement error model of the trait of interest.

Other limitations of this work include the assumption of one tGFR and constant covariance of outcomes in the measurement error models. The underlying latent trait may change over time, and the covariance structure among outcome measures may be complex.

However, the basic conclusion of this work holds for multiple latent traits and complex covariance structures. Regardless of the specific covariance structure, correlated systematic errors reduce the gain in precision when using multiple measures.

GWAS provide a systematic, unbiased way to identify genes and pathways underlying a biological process.<sup>24,25</sup> Very large sample sizes have been used to increase the precision of the genetic parameter estimates, thus increasing the power to identify loci of modest effect sizes.<sup>26</sup> Increasing sample size through additional participant recruitment can be expensive and sometimes not feasible. Therefore, using multiple measures of an outcome is another way to increase the statistical power of a study, especially for population-based cohort studies that have often collected multiple measures for prospective analyses of an outcome. Our findings can inform the choice of measures in the design of a multiple-measure study.

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

### ACKNOWLEDGEMENTS

We thank the staff and participants of the ARIC study for their important contributions. The Atherosclerosis Risk in Communities Study was carried out as a collaborative study supported by National Heart, Lung and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

- 1 Hindorf, L., MacArthur, J. E. B. I., Wise, A., Junkins, H., Hall, P., Klemm, A. *et al*. A Catalog of Published Genome-Wide Association Studies. Available at [www.genome.gov/gwastudies](http://www.genome.gov/gwastudies) (accessed on 12 June 2012).
- 2 Liu, G. Jr. & Liang, K. Y. Sample size calculations for studies with correlated observations. *Biometrics* **53**, 937–947 (1997).
- 3 Diggle, P., Heagerty, P., Liang, K. Y. & Zeger, S. *Analysis of Longitudinal Data*. 2nd edn (Oxford University Press, Oxford, UK, 2002).
- 4 National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Am. J. Kidney Dis.* **39** (suppl. 1), S1–S266 (2002).
- 5 Stevens, L. A., Coresh, J., Schmid, C. H., Feldman, H. I., Froissart, M., Kusek, J. *et al*. Estimating GFR using serum cystatin C alone and in combination with serum

- creatinine: a pooled analysis of 3418 individuals with CKD. *Am. J. Kidney Dis.* **51**, 395–406 (2008).
- 6 Bianchi, C., Donadio, C., Tramonti, G., Consani, C., Lorusso, P. & Rossi, G. Reappraisal of serum beta2-microglobulin as marker of GFR. *Ren. Fail.* **23**, 419–429 (2001).
- 7 White, C. A., Akbari, A., Doucette, S., Fergusson, D., Hussain, N., Dinh, L. *et al*. Estimating GFR using serum beta trace protein: accuracy and validation in kidney transplant and pediatric populations. *Kidney Int.* **76**, 784–791 (2009).
- 8 Astor, B. C., Shafi, T., Hoogeveen, R. C., Matsushita, K., Ballantyne, C. M., Inker, L. A. *et al*. Novel markers of kidney function as predictors of ESRD, cardiovascular disease, and mortality in the general population. *Am. J. Kidney Dis.* **59**, 653–662 (2012).
- 9 Kottgen, A., Pattaro, C., Boger, C. A., Fuchsberger, C., Olden, M., Glazer, N. L. *et al*. New loci associated with kidney function and chronic kidney disease. *Nat. Genet.* **42**, 376–384 (2010).
- 10 Levey, A. S. Measurement of renal function in chronic renal disease. *Kidney Int.* **38**, 167–184 (1990).
- 11 Myers, G. L., Miller, W. G., Coresh, J., Fleming, J., Greenberg, N., Greene, T. *et al*. Recommendations for improving serum creatinine measurement: a report from the Laboratory Working Group of the National Kidney Disease Education Program. *Clin. Chem.* **52**, 5–18 (2006).
- 12 Kwong, Y. T., Stevens, L. A., Selvin, E., Zhang, Y. L., Greene, T., Van Lente, F. *et al*. Imprecision of urinary iothalamate clearance as a gold-standard measure of GFR decreases the diagnostic accuracy of kidney function estimating equations. *Am. J. Kidney Dis.* **56**, 39–49 (2010).
- 13 Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* (Wiley, New York, NY, USA, 2002).
- 14 ARIC. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.* **129**, 687–702 (1989).
- 15 Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinform.* **11**, 134 (2010).
- 16 Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.* **60**, 155–166 (2001).
- 17 Sandell, L. L., Guan, X. J., Ingram, R. & Tilghman, S. M. Gattm, a creatine synthesis enzyme, is imprinted in mouse placenta. *Proc. Natl Acad. Sci. USA* **100**, 4622–4627 (2003).
- 18 Lui, K. J. & Cumberland, W. G. Sample size requirement for repeated measurements in continuous data. *Stat. Med.* **11**, 633–641 (1992).
- 19 Wojczynski, M. K. & Tiwari, H. K. Definition of phenotype. *Adv. Genet.* **60**, 75–105 (2008).
- 20 Rasmussen-Torvik, L. J., Alonso, A., Li, M., Kao, W., Kottgen, A., Yan, Y. *et al*. Impact of repeated measures and sample selection on genome-wide association studies of fasting glucose. *Genet. Epidemiol.* **34**, 665–673 (2010).
- 21 Malhotra, A., Kobes, S., Knowler, W. C., Baier, L. J., Bogardus, C. & Hanson, R. L. A genome-wide association study of BMI in American Indians. *Obesity (Silver Spring, MD)* **19**, 2102–2106 (2011).
- 22 Ferreira, M. A. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25**, 132–133 (2009).
- 23 Coin, L., O'Reilly, P., Pompyen, Y. & Hoggart, C. F. C. MultiPhen: joint model of multiple phenotype increases discovery in GWAS. Presented at ASHG/ICHG. 13 October 2011, Montreal, Canada.
- 24 Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
- 25 Hirschhorn, J. N. Genomewide association studies—illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).
- 26 Park, J. H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J. *et al*. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)