Estimating the overall fraction of phenotypic variance attributed to high-dimensional predictors measured with error

Soutrik Mandal¹, Do Hyun Kim², Xing Hua³, Shilan Li¹, Jianxin Shi¹

¹Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA. ²: Department of Biostatistics, Fielding School of Public Health, University of California at Los Angeles (UCLA), Los Angeles, CA 90095. ³: Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109.

Correspondence to:

Jianxin Shi, Ph.D.

Senior Investigator

Biostatistics Branch

Division of Cancer Epidemiology and Genetics

National Cancer Institute

Address: 9609 Medical Center Drive, RM 7E122, Bethesda, MD 20892.

Tel: (240) 276-7419

Email: Jianxin.Shi@nih.gov

Abstract

In prospective genomic studies (e.g., DNA methylation, metagenomics, and transcriptomics), it is crucial to estimate the overall fraction of phenotypic variance (OFPV) attributed to the high-dimensional genomic variables, a concept similar to heritability analyses in genome-wide association studies (GWAS). Unlike genetic variants in GWAS, these genomic variables are typically measured with error due to technical limitation and temporal instability. While the existing methods developed for GWAS can be used, ignoring measurement error may severely underestimate OFPV and mislead the design of future studies. Assuming that measurement error variances are distributed similarly between causal and noncausal variables, we show that the asymptotic attenuation factor equals to the average intraclass correlation coefficients of all genomic variables, which can be estimated based on a pilot study with repeated measurements. We illustrate the method by estimating the contribution of microbiome taxa to body mass index and multiple allergy traits in the American Gut Project. Finally, we show that measurement error does not cause meaningful bias when estimating the correlation of effect sizes for two traits.

1. Introduction

Genome-wide association studies (GWAS) aim to identify genetic variants associated with a specific phenotype by examining millions of genetic variants across the genomes^{1,2}. Since the first GWAS was successfully performed for age-related macular degeneration in 2005³, hundreds of associations for many diseases have been identified. These findings have improved our understanding of disease etiology and genetic risk prediction. While GWAS have been successful as a revolutionary method for genetic mapping of complex traits and diseases, the identified few variants can only explain a small fraction of heritability estimated based on twin or family studies. This so-called “missing heritability” problem puzzled scientists. By that time, the identified few variants could only explain a small fraction of heritability estimated based on twin or family studies. For example, the approximately 180 genetic variants identified in large-scale GWAS of human height published in 2010 explained about 10% of phenotype variance⁴, much less than the 80% heritability estimated based on twin- or family-based studies. This raised a question whether GWAS would be highly productive and provide important tools for genetic risk prediction eventually.

To answer this question, linear mixed models (LMM) were developed to estimate the overall fraction of phenotypic variation (OFPV), also called narrow sense heritability in genetics, explained by all common variants that are genotyped or imputed in a GWAS⁵. Here, OFPV corresponds to the variance component in LMMs. For human height, they estimated the heritability to be approximately 45%⁵, much higher than the 10% based on the 180 variants identified by that time⁶. This analysis provided compelling evidence that majority of causal variants were not detected because of insufficient power. For case-control studies, we can use

this framework to estimate OFPV at the observational scale and then convert it to population-level OFPV by adjusting for ascertainment and disease prevalence in the target population⁷.

Recently, large-scale prospective studies with high-dimensional genomic variables have been performed or planned for various diseases. These prospective studies aim to identify genomic variables associated with disease risk to better understand disease etiology and to improve risk prediction. Examples of the genomic variables include but are not limited to microbiome taxa relative abundances, gene expression, microRNA expression, and DNA methylation. Typically, none or just a few genomic variables can be identified to be statistically significant even in a reasonably large-scale study, and the resulting predictive models have a poor performance. For example, a recent prospective peripheral blood DNA methylation study⁸ with 1,663 breast cancer cases and 1,885 controls at 365,145 CpG markers did not identify any significant association after multiple testing correction. In our recent study for never smokers from two cohorts – Shanghai Women's Health Study and the Shanghai Men's Health Study, we did not identify significant associations between oral microbiome and lung cancer risk⁹. In another oral microbiome study nested in cohorts, we identified a few significant associations for lung cancer risk¹⁰, but these associations only explained a small fraction of lung cancer risk. Similar to the puzzling “missing heritability” problem in GWAS, the lack of findings in these genomic studies may be caused by limited statistical power due to insufficient sample size or by the fact that the genomic variables do not contribute to the disease risk. Thus, it is informative to estimate OFPV, which helps decide the necessity and the scale of future studies.

While the statistical framework and software tools^{5,11} developed for GWAS can be readily used, there is important difference between these genomic variables and genetic variants in GWAS.

Genotypes of common genetic variants remain unchanged over the whole life and can be

accurately captured by genotyping arrays. However, genomic variables are typically measured with error, which comes from two sources: technical variability and temporal instability.

Ignoring measurement error in genomic variables will cause a biased estimate of OFPV.

The problem of measurement error in independent variables, or errors-in-variables models, have been intensively investigated for regression analysis¹², LMMs¹³ and generalized LMMs^{14,15}. For a simple linear regression, measurement error in the independent variables attenuates the estimate toward zero; the attenuation factor or the reliability ratio¹² equals to the intraclass correlation coefficient (ICC) of the variable. Treatment of measurement error in generalized linear models and other nonlinear models has been well discussed^{16,17}. For LMMs or generalized LMMs, methods have been developed to correct for the bias in both fixed effects and variance components when there is measurement error in fixed effect variables^{13-15,18}. When there is measurement error in both fixed and random effects, Cui and his colleagues developed moment estimators to consistently estimate both fixed and random effects¹⁹. While this framework is most relevant to our problem, it cannot be directly applied. The LMMs assume that all random effect variables are causal; however, signals are expected to be sparse in genomic studies²⁰. As was pointed out by Jiang and his colleagues²¹, the LMMs⁵ used for estimating heritability are misspecified and statistical inference needs to be performed under the assumption of sparse signals. Moreover, the covariance structure of measurement error is assumed to be known¹⁹. In practice, this needs to be estimated using a dataset of limited sample size and the uncertainty in the estimate needs to be modeled when constructing confidence intervals.

In this manuscript, we aim to derive the asymptotic attenuation factor (AAF) for estimating OFPV when the high-dimensional genomic variables are measured with error. Because bias analysis is complicated based on the LMMs using restricted maximum likelihood (REML), we

proceed with a recently developed pairwise regression that is similar to the Haseman-Elston regression^{22,23} in statistical genetics. We will show that AAF equals to the average ICC of all genomic variables under the assumption that the distributions of ICCs are identical between causal and non-causal variables. Thus, to derive a consistent estimator for OFPV, we need to perform a pilot study with repeated measurements to estimate the average ICC.

A parallel question is how to estimate the correlation of effect sizes between two traits, which was also motivated in GWAS for analyzing genetic correlation between two traits^{24,25}. We assume that the two traits are influenced by the same set of K genomic variables with effect sizes following a bivariate normal distribution with correlation ρ . Estimating ρ is crucial for understanding the degree of the shared etiology between the two traits. A joint study can be performed to achieve a high statistical power for detecting individual associations or risk prediction²⁶⁻²⁹ if multiple traits share the same etiology with a large value of ρ between primary and the secondary traits. Our analysis shows that measurement error has little impact for estimating ρ even if we ignore measurement error.

This manuscript is organized as follows. In Section 2, we derive AAF in the presence of independent and additive measurement error in genomic variables. In Section 3, we discuss estimating AAF using repeated measurements when ICCs are unknown. In Section 4, we investigate the impact of measurement error on estimating ρ for two traits. In Section 5, we present numerical studies. In Section 6, we present a data example estimating OFPV for body mass index (BMI) and multiple allergy traits explained by gut microbiome using the data from the American Gut Project. Limitations and future work are discussed in Section 7.

2. Estimating the overall fraction of phenotypic variance attributed to high-dimensional genomic variables

2.1 Notations and models

For subject i ($i = 1, \dots, N$), let y_i denote the phenotypic value and $(x_{i1}, \dots, x_{iM})$ denote the values of M independent genomic variables. Here, y_i and x_{im} have been normalized to have mean zero and unit variance. We consider an additive model $y_i = \sum_{m \in C} \beta_m x_{im} + \varepsilon_i$, where C denotes the set of K causal variables. In the narrow sense, $\lambda = \sum_{m \in C} \beta_m^2$ represents the OFPV explained by all K causal variables. Assuming a random effect model $\beta_m \sim N(0, \sigma^2)$ *i.i.d.*, then $\lambda = K\sigma^2$. One can show that $\text{cor}(y_i, y_j) = E(y_i y_j) = \lambda g_{ij}$, where

$$g_{ij} = \frac{1}{K} \sum_{m \in C} x_{im} x_{jm}. \quad (1)$$

We define an $N \times N$ similarity matrix G with $(G)_{ij} = g_{ij}$. Then, $(y_1, \dots, y_N)$ follows a multivariate normal distribution with covariance matrix $\lambda G + (1 - \lambda)I$.

Two types of methods have been proposed to estimate λ . The first method is the restricted maximum likelihood (REML) method. For GWAS, this method was implemented in a software package Genome-wide Complex Trait Analysis (GCTA)⁵ and has been used extensively. The second method regresses the empirical Phenotypic Correlations $y_i y_j$ onto the Genetic Correlation (PCGC)³⁰ $y_i y_j \sim \lambda g_{ij}$ for all $N(N - 1)/2$ pairs of subjects to obtain

$$\hat{\lambda}_g = \frac{\sum_{i < j} y_i y_j g_{ij}}{\sum_{i < j} g_{ij}^2}, \quad (2)$$

which is also called the Haseman-Elston regression in statistical genetics literature^{22,23}. Existing studies showed that both methods are consistent^{21,30}. We call $\hat{\lambda}_g$ as the PCGC estimator. Note that the REML method is based on the full likelihood and provides a more efficient estimate than the PCGC method that uses only pairwise information. However, technical analysis of measurement error is easier for the PCGC estimator; thus, our investigation will be based on the PCGC estimator.

2.2 Estimating OFPV when the causal set C is known

We extend the PCGC method to the scenario when genomic variables are measured with error. For the genomic variable m and subject i , let x_{im} be the underlying unobserved true value, e_{im} be the measurement error and $x'_{im} = x_{im} + e_{im}$ be the observed value of the error-prone variable. We make three assumptions: (1) e_{im} and x_{im} are independent, (2) e_{im} and e_{in} are independent for any (m, n) , and (3) $e_{im} \sim N(0, \theta_m^2)$. Define $\tau_m = (1 + \theta_m^2)^{-1}$, the ICC of the variable. Again, we normalize the observed value to have mean zero and unit variance

$$x_{im}^* = \frac{x_{im} + e_{im}}{\sqrt{1 + \theta_m^2}}. \quad (3)$$

The error-prone version of g_{ij} in (3) is given by

$$g_{ij}^* = \frac{1}{K} \sum_{m \in C} x_{im}^* x_{jm}^*. \quad (4)$$

Let $\hat{\lambda}_g^*$ be the PCGC estimator based on g_{ij}^* using formula (4). We will derive AAF for $\hat{\lambda}_g^*$.

Theorem 1. *We assume that x_{im}, e_{im} ($i = 1, \dots, N; m = 1, \dots, M$) are independent random variables with $x_{im} \sim N(0, 1)$ and $e_{im} \sim N(0, \theta_m^2)$. Define $\bar{\tau}_c = \sum_{m \in C} \tau_m / K$ be the average ICC for the causal genomic variables. Let $\hat{\lambda}_g^*$ be the PCGC estimator based on error-prone x_{im}^* defined in (3). As $N, K \rightarrow \infty$ in a way that $K/N^2 \rightarrow 0$, $\hat{\lambda}_g^* / \bar{\tau}_c \xrightarrow{P} \lambda$.*

Thus, $\hat{\lambda}_g^*$ using error-prone predictors underestimates λ with $\text{AAF} = \bar{\tau}_c$. If we know $\bar{\tau}_c$, $\hat{\lambda}_g^* / \bar{\tau}_c$ provides a consistent estimate of λ . As a special case, when all causal genomic variables have the same ICC τ_0 , $\hat{\lambda}_g^* / \tau_0$ provides a consistent estimate of λ . Of note, the convergence requires $K/N^2 \rightarrow 0$ as $N, K \rightarrow \infty$, i.e., the number of causal genomic variables does not increase too fast compared to the sample size N .

2.3 Genomic variables are measured with error and the causal set C is unknown

In reality, the set of causal genomic variables C is usually unknown. In GWAS, g_{ij} in (1) was replaced by h_{ij} that is defined based on all M genetic variants:

$$h_{ij} = \frac{1}{M} \sum_{m=1}^M x_{im} x_{jm}. \quad (5)$$

Obviously, h_{ij} is a noised version of g_{ij} by including $M - K$ non-causal genomic variables.

Define a $N \times N$ similarity matrix H with $(H)_{ij} = h_{ij}$. We can derive an estimate based on either REML or PCGC using H . Simulations for the REML estimator using GWAS data in early years suggested that these estimates based on H provided a practically unbiased estimate for λ , which was justified by a theoretical analysis under appropriate assumptions²¹. In what follows, let $\hat{\lambda}_h$ be

the PCGC estimator using h_{ij} . In Theorem 2, we will show that $\hat{\lambda}_h \rightarrow \lambda$ in probability as $M, N \rightarrow \infty$ in a way that $M/N^2 \rightarrow 0$. This provides another proof of the work in Jiang et al. (2016)²¹, while their work was based on REML and technically more complicated.

Theorem 2. Let $\hat{\lambda}_h$ be the PCGC estimator based on h_{ij} defined in (5). When $M, N \rightarrow \infty$ in a way that $M/N^2 \rightarrow 0$, $\hat{\lambda}_h \xrightarrow{P} \lambda$.

Now, we consider the scenario when genomic markers are measured with error and the causal set is unknown. Let h_{ij}^* be calculated based on all M genomic variables that are measured with error and let $\hat{\lambda}_h^*$ be the PCGC estimator using h_{ij}^* . In Theorem 3, we show that the AAF of h_{ij}^* is equal to the average ICC of all genomic variables under appropriate assumptions.

Theorem 3. We assume that τ_m follows the same distribution for causal and non-causal genomic variables. Define $\bar{\tau} = \sum_{m=1}^M \tau_m / M$ to be the average ICC for all M genomic variables. Then, $\hat{\lambda}_h^* / \bar{\tau} \xrightarrow{P} \lambda$ when $M, N \rightarrow \infty$ in a way that $M/N^2 \rightarrow 0$.

3. Estimating AAF using repeated measurement experiment

When the ICC values τ_m are known for all genomic variables, $\hat{\lambda}_h^* / \bar{\tau}$ consistently estimates λ under appropriate conditions and $Var(\hat{\lambda}_h^* / \bar{\tau}) = Var(\hat{\lambda}_h^*) / \bar{\tau}^2$. When ICCs are unknown, we need to perform a repeated measurement experiment to derive $\hat{\theta}_m^2$, the estimate of the measurement error variance θ_m^2 . Let $\hat{\Theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m^2$ be the average across M genomic variables. Since $\tau_m = 1 / (1 + \theta_m^2)$, we have the first order Taylor's expansion:

$$\frac{\hat{\lambda}_h^*}{\bar{\tau}} = \frac{\hat{\lambda}_h^*}{\frac{1}{M} \sum_{m=1}^M \frac{1}{1+\hat{\theta}_m^2}} \approx \hat{\lambda}_h^* \left(1 + \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m^2 \right) = \hat{\lambda}_h^* (1 + \hat{\Theta}). \quad (6)$$

For two independent random variables z_k ($k = 1, 2$) with $\mu_k = E z_k$ and $\sigma_k^2 = Var(z_k)$,

$Var(z_1 z_2) = \sigma_1^2 \sigma_2^2 + \sigma_1^2 \mu_2^2 + \mu_1^2 \sigma_2^2$. Following this, we have

$$Var(\hat{\lambda}_h^* (1 + \hat{\Theta})) = Var(\hat{\lambda}_h^*) Var(\hat{\Theta}) + Var(\hat{\lambda}_h^*) (1 + E\hat{\Theta})^2 + (E\hat{\lambda}_h^*)^2 Var(\hat{\Theta}).$$

This leads to the following approximation:

$$Var\left(\frac{\hat{\lambda}_h^*}{\bar{\tau}}\right) \approx \left(Var(\hat{\lambda}_h^*) + (\hat{\lambda}_h^*)^2\right) Var(\hat{\Theta}) + Var(\hat{\lambda}_h^*) (1 + \hat{\Theta})^2 \quad (7)$$

if we replace $E\hat{\lambda}_h^*$ with $\hat{\lambda}_h^*$ and replace $E\hat{\Theta}$ with $\hat{\Theta}$. The first term $\left(Var(\hat{\lambda}_h^*) + (\hat{\lambda}_h^*)^2\right) Var(\hat{\Theta})$ in (7) is the extra variability caused by the uncertainty of the estimated ICC values for M genomic variables. Of note, $Var(\hat{\Theta}) = \frac{1}{M^2} \sum_{m=1}^M Var(\hat{\theta}_m^2)$ reduces with M . Typically, $Var(\hat{\theta}_m^2)$ may be quite large for an individual genomic variable due to the limited sample size in the repeated measurement experiment; however, $Var(\hat{\Theta})$ may be small when M is large.

4. Estimating the correlation of effect sizes between two traits

We consider two studies investigating different traits with N_1 and N_2 subjects, respectively. All subjects are profiled for M genomic markers. For subject i in study k ($k = 1, 2$), let y_{ki} be the trait value and x_{kim} be the uncontaminated genomic value. All variables are normalized to have mean zero and unit variance. We assume that the same set of K genomic markers, denoted as C , are causal for the two traits under the following additive models:

$$y_{1i} = \sum_{m \in C} \beta_{1m} x_{1im} + \varepsilon_{1i} \quad \text{and} \quad y_{2j} = \sum_{m \in C} \beta_{2m} x_{2jm} + \varepsilon_{2j}.$$

Furthermore, we assume $\beta_{1m} \sim N(0, \sigma_1^2)$ i.i.d., $\beta_{2m} \sim N(0, \sigma_2^2)$ i.i.d. and $\text{cor}(\beta_{1m}, \beta_{2m}) = \rho$ for $m \in C$. For each trait, $\lambda_k = K\sigma_k^2$ can be estimated using methods in the previous section. Our interest in this section centers on estimating ρ when genomic variables are measured with error.

Let $g_{ij}^{12} = \frac{1}{K} \sum_{m \in C} x_{1im} x_{2jm}$ for two subjects i and j from two studies. We have $E y_{1i} y_{2j} = \sigma_{12} g_{ij}^{12}$, where $\sigma_{12} = \rho \sqrt{\lambda_1 \lambda_2}$. Similarly, we consider regression $y_{1i} y_{2j} \sim \sigma_{12} g_{ij}^{12}$ to derive the PCGC estimator

$$\hat{\sigma}_{12} = \frac{\sum_{i,j} y_{1i} y_{2j} g_{ij}^{12}}{\sum_{i,j} (g_{ij}^{12})^2}. \quad (8)$$

Now, we assume that genomic variables are measured with error, e.g., x_{kim} is contaminated as $x'_{kim} = x_{kim} + e_{kim}$ with $e_{kim} \sim N(0, \theta_{km}^2)$. Similarly, we define $x_{kim}^* = x'_{kim} / \sqrt{1 + \theta_{km}^2}$ and $(h_{ij}^{12})^* = \frac{1}{M} \sum_m x_{1im}^* x_{2jm}^*$ using all M genomic variables, and derive the PCGC estimator

$$\hat{\sigma}_{12}^* = \frac{\sum_{i,j} y_{1i} y_{2j} (h_{ij}^{12})^*}{\sum_{i,j} ((h_{ij}^{12})^*)^2}. \quad (9)$$

We define $\tau_{km} = \frac{1}{1 + \theta_{km}^2}$ to be the ICC for the m^{th} genomic marker in study k and denote $\bar{\tau}_k$ as

the average ICC across all M genomic variables. Let $\hat{\lambda}_{kh}^*$ be the PCGC estimator based on all M contaminated genomic variables for study k . Based on Theorem 3, we have $\hat{\lambda}_{kh}^* / \bar{\tau}_k \xrightarrow{P} \lambda_k$. Then,

ρ can be estimated as $\hat{\rho}^* = \hat{\sigma}_{12}^* / \sqrt{\hat{\lambda}_{1h}^* \hat{\lambda}_{2h}^*}$.

Theorem 4. For both studies, we assume that the ICC values τ_{km} are equally distributed

between causal and non-causal genomic markers. As $N_1, N_2, M \rightarrow \infty$ in a way that $\frac{M}{N_1 N_2} \rightarrow 0$,

$$\frac{\hat{\sigma}_{12}^*}{\frac{1}{M} \sum_{m=1}^M \sqrt{\tau_{1m} \tau_{2m}}} \xrightarrow{P} \sigma_{12} \text{ and } \hat{\rho}^* \frac{\sqrt{\frac{1}{M} \sum_{m=1}^M \tau_{1m} \frac{1}{M} \sum_{m=1}^M \tau_{2m}}}{\frac{1}{M} \sum_{i=1}^M \sqrt{\tau_{1m} \tau_{2m}}} \xrightarrow{P} \rho.$$

Theorem 4 suggests that the AAF for $\hat{\rho}^*$ is $\frac{E\sqrt{\tau_{1m}\tau_{2m}}}{\sqrt{E\tau_{1m}E\tau_{2m}}}$. By the Cauchy-Schwarz inequality, AAF ≤ 1 and “=” holds iff $\frac{\tau_{1m}}{\tau_{2m}}$ is a constant. In particular, if $\tau_{1m} = \tau_{2m}$ for all markers (a reasonable assumption), AAF=1, i.e., $\hat{\rho}^*$ is consistent even when genomic markers are measured with error.

5. Simulation studies

In the first set of simulations, we assumed the same measurement error variance ($\theta^2 = 1$ or 0.25, ICC=0.5 or 0.8) across $M = 10,000$ independent variables $x_{im} \sim N(0,1)$. The fraction of causal variables was set as 100%, 20%, 10% and 5%. Sample size was set at 5000. For each setting, we performed 500 simulations. The results are summarized in Table 1. The naïve estimate without measurement error correction underestimated λ with the ICC as the attenuation factor. Our methods provided unbiased estimates after correction.

In the second set of simulations, we allowed the M independent variables to have different measurement error variances $e_m \sim N(0, \theta_m^2)$, where θ_m were generated from a $Beta(a, b)$ distribution. We chose $(a, b) = (3, 3)$ and $(5, 2)$ to include symmetric and left skewed distributions, respectively. The average ICC across M predictors was 0.793 and 0.664 for the two sets of (a, b) parameters, respectively. We repeated the different set-ups where different fractions of predictors were causal. Again, sample size was 5000 and $M = 10,000$. These results

are summarized in Table 2. Consistent with our theoretical analysis, the attenuation factor was the average ICC across M predictors. After correction, we obtained unbiased estimates of λ for all simulations.

In the third set of simulations, we investigated the impact of correlation among predictors. Parameters were the same as the second set of simulations; however, the predictors were simulated to be dependent with correlation coefficient 0.2 or 0.5 for all pairs of predictors. The results are summarized in Table 3. Modest correlation $r = 0.2$ did not impact the estimate of λ and we obtained unbiased estimates of λ after correction. For the scenario with extensive and strong correlations ($r = 0.5$), the original statistical framework overestimated λ . This rarely happens in reality because only few genomic variables are highly correlated and highly correlated variables are typically excluded from analysis. Interestingly, the attenuation factor was still the average of ICC (~ 0.667 in current simulations) regardless of the correlation among predictors.

6. Real data analysis

The human gut microbiome, also called the second human genome, is the collection of microbes inhabiting the human gut and has been shown to be implicated in many complex diseases in cross-sectional and prospective studies³¹. Estimating the OFPV of human microbiome to the risk of developing a specific disease in prospective studies can help us understand prognostic potential of human microbiome and design future studies appropriately. However, very few prospective microbiome studies are publicly available, and for the ones whose data are available, the sample size is too small to estimate OFPV.

To illustrate the idea of measurement error correction in estimating OFPV, we have used the human gut microbiome data from the American Gut Project (AGP)³² to analyze body mass index

(BMI) and multiple allergy traits (Table 4). Note that AGP is a cross-sectional study, the estimated λ cannot be interpreted as the causal effect of gut microbiome to these traits. Even if the human gut microbiome explains a large fraction of phenotypic variance of, for example, peanut allergy, this may be because individuals with peanut allergy avoid peanut-related food that impacts the composition of gut microbiome communities. Causality needs to be established based on large scale prospective studies. While the causality cannot be established, this type of analyses is important and informative for microbiome studies³³.

The AGP data, including the operational taxonomy units (OTU) table and clinical variables, were downloaded from the Github website (<https://github.com/biocore/American-Gut/tree/master/data/AG>). We calculated the relative abundance (RA) table based on the OTU table. In total, 2257 subjects were available for analysis after filtering (sample exclusion: age missing or ≤ 3 , history of using antibiotics in the past month, patients with type 2 diabetes or inflammatory bowel disease, or the number of reads < 1000 after quality control).

We estimated measurement error variances and ICCs for taxa using data with the repeated measurement data separated by approximately six months from the Human Microbiome Project³⁴ (HMP). For 887 taxa present in at least 5% of HMP subjects and at least 1% in AGP subjects, we fitted a random effect model using an R package *lme4*³⁵ to estimate measurement error variances and calculated ICCs. The average ICC for those taxa was found to be 0.27.

We performed two sets of analyses. The first set included 887 OTUs present in at least 1% of AGP subjects while the second set included 507 OTUs present in at least 5% of HMP subjects (average ICC=27%). Results are summarized in Table 4. Overall, human gut microbiome was significantly associated with BMI and all allergy traits except for dander and poison ivy allergy, consistent with our previous beta-diversity analyses³⁶. ICCs were relatively low in general for

microbiome primarily due to temporal instability; ignoring measurement error may severely underestimate the overall contribution of microbiome to the trait. For example, correcting for measurement error increased the estimate for seasonally allergy from 11.6% to 43.1% using 887 taxa, suggesting that a very large fraction of seasonally allergy is related to gut microbiome. **In our analyses, gut microbiome explained 12.3% of BMI variance without correction for measurement error and 45.5% after correction. A recent paper estimated that 25% of BMI variance was attributed to gut microbiome³³ without considering correcting for measurement error. This uncorrected estimate was higher than our estimate possibly because of different covariates adjusted in the model.**

Finally, we point out the estimates would be higher when we include more predictors at the price of increased standard error. For example, for seasonal allergy, the measurement error-corrected estimate $\hat{\lambda} = 32\%$ ($s.e. = 4.4\%$) using 507 OTUs present in at least 5% subjects and $\hat{\lambda} = 43.1\%$ ($s.e. = 6.1\%$) using 887 OTUs present in at least 1% subjects.

7. Discussion

Prospective studies examining the association between high-dimensional genomic variables and different diseases are essential for understanding disease etiology and for building risk prediction models. Estimating the OFPV of a disease/trait attributed to all genomic variables is a crucial step to understand the upper limit of the potential of genomic variables for predicting a disease and design future projects. We have investigated the impact of measurement error on estimating

OFPV. Our analysis shows that AAF equals to the average ICC across all causal genomic variables. Under the assumption ICCs are distributed similarly between causal and non-causal variables, AAF equals to the average ICC of all genomic variables.

When ICCs are known in advance, we can consistently estimate OFPV and adjust the variance of the estimator. When ICCs are unknown, we have to estimate ICCs using pilot studies with repeated measurements. The sample size of a pilot study is often limited; thus, the uncertainty of the estimated measurement error variances is large for individual variables. However, since AAF is the average ICC across all genomic markers, the extra variance due to the uncertainty of ICC estimate is small if the number of genomic variables is large.

In GWAS, measurement error comes from both genotyping and imputation. For common variants, since only well-genotyped and well-imputed variants are used, measurement error is expected to have a relatively small impact on heritability analysis. As large-scale sequencing studies are developed, heritability analyses are performed for rare variants derived from sequencing or imputation. In this case, the impact of measurement error of heritability analysis must be evaluated. For other genomic variables (e.g., microbiome), measurement error would have relatively large impact on estimating OFPV and must be accounted for.

Further, we have derived the AAF for estimating the correlation of effect sizes of M variables on two traits (also called co-heritability in genetics). We have considered two independent sets of subjects profiled for the same set of genomic markers. A similar analysis can be applied to a single set of subjects with both traits and the conclusion remains the same. As we show, when

$\frac{\tau_{1m}}{\tau_{2m}} = \text{constant}$, i.e., the ratios of ICCs between two studies are the same across all variables,

AAF equals to one. In reality, we expect the same ICC for a genomic variable in both studies; thus, we do not need to adjust for measurement error for estimating this parameter.

There are several limitations in the current analysis. First, we assume that genomic variables are independent, which may not hold for most of genomic studies. Our numerical study shows that modest correlation does not seem to bias the estimate and the conclusion of measurement error correction is still valid. Thus, for real data analysis, we would suggest performing correlation-based pruning so that no extensively highly correlated genomic variables are included for the analysis. Future theoretical analysis is required for any correlation structure among genomic variables. Second, our analysis was based on the PCGC estimator, a type of moment estimator. This estimator is less efficient compared to the REML estimator; it would be interesting to extend current analysis to the REML framework. Third, the current analysis is for a quantitative trait, and it will be interesting to extend it to survival traits in cancer genomics and case-control studies by accounting for ascertainment^{7,11}. Moreover, we did not consider the measurement error in fixed effects. Finally, we assume that measurement error is additive and independent, which needs to be investigated empirically. Cui and his colleagues¹⁹ assume a general covariance structure for measurement error; however, it is impossible to apply their approach to a study with a large number of genomic variables. Future research is warranted in this direction.

Appendix A. Notations and lemmas

For convenience, we define

$$\hat{\eta}_m = \frac{2}{N(N-1)} \sum_{i < j} y_i y_j x_{im} x_{jm}, \quad (A1)$$

the numerator of the PCGC estimator based on a single genomic variable. One can verify that

$\sigma_N^2 = \text{Var}(\hat{\eta}_m) \propto \frac{2}{N(N-1)}$. As $N \rightarrow +\infty$, $N^2 \sigma_N^2 \rightarrow \pi$, a finite constant. Some algebra leads to

$$\hat{\lambda}_g = \frac{\sum_{m \in C} \hat{\eta}_m}{\frac{2K}{N(N-1)} \sum_{i < j} g_{ij}^2}. \quad (A2)$$

We will show that the denominator $\frac{2K}{N(N-1)} \sum_{i < j} g_{ij}^2$ converges to 1 when g_{ij} is calculated based

on K independent predictors. Thus, $\hat{\lambda}_g \approx \sum_{m \in C} \hat{\eta}_m$ and the representation (A2) greatly simplifies the presentation for our technical analysis.

Lemma 1. Assume random variables (RVs) $X_1, \dots, X_N$ i.i.d. with $EX_i = 0$, $\text{Var} X_i = 1$ and

$EX_i^4 < +\infty$. Then $\lim_{N \rightarrow \infty} \frac{2}{N(N-1)} \sum_{i < j} X_i^2 X_j^2 \xrightarrow{a.s.} 1$.

Proof. Some algebra leads to $\frac{2}{N(N-1)} \sum_{i < j} X_i^2 X_j^2 = \frac{N}{N-1} \left(\frac{1}{N} \sum_{i=1}^N X_i^2 \right)^2 - \frac{1}{N-1} \frac{1}{N} \sum_{i=1}^N X_i^4$. As

$N \rightarrow \infty$, $\frac{1}{N} \sum_{i=1}^N X_i^2 \xrightarrow{a.s.} EX_i^2 = 1$ and $\frac{1}{N} \sum_{i=1}^N X_i^4 \xrightarrow{a.s.} EX_i^4 < +\infty$. This completes the proof.

Lemma 2. Assume RVs X_{im} ($i = 1, \dots, N$; $m = 1, \dots, M$) i.i.d. with $EX_{im} = 0$, $\text{Var} X_{im} = 1$

and $EX_{im}^4 < +\infty$. Then $\lim_{N \rightarrow \infty} \frac{2}{N(N-1)} \sum_{i < j} X_{im} X_{jm} X_{il} X_{jl} \xrightarrow{a.s.} 0$.

Proof. Note that $\frac{2}{N(N-1)} \sum_{i < j} X_{im} X_{jm} X_{il} X_{jl} = \frac{N}{N-1} \left(\frac{1}{N} \sum_{i=1}^N X_{im} X_{il} \right)^2 - \frac{1}{N-1} \left(\frac{1}{N} \sum_{i=1}^N X_{im}^2 X_{il}^2 \right)$. As

$N \rightarrow \infty$, this converges to $(E X_{im} X_{il})^2 - E X_{im}^2 X_{il}^2 = 0$ because of independence.

Lemma 3. Assume RVs X_{im} ($i = 1, \dots, N; m = 1, \dots, M$) i.i.d. with $E X_{im} = 0$ and $Var X_{im} = 1$ and $E X_{im}^4 < +\infty$. Let $g_{ij} = \sum_{m=1}^M X_{im} X_{jm} / M$. Then,

$$\lim_{N \rightarrow \infty} \frac{2M}{N(N-1)} \sum_{i < j} g_{ij}^2 \xrightarrow{a.s.} 1.$$

Proof. One can verify that the above equation equals to

$$\frac{1}{M} \sum_{m=1}^M \left(\frac{2}{N(N-1)} \sum_{i < j} X_{im}^2 X_{jm}^2 \right) + \frac{1}{M} \sum_{m \neq l} \left(\frac{2}{N(N-1)} \sum_{i < j} X_{im} X_{jm} X_{il} X_{jl} \right).$$

As $N \rightarrow \infty$, the first term $\xrightarrow{a.s.} 1$ by Lemma 1 and the second term $\xrightarrow{a.s.} 0$ by Lemma 2. This completes the proof.

Lemma 4. Suppose independent RVs $V_{N,k} \sim N(0, \sigma_N^2)$ with $\lim_{N \rightarrow \infty} N^2 \sigma_N^2 = \pi < \infty$. Let

$S_{N,K} = \sum_{k=1}^K V_{N,k}$. Then, $S_{N,K} \xrightarrow{P} 0$ as when $N, K \rightarrow \infty$ in a way that $K/N^2 \rightarrow 0$.

Proof. When $N, K \rightarrow \infty$ in a way that $K/N^2 \rightarrow 0$, $Var(S_{N,K}) = Var(\sum_{k=1}^K V_{N,k}) = K \sigma_N^2 =$

$\frac{K}{N^2} N^2 \sigma_N^2 \rightarrow 0$ because $\lim_{N \rightarrow \infty} N^2 \sigma_N^2 = \pi < \infty$. This implies $S_{N,K} \xrightarrow{L_2} 0$ and thus $S_{N,K} \xrightarrow{P} 0$.

Lemma 5. Let $\lambda > 0$ be a fixed constant. Given K , let $V_1, \dots, V_K$ be independent RVs with

$V_i \sim N(0, \sqrt{\lambda/K})$. Given V_i , define W_i to be a RV with $W_i \sim N(V_i^2, \sigma_N^2)$ and $\lim_{N \rightarrow \infty} N^2 \sigma_N^2 = \pi <$

∞ . Let $U_1, \dots, U_K$ be i.i.d. RVs with $U_i \sim N(u, \sigma_U^2)$ and $\bar{U}_K = \sum_{i=1}^K U_i / K$. When $N, K \rightarrow \infty$ in a

way that $K/N^2 \rightarrow 0$, $\sum_{i=1}^K (U_i - \bar{U}_K) W_i \xrightarrow{P} 0$.

Proof. Let $V'_i = \sqrt{K/\lambda}V_i$, then $V'_i \sim N(0, 1)$. Since $W_i \sim N(V_i^2, \sigma_N^2)$, we assume $W_i = V_i^2 + E_i =$

$\frac{\lambda}{K} V_i'^2 + E_i$, where $E_i \sim N(0, \sigma_N^2)$. Thus,

$$\sum_{i=1}^K (U_i - \bar{U}_K) W_i = \lambda \underbrace{\frac{\sum_{i=1}^K U_i V_i'^2}{K}}_{A_1} + \underbrace{\sum_{i=1}^K U_i E_i}_{A_2} - \lambda \bar{U}_K \underbrace{\frac{\sum_{i=1}^K V_i'^2}{K}}_{A_3} - \bar{U}_K \underbrace{\sum_{i=1}^K E_i}_{A_4}.$$

By the law of large numbers, as $K \rightarrow \infty$, $A_1 \xrightarrow{a.s.} E(U_i V_i'^2) = u$ and $A_3 \xrightarrow{a.s.} \lambda E U_i E V_i'^2 = u$; thus,

$A_1 - A_3 \xrightarrow{a.s.} 0$. Moreover, because $\lim_{N \rightarrow \infty} N^2 \sigma_N^2 = \pi < \infty$ and $K/N^2 \rightarrow 0$, $\sum_{i=1}^K E_i \xrightarrow{P.} 0$ (i.e.,

$A_4 \xrightarrow{P.} 0$) and $A_2 \rightarrow 0$ by Lemma 4. This completes the proof.

Appendix B. Proof of Theorem 1.

Proof. By definition,

$$\hat{\lambda}_g^* = \frac{\frac{2}{N(N-1)} \sum_{i,j} y_i y_j g_{ij}^*}{\frac{2}{N(N-1)} \sum_{i,j} [g_{ij}^*]^2}, \quad (B1)$$

where g_{ij}^* defined in (4) is based on error-prone predictors. Combining (3) and (4), we have

$$g_{ij}^* = \frac{1}{K} \sum_{m \in C} [\tau_m(x_{im} + e_{im})(x_{jm} + e_{jm})] = \frac{1}{K} \sum_{m \in C} \tau_m(x_{im}x_{jm} + \xi_{ijm}), \quad (B2)$$

where $\xi_{ijm} = x_{im}e_{jm} + e_{im}x_{jm} + e_{im}e_{jm}$. Plugging (B2) into (2) and noticing $\hat{\eta}_m$ defined in

(A1), we found that $\hat{\lambda}_g^*$ can be rewritten as:

$$\hat{\lambda}_g^* = \underbrace{\frac{\sum_{m \in C} \tau_m \hat{\eta}_m}{\frac{2K}{N(N-1)} \sum_{i,j} g_{ij}^2}}_{A_1} \cdot \underbrace{\frac{\frac{2K}{N(N-1)} \sum_{i,j} g_{ij}^2}{\frac{2K}{N(N-1)} \sum_{i,j} [g_{ij}^*]^2}}_{A_2} + \underbrace{\frac{\sum_{m \in C} \left(\frac{2}{N(N-1)} \sum_{i,j} y_i y_j \xi_{ijm} \right)}{\frac{2K}{N(N-1)} \sum_{i,j} [g_{ij}^*]^2}}_{A_3}. \quad (B3)$$

By Lemma 3, the numerator and denominator of A_2 converge to 1 and thus $A_2 \rightarrow 1$ as $N \rightarrow \infty$.

We next show that $A_3 \xrightarrow{P} 0$. Since the denominator of $A_3 \xrightarrow{a.s.} 1$ by Lemma 3, it remains to prove

that the numerator converges to zero in probability. It is easy to verify that $cov(y_i y_j, \xi_{ijm}) = 0$,

thus $\hat{R}_m := \frac{2}{N(N-1)} \sum_{i < j} y_i y_j \xi_{ijm}$ is the PCGC estimator for zero using data ξ_{ijm} ; thus,

$\hat{R}_m \sim N(0, \sigma_N^2)$ asymptotically with $\lim_{N \rightarrow \infty} N^2 \sigma_N^2 = \pi < \infty$. Thus, the numerator of A_3 becomes

$\sum_{m \in \mathcal{C}} \hat{R}_m$ and converges to zero in probability by Lemma 4.

Finally, we prove that $\lim_{K, N \rightarrow \infty} A_1 / \bar{\tau}_g = \lambda$. Note that A_1 can be rewritten as

$$A_1 = \frac{\sum_{m \in \mathcal{C}} (\tau_m - \bar{\tau}_g) \hat{\eta}_m}{\frac{2K}{N(N-1)} \sum_{i < j} g_{ij}^2} + \bar{\tau}_g \frac{\sum_{m \in \mathcal{C}} \hat{\eta}_m}{\frac{2K}{N(N-1)} \sum_{i < j} g_{ij}^2} = \frac{\sum_{m \in \mathcal{C}} (\tau_m - \bar{\tau}_g) \hat{\eta}_m}{\frac{2K}{N(N-1)} \sum_{i < j} g_{ij}^2} + \bar{\tau}_g \hat{\lambda}_g.$$

The second equation holds because of (A2). Note that $\hat{\eta}_m$ is the PCGC estimator for β_m^2 . This

suggests that $\hat{\eta}_m$ can be written as $\hat{\eta}_m = \beta_m^2 + E_m$, where $\beta_m \sim N(0, \lambda/K)$ is the effect size for

the variable in the random effect model and $E_m \sim N(0, \sigma_N^2)$. By Lemma 5, $\sum_{m \in \mathcal{C}} (\tau_m - \bar{\tau}_g) \hat{\eta}_m \rightarrow 0$

(and thus the first term of $A_1 \rightarrow 0$) as $N, K \rightarrow \infty$ in a way that $K/N^2 \rightarrow 0$. Since it is established

that $\lim_{N \rightarrow \infty} \hat{\lambda}_g = \lambda$, $\lim_{N \rightarrow \infty} A_1 / \bar{\tau}_g = \lambda$. This completes the proof.

Appendix C. Proof Theorem 2.

Proof. By definition,

$$\hat{\lambda}_h = \frac{\sum_{m=1}^M \hat{\eta}_m}{\frac{2M}{N(N-1)} \sum_{i < j} h_{ij}^2} = \frac{\sum_{m \in \mathcal{C}} \hat{\eta}_m + \sum_{m \notin \mathcal{C}} \hat{\eta}_m}{\frac{2M}{N(N-1)} \sum_{i < j} h_{ij}^2},$$

which can be further written as

$$\hat{\lambda}_h = \underbrace{\frac{\sum_{m \in C} \hat{\eta}_m}{\frac{2K}{N(N-1)} \sum_{i < j} g_{ij}^2}}_{A_1} + \underbrace{\frac{\sum_{m \notin C} \hat{\eta}_m}{\frac{2M}{N(N-1)} \sum_{i < j} h_{ij}^2}}_{A_2} + \underbrace{\frac{\sum_{m \notin C} \hat{\eta}_m}{\frac{2M}{N(N-1)} \sum_{i < j} h_{ij}^2}}_{A_3}.$$

It is established that $A_1 \rightarrow \hat{\lambda}_g$. Also, $A_2 \rightarrow 1$ because both the numerator and the denominator converge to 1 as $N \rightarrow +\infty$. Thus, $A_1 A_2 \rightarrow \lambda$ as $N \rightarrow +\infty$. Moreover, by Lemma 4, we have $\sum_{m \notin C} \hat{\eta}_m \xrightarrow{P} 0$ (as thus $A_3 \xrightarrow{P} 0$) as $M, N \rightarrow +\infty$ in a way that $M/N^2 \rightarrow 0$. This completes the proof.

Appendix D. Proof of Theorem 3.

Proof. Let $\tau_0 = E\tau_m$. Note that,

$$h_{ij}^* = \frac{1}{K} \sum_{m=1}^M \tau_m (x_{im} + e_{im})(x_{jm} + e_{jm}) = \frac{1}{K} \sum_{m=1}^M \tau_m (x_{im} x_{jm} + \xi_{ijm}) \quad (C1)$$

where $\xi_{ijm} = x_{im} e_{jm} + e_{im} x_{jm} + e_{im} e_{jm}$. Then, $\hat{\lambda}_h^*$ can be rewritten as:

$$\hat{\lambda}_h^* = \underbrace{\frac{\sum_{m \in C} \tau_m \hat{\eta}_m}{\frac{2K}{N(N-1)} \sum_{i < j} [h_{ij}^*]^2}}_{A_1} + \underbrace{\frac{\sum_{m \notin C} \tau_m \hat{\eta}_m}{\frac{2M}{N(N-1)} \sum_{i < j} [h_{ij}^*]^2}}_{A_2} + \underbrace{\frac{\sum_{m=1}^M \frac{2}{N(N-1)} \sum_{i < j} y_i y_j \xi_{ijm}}{\frac{2M}{N(N-1)} \sum_{i < j} [h_{ij}^*]^2}}_{A_3}.$$

Note that $\sum_{m \in C} \tau_m \hat{\eta}_m = \sum_{m \in C} (\tau_m - \tau_0) \hat{\eta}_m + \sum_{m \in C} \tau_0 \hat{\eta}_m$. Using a similar argument as in

Theorem 1, we have $A_1 \rightarrow \tau_0 \lambda$. Moreover, because $0 \leq \tau_m \leq 1$, we have $A_2 \xrightarrow{P} 0$ based on

Lemma 4. Finally, using the same argument for A_3 in Theorem 1, $A_3 \xrightarrow{P} 0$ when $M, N \rightarrow +\infty$ in a way that $M/N^2 \rightarrow 0$. This completes the proof.

Appendix E. Proof of Theorem 4.

Proof. Using a similar argument as in Lemma 3, we can prove that the denominator of $\hat{\sigma}_{12}^*$,

$\frac{M}{N_1 N_2} \sum_{i,j} ((h_{ij}^{12})^*)^2 \rightarrow 0$ as $N_1, N_2 \rightarrow +\infty$. So, we ignore the denominator for simplicity.

Let $\hat{\delta}_m = \frac{1}{N_1 N_2} \sum_{i,j} y_{1i} y_{2j} x_{1im} x_{2jm}$. Let $\gamma_m = \sqrt{\tau_{1m} \tau_{2m}}$ and $\gamma_0 = E\gamma_m$. Then,

$$\hat{\sigma}_{12}^* = \underbrace{\sum_{m \in C} \gamma_m \hat{\delta}_m}_{A_1} + \underbrace{\sum_{m \notin C} \gamma_m \hat{\delta}_m}_{A_2} + \underbrace{\sum_{m=1}^M \frac{1}{N_1 N_2} \sum_{i \leq N_1, j \leq N_2} y_{1i} y_{2j} \xi_{ijm}}_{A_3}$$

Note that $A_1 = \sum_{m \in C} \gamma_m \hat{\delta}_m = \sum_{m \in C} (\gamma_m - \gamma_0) \hat{\delta}_m + \sum_{m \in C} \gamma_0 \hat{\delta}_m$. Using a similar argument as in

Theorem 1, we have $\sum_{m \in C} (\gamma_m - \gamma_0) \hat{\delta}_m \xrightarrow{P} 0$ and thus $A_1 \xrightarrow{P} \gamma_0 \sigma_{12}$ as $N_1, N_2, M \rightarrow \infty$ in a way that

$M/N_1 N_2 \rightarrow 0$. Moreover, because $0 \leq \gamma_m \leq 1$, we have $A_2 \xrightarrow{P} 0$ based on Lemma 4. Finally,

using the same argument for A_3 in Theorem 1, $A_3 \xrightarrow{P} 0$ as $N_1, N_2, M \rightarrow \infty$ in a way that $M/$

$N_1 N_2 \rightarrow 0$. This completes the proof.

Acknowledgements

This work utilized the computational resources of the NIH HPC Biowulf cluster. (<http://hpc.nih.gov>). The authors are supported by the NIH Intramural Research Program.

Competing interests

The authors declare no competing interests.

References

1. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467-484 (2019).
2. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-76 (2010).

3. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-9 (2005).
4. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).
5. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
6. Allen, H.L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838 (2010).
7. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
8. Bodelon, C. *et al.* Blood DNA methylation and breast cancer risk: a meta-analysis of four prospective cohort studies. *Breast Cancer Res* **21**, 62 (2019).
9. Hosgood, H.D. *et al.* Variation in oral microbiome is associated with future risk of lung cancer among never-smokers. *Thorax* **76**, 256-263 (2021).
10. Vogtmann E, H.X., Yu G, Purandare V, Hullings AG, Shao D, Wan Y, Li S, Dagnall CL, Jones K, Hicks BD, Hutchinson A, Caporaso JG, Wheeler W, Sandler DP, Beane Freeman LE, Liao LM, Huang WY, Freedman ND, Caporaso N, Sinha R, Gail MH, Shi J, Abnet CC. . The human oral microbiome and risk of lung cancer: An analysis of three prospective cohort studies. . *JAMA Oncology* **2022;ONC21-3241**. (2022).
11. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
12. Fuller. Measurement Error Models. *New York: Wiley* (1987).
13. Zhong XP, F.W., Wei BC. Estimation in linear models with random effects and errors-in-variables. *Ann Inst Math Stat* **54**, 595-606 (2002).
14. Wang, N.Y., Lin, X.H., Gutierrez, R.G. & Carroll, R.J. Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association* **93**, 249-261 (1998).
15. Lin, X.H., Carroll, R.J. SIMEX Variance Component Tests in Generalized Linear Mixed Measurement Error Models. *Biometrics*, 613-619 (1998).
16. Carroll, R.J., Kuchenhoff, H., Lombard, F. & Stefanski, L.A. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association* **91**, 242-250 (1996).
17. Carroll, R.J.R., David; Stefanski, Leonard A.; Crainiceanu, Ciprian. Measurement Error in Nonlinear Models: A Modern Perspective. (2006).
18. Zare, K., Rasekh, A. & Rasekhi, A.A. Estimation of variance components in linear mixed measurement error models. *Statistical Papers* **53**, 849-863 (2012).
19. Cui, H.J., Ng, K.W. & Zhu, L.X. Estimation in mixed effects model with errors in variables. *Journal of Multivariate Analysis* **91**, 53-73 (2004).
20. Zhang, Y.D. *et al.* Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat Commun* **11**, 3353 (2020).
21. Jiang, J.M., Li, C., Paul, D., Yang, C. & Zhao, H.Y. On High-Dimensional Misspecified Mixed Model Analysis in Genome-Wide Association Study. *Annals of Statistics* **44**, 2127-2160 (2016).
22. Haseman, J.K. & Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**, 3-19 (1972).
23. Elston, R.C., Buxbaum, S., Jacobs, K.B. & Olson, J.M. Haseman and Elston revisited. *Genet Epidemiol* **19**, 1-17 (2000).

24. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2 (2012).
25. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* **45**, 984-+ (2013).
26. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50**, 229-237 (2018).
27. Hu, Y. *et al.* Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet* **13**, e1006836 (2017).
28. Ting-Huei Chen, N.C., Maria Teresa Landi, Jianxin Shi. A Penalized Regression Framework for Building Polygenic Risk Models Based on Summary Statistics From Genome-Wide Association Studies and Incorporating External Information. *Journal of the American Statistical Association* **116**, 113-143 (2021).
29. Li, C., Yang, C., Gelernter, J. & Zhao, H. Improving genetic risk prediction by leveraging pleiotropy. *Hum Genet* **133**, 639-50 (2014).
30. Golan, D., Lander, E.S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc Natl Acad Sci U S A* **111**, E5272-81 (2014).
31. Turnbaugh, P.J. *et al.* The human microbiome project. *Nature* **449**, 804-10 (2007).
32. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**(2018).
33. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210-215 (2018).
34. Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-14 (2012).
35. Bates D, M.M., Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1-48 (2015).
36. Hua, X., Goedert, J.J., Pu, A., Yu, G.Q. & Shi, J.X. Allergy associations with the adult fecal microbiota: Analysis of the American Gut Project. *Ebiomedicine* **3**, 172-179 (2016).

Table 1. Estimates of OFPV λ of 10,000 independent predictors measured with a fixed measurement error (ME) standard deviation (θ).

Percentage of causal variables	True λ	$\hat{\lambda}$, not corrected for ME		$\hat{\lambda}$, corrected for ME	
		$\theta = 0.5$	$\theta = 1$	$\theta = 0.5$	$\theta = 1$
100%	0.1	0.082 (0.03)	0.051 (0.03)	0.102 (0.04)	0.103 (0.06)
	0.5	0.400 (0.04)	0.250 (0.03)	0.500 (0.05)	0.500 (0.07)
	0.9	0.719 (0.04)	0.449 (0.04)	0.898 (0.06)	0.898 (0.08)
20%	0.1	0.078 (0.03)	0.048 (0.03)	0.097 (0.04)	0.096 (0.06)
	0.5	0.399 (0.04)	0.249 (0.03)	0.499 (0.05)	0.498 (0.07)
	0.9	0.722 (0.05)	0.451 (0.04)	0.902 (0.06)	0.901 (0.08)
10%	0.1	0.079 (0.03)	0.050 (0.03)	0.098 (0.04)	0.099 (0.06)
	0.5	0.397 (0.04)	0.248 (0.04)	0.497 (0.05)	0.496 (0.07)
	0.9	0.717 (0.05)	0.447 (0.04)	0.896 (0.07)	0.894 (0.08)
5%	0.1	0.079 (0.03)	0.049 (0.03)	0.099 (0.04)	0.099 (0.06)
	0.5	0.400 (0.04)	0.250 (0.04)	0.500 (0.06)	0.500 (0.07)
	0.9	0.722 (0.06)	0.452 (0.05)	0.903 (0.08)	0.904 (0.10)

Table 2. Estimates of OFPV λ of 10,000 predictors measured with measurement error (ME). The standard deviation of ME follows a beta distribution $\theta \sim \text{Beta}(3,3)$ or $\theta \sim \text{Beta}(5,2)$.

Percentage of causal variables	True λ	$\hat{\lambda}$, not corrected for ME		$\hat{\lambda}$, corrected for ME	
		$\text{Beta}(3, 3)$	$\text{Beta}(5, 2)$	$\text{Beta}(3, 3)$	$\text{Beta}(5, 2)$
100%	0.1	0.078 (0.03)	0.065 (0.03)	0.098 (0.04)	0.098 (0.05)
	0.5	0.396 (0.04)	0.333 (0.03)	0.499 (0.05)	0.501 (0.05)
	0.9	0.715 (0.04)	0.601 (0.04)	0.900 (0.05)	0.904 (0.06)
20%	0.1	0.081 (0.03)	0.066 (0.03)	0.102 (0.04)	0.099 (0.04)
	0.5	0.398 (0.04)	0.331 (0.03)	0.501 (0.05)	0.497 (0.05)
	0.9	0.715 (0.05)	0.596 (0.04)	0.900 (0.06)	0.897 (0.06)
10%	0.1	0.078 (0.03)	0.065 (0.03)	0.099 (0.04)	0.097 (0.04)
	0.5	0.394 (0.04)	0.33 (0.04)	0.496 (0.05)	0.497 (0.06)
	0.9	0.712 (0.06)	0.597 (0.05)	0.897 (0.07)	0.898 (0.07)
5%	0.1	0.083 (0.03)	0.069 (0.03)	0.105 (0.04)	0.105 (0.04)
	0.5	0.397 (0.04)	0.334 (0.04)	0.500 (0.05)	0.503 (0.06)
	0.9	0.710 (0.06)	0.600 (0.06)	0.894 (0.08)	0.902 (0.08)

Table 3. Estimates of OFPV λ of 10,000 dependent predictors measured with measurement error (ME). The standard deviation of ME follows a beta distribution $\theta \sim \text{Beta}(5,2)$ with average ICC ~ 0.667 . Correlation coefficients between predictors were set at $r = 0.2$ or 0.5 .

Percentage of causal variables	True λ	$\hat{\lambda}$, not corrected for ME		$\hat{\lambda}$, corrected for ME	
		$r = 0.2$	$r = 0.5$	$r = 0.2$	$r = 0.5$
100%	0.1	0.066 (0.029)	0.085 (0.026)	0.100 (0.044)	0.127 (0.039)
	0.5	0.337 (0.034)	0.429 (0.035)	0.506 (0.052)	0.645 (0.053)
20%	0.1	0.066 (0.029)	0.085 (0.028)	0.100 (0.043)	0.127 (0.042)
	0.5	0.334 (0.034)	0.426 (0.044)	0.503 (0.051)	0.641 (0.067)
10%	0.9	0.065 (0.029)	0.084 (0.028)	0.098 (0.044)	0.127 (0.043)
	0.5	0.334 (0.037)	0.425 (0.05)	0.503 (0.051)	0.64 (0.076)
5%	0.1	0.070 (0.031)	0.087 (0.034)	0.105 (0.046)	0.130 (0.051)
	0.5	0.340 (0.062)	0.427 (0.108)	0.512 (0.093)	0.642 (0.163)

Table 4. Estimated phenotypic variance attributed to gut microbiome in American Gut Project.

		# no allergy	# with allergy	Based on 887 OTUs present in > 1% of subjects				Based on 507 OTUs present in > 5%			
				$\hat{\lambda}_{uncorrected}$	P-value	$\hat{\lambda}_{corrected}$	s.e.	$\hat{\lambda}_{uncorrected}$	P-value	$\hat{\lambda}_{corrected}$	s.e.
Allergy traits	Drug	1835	422	18.0%	1.5E-66	66.8%	8.1%	13.8%	7.5E-76	51.0%	6.1%
	Bee sting	2135	122	10.7%	5.4E-25	39.7%	5.7%	5.4%	1.4E-13	20.2%	3.5%
	Dander	1928	329	0.8%	2.3E-01	2.9%	4.0%	0.5%	2.5E-01	1.9%	2.8%
	Nonfood allergy	1601	656	9.0%	4.0E-18	33.4%	5.3%	5.4%	3.1E-13	19.9%	3.5%
	Poison ivy	1841	416	0.0%	5.0E-01	0.0%	3.9%	0.5%	2.3E-01	2.0%	2.8%
	Seasonal allergy	1315	887	11.6%	8.1E-28	43.1%	6.1%	8.6%	9.0E-30	32.0%	4.4%
	Tree nuts	2196	61	2.9%	2.4E-03	10.9%	4.1%	2.2%	1.2E-03	8.3%	2.9%
	Shellfish	2199	58	5.4%	8.6E-08	20.2%	4.5%	4.3%	3.6E-09	15.9%	3.3%
	Pea nuts	2204	53	6.8%	4.3E-11	25.0%	4.7%	5.5%	5.1E-14	20.4%	3.5%
BMI			2257	12.3%	4.5E-32	45.5%	6.2%	8.2%	2.7E-28	30.3%	4.3%