UW Biostatistics Working Paper Series

11-9-2016

Confidence Intervals for Heritability via Haseman-Elston Regression

Tamar Sofer
University of Washington, tsofer@uw.edu

Confidence intervals for heritability via Haseman-Elston regression

Tamar Sofer*

Department of Biostatistics, University of Washington, Seattle, WA, United States of America

*Correspondence to: Tamar Sofer, Department of Biostatistics, University of Washington, UW Tower, 15th Floor, 4333 Brooklyn Ave. NE, Seattle, 98105, USA. E-mail: tsofer@uw.edu. Tel: (206) 543-1490.

*Correspondence: tsofer@uw.edu

Abstract

Heritability is the proportion of phenotypic variance in a population that is attributable to individual genotypes. Heritability is considered an important measure in both evolutionary biology and in medicine, and is routinely estimated and reported in genetic epidemiology studies. In population-based genome-wide association studies (GWAS), mixed models are used to estimate variance components, from which a heritability estimate is obtained. The estimated heritability is the proportion of the model's total variance that is due to the genetic relatedness matrix (kinship measured from genotypes). Current practice is to use bootstrapping, which is slow, or normal asymptotic approximation to estimate the precision of the heritability estimate; however, this approximation fails to hold near the boundaries of the parameter space or when the sample size is small. In this paper we propose to estimate variance components via a Haseman-Elston regression, find the asymptotic distribution of the variance components and proportions of variance, and use them to construct confidence intervals (CIs). Our method is further developed to estimate unbiased variance components and construct CIs by meta-analyzing information from multiple studies. We demonstrate our approach on data from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).

KEY WORDS: Variance components; proportions of variance; quadratic forms; eigenvalues; meta-analysis.

Introduction

Heritability is the proportion of phenotypic variance that is due to genetic variation among individuals in a population. Heritability is often estimated using mixed models (Zaitlen and Kraft, 2012), where the genetic relatedness between any two individuals in a given study population is estimated (e.g. kinship coefficients may be calculated from GWAS data, or inferred from pedigrees) and then taken as fixed. Then, a variance component due to genetic variation is estimated, and the estimated heritability is the ratio between this variance component and the total variance in the model.

Inference about heritability when estimated from mixed models, and more generally, about other proportions of variance, usually relies on asymptotic normal approximation to the distribution of the estimators. However, multiple studies showed (e.g., Burch (2011); Kruijer et al. (2015)) that such confidence intervals are inaccurate, and may yield values that are not permissible (e.g. negative values). Recently, Schweiger et al. (2016) proposed a bootstrap approach for estimating confidence intervals for heritability, and a numerical approximation that does not require bootstrapping under a specific way of calculating the genetic relatedness matrix. While they show that their confidence intervals are accurate, their method is limited by computation time, by requiring a single modeled variance parameter, and by requiring a specific form for the genetic relatedness matrix when using the numerical approximation. In addition, current meta-analysis approaches for heritability estimates rely on the inaccurate normal asymptotic approximation.

In this work we proposes to use Haseman-Elston regression for estimating variance components. This approach entails regressing multiplied residuals against entries of covariance

matrices. We find the asymptotic distribution of the variance component estimators as well as the distributions of the proportions of variance, in a general model that allows for multiple sources of variation. We provide an algorithm to estimate the confidence intervals, and to obtain an unbiased meta-analytic estimator of heritability that accurately combines information from multiple studies. In the case where genetic relatedness (or kinship) is the only sources of variation, our algorithm is very quick, with the computationally demanding step being the calculation of eigenvalues from a sub-matrix of the kinship matrix. We demonstrate our method by estimating heritability of height, depression score, systolic blood pressure, lung function, and dental decay measure in the Hispanics Community Health Study/Study of Latinos.

The mathematical model

Haseman-Elston regression

Suppose that a quantitative trait Y, measured on n individuals, follows the regression model

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + b_{i,a} + \ldots + b_{i,k} + e_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n$$

with $\boldsymbol{\beta}$ a vector of fixed effects of a covariates vector $\boldsymbol{w}$, $b_{i,l}, l = a, \ldots, k, i = 1, \ldots, n$ are mean-zero random effects with $\boldsymbol{b}_l = (b_{i,1}, \ldots, b_{n,l})$ and $\mathrm{cov}(\boldsymbol{b}_l) = \sigma_l^2 \boldsymbol{L}$, so that $\sigma_a^2, \ldots, \sigma_k^2$ are variances corresponding to $a, \ldots, k$ independent sources of variation, and $\mathbf{A}, \ldots, \mathbf{K}$ are $n \times n$ matrices with i, j entries $a_{i,j}, \ldots, k_{i,j}$ modeling the correlation between the individuals' random effects. Also $e_i, i = 1, \ldots, n$ are independent errors with variance σ_e^2. In genetic

association studies one of these matrices, say $\mathbf{K}$, is a kinship, or genetic relatedness, matrix. Then

$$\begin{aligned} E[y_i - \boldsymbol{w}_i\boldsymbol{\beta}] &= E[\boldsymbol{\epsilon}] = \mathbf{0} \\ \text{var}[\boldsymbol{\epsilon}] &= \sigma_e^2\mathbf{I}_{n\times n} + \sigma_a^2\mathbf{A} + \ldots + \sigma_k^2\mathbf{K} = \boldsymbol{\Sigma}, \text{ and} \\ E[\epsilon_i\epsilon_j] &= \text{cov}(\epsilon_i, \epsilon_j) = \sigma_e^2\mathcal{I}_{(i=j)} + \sigma_a^2 a_{i,j} + \ldots + \sigma_k^2 k_{i,j}, \end{aligned}$$

where here $\sigma_k^2/(\sigma_a^2 + \ldots + \sigma_k^2 + \sigma_e^2) \equiv \sigma_k^2/\sigma_T^2$ is the heritability.

Let $\widehat{\boldsymbol{\beta}}$ be an unbiased estimator of $\boldsymbol{\beta}$, and let $\hat{\epsilon}_i = y_i - \boldsymbol{w}_i^T\widehat{\boldsymbol{\beta}}$ be an estimator $\epsilon_i, i = 1, \ldots, n$. We estimate the variance components in a residual regression, i.e. by taking the vector of all unique pairs of residuals $\hat{\epsilon}_i\hat{\epsilon}_j, i \leq j$ (we can do it by taking the upper diagonal sub-matrix of $\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$ that includes the diagonal), denoted by $\tilde{\boldsymbol{\epsilon}}^d$ and regressing it according to the above model. The regression design matrix is given by:

$$\mathbf{X} = \begin{pmatrix} 1 & a_{1,1} & \ldots & k_{1,1} \\ 0 & a_{1,2} & \ldots & k_{1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{1,n} & \ldots & k_{1,n} \\ 1 & a_{2,2} & \ldots & k_{2,2} \\ 0 & a_{2,3} & \ldots & k_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{2,n} & \ldots & k_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{n-1,n-1} & \ldots & k_{n-1,n-1} \\ 0 & a_{n-1,n} & \ldots & k_{n-1,n} \\ 1 & a_{n,n} & \ldots & k_{n,n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & a_{1,2} & \ldots & k_{1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{1,n} & \ldots & k_{1,n} \\ 1 & 1 & 1 & 1 \\ 0 & a_{2,3} & \ldots & k_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{2,n} & \ldots & k_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \\ 0 & a_{n-1,n} & \ldots & k_{n-1,n} \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

where the second equality is because $a_{i,i}, \ldots, k_{i,i} = 1$ for all i. Denote the vector of variance components estimated from the Haseman-Elston regression by $\widehat{\boldsymbol{\sigma}}^2 = (\widehat{\sigma}_e^2, \widehat{\sigma}_a^2, \ldots, \widehat{\sigma}_k^2)^T$.

Properties of the variance components and proportions of variance estimators

Complete mathematical derivations are provided in the appendix. Below are statements of some of the results to provide intuition to the findings and methods.

Lemma 2: *Variance component estimators corresponding to the matrices* $\mathbf{A}, \ldots, \mathbf{K}$ *depend only on the between-observation multiplied residuals of the form* $\hat{\epsilon}_i \hat{\epsilon}_j$ *for* $i \neq j$.

Lemma 3: *Denote by* $\sigma_T^2 = \sigma_e^2 + \sigma_a^2 + \ldots + \sigma_k^2$. *Then* $\widehat{\sigma}_T^2 = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2$.

Theorem: *We say that two matrices* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *are orthogonal in the trace inner product, or "trace orthogonal" if* $tr(\mathbf{C}_1 \mathbf{C}_2) = 0$. *Let the matrix* $\boldsymbol{L}^-$ *be the matrix* $\boldsymbol{L}$ *with all diagonal values set to* 0. *If a matrix* $\boldsymbol{L}^-$ *is trace orthogonal to all other matrices in the set* $\{\mathbf{A}^-, \ldots, \mathbf{K}^-\}$, *then*

$$\widehat{\sigma}_l^2 = \frac{1}{\sum_{j>i} l_{i,j}^2} \sum_{j>i} l_{i,j} \hat{\epsilon}_i \hat{\epsilon}_j = \frac{\hat{\boldsymbol{\epsilon}}^T \boldsymbol{L}^- \hat{\boldsymbol{\epsilon}}}{tr(\boldsymbol{L}^- \boldsymbol{L}^-)},$$

and the estimator of the proportion of variance modeled in $\boldsymbol{L}$ *is the ratio between two quadratic forms given by:*

$$\frac{\widehat{\sigma}_l^2}{\widehat{\sigma}_T^2} = \frac{\hat{\boldsymbol{\epsilon}}^T \boldsymbol{L}^- \hat{\boldsymbol{\epsilon}}}{\frac{1}{n} tr(\boldsymbol{L}^- \boldsymbol{L}^-) \hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}.$$

The above theorem provides a closed form estimator for a variance component and the proportion of variance corresponding to a covariance matrix $\boldsymbol{L}$ when it represents either the only modeled source of variation in the model, or when it is orthogonal to all other modeled sources of variation. It is straightforward to obtain closed form expressions in the

more complicated case of multiple modeled sources of variation that are not orthogonal.

Computation

Variance component estimators

While any unbiased estimator of $\widehat{\boldsymbol{\beta}}$ suffices to generate residuals $\hat{\boldsymbol{\epsilon}}$ and use them to obtain variance component estimators as $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}^d$, a more efficient estimator iterates between estimating $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ as follows:

1. Initialization step: set $\widehat{\boldsymbol{\beta}}^{(0)} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y}$.

2. Iteration step:

   (a) Given the kth estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}^{(k)}$, set $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \mathbf{W}\boldsymbol{\beta}^{(k)}$. Let $\tilde{\boldsymbol{\epsilon}}$ denote the vector of upper diagonal matrix (including the diagonal) of $\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$. Set $\widehat{\boldsymbol{\sigma}}^{2,(k)} = (\widehat{\sigma}_e^{2,(k)}, \widehat{\sigma}_a^{2,(k)}, \ldots, \widehat{\sigma}_k^{2,(k)}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}$.

   (b) Given the kth estimator of $\boldsymbol{\sigma}^2$, $\widehat{\boldsymbol{\sigma}}^{2,(k)}$, let $\widehat{\boldsymbol{\Sigma}}^{(k)} = \widehat{\sigma}_e^{2,(k)}\mathbf{I}_{n\times n} + \widehat{\sigma}_a^{2,(k)}\mathbf{A} + \ldots + \widehat{\sigma}_k^{2,(k)}\mathbf{K}$ with inverse $\widehat{\boldsymbol{\Sigma}}^{-1,(k)}$. Set $\widehat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{W}^T\widehat{\boldsymbol{\Sigma}}^{-1,(k)}\mathbf{W})^{-1}\mathbf{W}^T\widehat{\boldsymbol{\Sigma}}^{-1,(k)}\boldsymbol{y}$

The iteration step repeats until convergence.

Confidence intervals for the variance components

From Lemma 4 in the appendix, any variance component (or sum of variance components) is given as a quadratic form. Let $\mathbf{Q}$ be the quadratic form corresponding to a variance component estimate $\widehat{\sigma}_l^2$, such that $\widehat{\sigma}_l^2 = \hat{\boldsymbol{\epsilon}}^T\mathbf{Q}\hat{\boldsymbol{\epsilon}}$. This $\widehat{\sigma}_l^2$ is distributed as the sum of

independent $\chi^2_{(1)}$ variables in $\sum_{i=1}^n \lambda_i \chi^2_{(1)}$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{Q}\text{cov}(\hat{\boldsymbol{\epsilon}})$. In practice, for $\text{cov}(\hat{\boldsymbol{\epsilon}})$ we use the estimated $\widehat{\boldsymbol{\Sigma}}(\widehat{\sigma}^2_e, \ldots, \widehat{\sigma}^2_k)$. Functions in the R package CompQuadForm (Duchesne and de Micheaux, 2010) calculate the probability function (or survival function) of this quadratic form based on $\lambda_1, \ldots, \lambda_n$. While it takes times to compute the eigenvalues, once they are computed, calculating the probabilities associated with the quadratic form is quick. We can test the hypothesis $H_0 : \sigma^2_l = 0$ by calculating the probability

$$\Pr\left(\boldsymbol{\epsilon}^T \mathbf{Q} \boldsymbol{\epsilon} = 0\right) = 1 - \Pr\left(\boldsymbol{\epsilon}^T \mathbf{Q} \boldsymbol{\epsilon} > 0\right),$$

and calculate two-sided confidence intervals for $\widehat{\sigma}^2_l$ by calculating the appropriate quantiles of the survival probability. For example, for a 95% confidence interval we take the values (c_1, c_2) for which

$$\begin{aligned} c_1 &= u : \Pr\left(\boldsymbol{\epsilon}^T \mathbf{Q} \boldsymbol{\epsilon} > u\right) = 0.025 \\ c_2 &= u : \Pr\left(\boldsymbol{\epsilon}^T \mathbf{Q} \boldsymbol{\epsilon} > u\right) = 0.975. \end{aligned}$$

We find these values using a binary search on the interval $[0, \widehat{\sigma}^2_T]$.

Computing heritability estimates and their confidence intervals

Suppose that the variance component corresponding to the kinship matrix is σ^2_k, with quadratic form denoted by $\mathbf{Q}_k$. We estimate heritability as $\widehat{h}_k = \widehat{\sigma}^2_k / \widehat{\sigma}^2_T$. However, we cannot use the confidence intervals for σ^2_k to construct confidence intervals for h_k. Instead, we note that the point estimate $\widehat{h}_k$ is given by:

$$\widehat{h}_k = \frac{\hat{\boldsymbol{\epsilon}}^T \mathbf{Q}_k \hat{\boldsymbol{\epsilon}}}{\frac{1}{n} \hat{\boldsymbol{\epsilon}}^T \mathbf{I} \hat{\boldsymbol{\epsilon}}} \sim \frac{\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{1/2} \mathbf{Q}_k \widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{x}}{\frac{1}{n} \boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{x}} = \frac{\boldsymbol{x}^T \mathbf{F} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{G} \boldsymbol{x}}$$

where $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, for $\mathbf{F} = \widehat{\boldsymbol{\Sigma}}^{1/2}\mathbf{Q}_k\widehat{\boldsymbol{\Sigma}}^{1/2}$ and $\boldsymbol{G} = \widehat{\boldsymbol{\Sigma}}/n$. Thus, it is a ratio between two quadratic forms in (what we assume are) normal variables. For the squared root $\widehat{\boldsymbol{\Sigma}}^{1/2}$, we use the Cholesky decomposition of $\widehat{\boldsymbol{\Sigma}}$.

We use the saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables, proposed by Lieberman (1994). Complete detailed are provided in the appendix. In brief, for each potential value of h_k, say h_k^*, we can calculate the survival probability $Pr\left(h_k \geq h_k^*\right)$ using the saddlepoint approximation. Each such calculation requires as input $d_1^*, \ldots, d_n^*$, the eigenvalues of the matrix $\mathbf{D}^* = \mathbf{F} - h_k^*\boldsymbol{G}$. We apply a binary search on the potential values $h_k^* \in [0, 1]$ to find end points c_1 and c_2 for the confidence intervals, as was done for calculating a confidence intervals for σ_k^2.

Fast computation when genetic relatedness is the only modeled source of correlation

If we have only have a single kinship matrix $\mathbf{K}$ modeling the phenotypic variance, we can compute the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix $\mathbf{K}^-$ once, and then transform these eigenvalues to obtain the eigenvalues $d_1^*(h_k^*), \ldots, d_n^*(h_k^*)$ for each value h_k^*, and save computation time. To see this, suppose that $\boldsymbol{u}$ is an eigenvector of $\mathbf{K}^-$ with an eigenvalue λ. Then, by definition, $\mathbf{K}^-\boldsymbol{u} = \lambda\boldsymbol{u}$. Since $\boldsymbol{\Sigma} = \sigma_k^2(\mathbf{K}^- + \mathbf{I}) + \sigma_e^2\mathbf{I}$, it is straightforward to see that $\boldsymbol{u}$ is also an eigenvector of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}\boldsymbol{u} = \left[\sigma_k^2(\mathbf{K}^- + \mathbf{I}) + \sigma_e^2\mathbf{I}\right]\boldsymbol{u} = \left(\sigma_k^2\lambda + \sigma_k^2 + \sigma_e^2\right)\boldsymbol{u}.$$

Similarly, $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{\Sigma}^{1/2}$ with an eigenvalue $\sqrt{\sigma_k^2\lambda + \sigma_k^2 + \sigma_e^2}$, which finally leads to the transformation between an eigenvalue λ_i of $\mathbf{K}^-$ to an eigenvalue of $\mathbf{D}^*(h_k^*)$:

$$d_i^*(h_k^*, \lambda_i) = \frac{1}{2\sum_{i<j} v_{ij}^2}\lambda_i(\lambda_i\sigma_k^2 + \sigma_k^2 + \sigma_e^2) - h_k^*(\lambda_i\sigma_k^2 + \sigma_k^2 + \sigma_e^2)/n.$$

As before, we use the estimated $\widehat{\sigma}_k^2, \widehat{\sigma}_e^2$ instead of the true unknown values.

Meta-analysis across studies when kinship is the only source of correlation

Suppose that there are S studies that we wanted to combine in meta-analysis. We assume that kinship is the only source of correlation. Each study has a vector of residuals $\hat{\boldsymbol{\epsilon}}_s = (\hat{\epsilon}_{s,1}, \ldots, \hat{\epsilon}_{s,n_s})^T, s = 1, \ldots, S$. Consider the Haseman-Elston regression, but incomplete, so that only the pairs of multiplied residuals within study $\hat{\epsilon}_{s,i}\hat{\epsilon}_{s,j}$ are regressed against entries of the kinship covariance matrix, but not $\hat{\epsilon}_{s,i}\hat{\epsilon}_{t,j}$ for $s \neq t$. For this, we do not need to assume that a participant in one study is genetically unrelated of a participant in another study. The meta-analytic estimator of σ_e^2 is given by $\widehat{\sigma}_e^2 = \sum_{s=1}^S \sum_{i=1}^{n_s} \hat{\epsilon}_{s,i}^2 / \sum_{s=1}^S n_s$. Let $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}}_1^T, \ldots, \hat{\boldsymbol{\epsilon}}_S)^T$. The meta-analytic kinship variance component estimator is given by

$$\widehat{\sigma}_k^2 = \frac{1}{\mathrm{tr}\left(\mathbf{K}_S^-\mathbf{K}_S^-\right)}\hat{\boldsymbol{\epsilon}}^T\mathbf{K}_S^-\hat{\boldsymbol{\epsilon}}$$

where $\mathbf{K}_S^-$ is the block diagonal matrix that have all the study-specific kinship matrix (without their diagonal values) arranged diagonally, as

$$\mathbf{K}_S^- = \begin{pmatrix} \mathbf{K}_1^- & \mathbf{0} & \ldots & \ldots & \\ \mathbf{0} & \mathbf{K}_2^- & & & \mathbf{0} \\ \vdots & & \ddots & & \\ \vdots & \mathbf{0} & \mathbf{0} & & \mathbf{K}_s^- \end{pmatrix}$$

To see that this meta-analytic estimator of σ_k^2 is unbiased, note first that $\text{cov}(\hat{\boldsymbol{\epsilon}}) = (\sigma_e^2 + \sigma_k^2)\mathbf{I} + \sigma_k^2\mathbf{K}^-$, where now $\mathbf{K}^-$ has kinship coefficients for individuals across studies (and diagonals set to zero). From characteristics of quadratic forms:

$$\begin{aligned} E\left[\widehat{\sigma}_k^2\right] &= E\left[\frac{1}{\text{tr}\left(\mathbf{K}_S^-\mathbf{K}_S^-\right)}\hat{\boldsymbol{\epsilon}}^T\mathbf{K}_S^-\hat{\boldsymbol{\epsilon}}\right] = \frac{1}{\text{tr}\left(\mathbf{K}_S^-\mathbf{K}_S^-\right)}\text{tr}\left(\mathbf{K}_S^-\text{cov}(\hat{\boldsymbol{\epsilon}})\right) \\ &= \frac{1}{\text{tr}\left(\mathbf{K}_S^-\mathbf{K}_S^-\right)}\text{tr}\left\{\mathbf{K}_S^-\left[(\sigma_e^2+\sigma_k^2)\mathbf{I} + \sigma_k^2\mathbf{K}^-\right]\right\} \\ &= \frac{1}{\text{tr}\left(\mathbf{K}_S^-\mathbf{K}_S^-\right)}\text{tr}\left(\mathbf{K}_S^-\sigma_k^2\mathbf{K}^-\right) = \frac{1}{\text{tr}\left(\mathbf{K}_S^-\mathbf{K}_S^-\right)}\text{tr}\left(\mathbf{K}_S^-\sigma_k^2\mathbf{K}_S^-\right) = \sigma_k^2. \end{aligned}$$

Let $\mathbf{K}^- = \mathbf{K}_S^- + \mathbf{K}_C^-$, where $\mathbf{K}_C^-$ is the matrix of cross-study relatedness. Although the variance components estimates and their ratios depend only on $\mathbf{K}_S^-$, their distribution of the depend on $\mathbf{K}_C^-$ as well.

Computing the meta-analytic heritability estimator and confidence intervals. Suppose that each of S independent studies calculated the residuals from a "null model" (a regression model without genetic fixed effects other than PCs). Each study s reports:

1. $\mathcal{K}^s = 2\sum_{i<j} k_{ij}^2$,
2. $\widehat{\sigma}_{k,s}^2$,
3. $\widehat{\sigma}_{e,s}^2$,
4. The number of participants in the study n_s,
5. The eigenvalues $\lambda_1^s, \ldots, \lambda_{n_s}^s$ of its matrix $\mathbf{K}_s^-$.

The meta-analysis estimates of the kinship and error variance components, and $\mathcal{K}^S$ are:

$$\begin{aligned}
\widehat{\sigma}^2_k &= \frac{\sum_{s=1}^S \mathcal{K}^s \widehat{\sigma}^2_{k,s}}{\sum_{s=1}^S \mathcal{K}^s} \\
\widehat{\sigma}^2_e &= \frac{\sum_{s=1}^S n_s \widehat{\sigma}^2_{e,s}}{\sum_{s=1}^S n_s}, \\
\mathcal{K}^S &= \sum_{s=1}^S \mathcal{K}^s,
\end{aligned}$$

and the eigenvalues of the across-studies matrix $\mathbf{K}^-$ ($= \mathbf{K}_S^-$ under independence between studies) are taken to be $\lambda^1_1, \ldots, \lambda^1_{n_1}, \ldots, \lambda^S_1, \ldots, \lambda^S_{n_S}$. Using these, the central location calculates heritability estimates and confidence intervals.

The Hispanic Community Health Study/Study of Latinos

The HCHS/SOL (LaVange et al., 2010; Sorlie et al., 2010)) is a community based cohort study, following self-identified Hispanic individuals from four field centers (Chicago, IL; Miami, FL; Bronx, NY; and San Diego, CA). Individuals were sampled via a two-stage sampling scheme, in which households were randomly sampled from sampled community block units. Almost 13,000 study participants consented for genotyping. Correlation matrices to model environmental variance due to households and community block units were generated so that the i, j entry of a given matrix was set to 1 if the i and j individuals live in the same household (or community block unit), and 0 otherwise.

HCHS/SOL individuals were classified into 'genetic analysis groups': Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American. These groups are based on self reported ethnicities and genetic similarity (Conomos et al., 2016). This study was approved by the institutional review boards at each field center, where all participants gave

written informed consent.

Genotyping, imputation and quality control

Blood samples from HCHS/SOL individuals were genotyped on a custom array consisting of Illumina Omni 2.5M content plus ∼150,000 custom markers selected to include ancestry-informative markers, variants characteristic of Amerindian populations, known GWAS hits and other candidate gene polymorphisms. Quality control was similar to the procedure described in Laurie et al. (2010) and included checks for sample identity, batch effects, missing call rate, chromosomal anomalies (Laurie et al., 2012), deviation from Hardy-Weinberg equilibrium, Mendelian errors, and duplicate sample discordance. 12,784 samples passed quality control, and 2,232,944 SNPs passed quality filters. Pairwise kinship coefficients and principal components reflecting ancestry were estimated in an iterative procedure which accounts for admixture (Conomos et al., 2016). All common variants were used to estimate kinship coefficients. Finally, we removed some individuals at random to generate a set of 10,255 individuals without any pair having kinship coefficient higher than 2^{-11}.

Heritability and proportion of variance estimation in the HCHS/SOL

Due to the sampling structure of the HCHS/SOL, the correlation between individuals is modeled via a kinship matrix, and two matrices modeling environmental effects: household and community block unit matrices. For each investigated trait we estimated variance components corresponding to the three correlation matrices via the Haseman-Elston regression and estimated 95% confidence intervals.

The considered traits were FEV1 (a measure of lung function), standing height, depression score (CESD10, a sum of ten questionnaire items related to depression in the week prior to the clinic visit), SBP (systolic blood pressure), and dental caries, a count of tooth decays and cavities across all participant's teeth. All regression models were adjusted (via the design matrix $\mathbf{W}$) to the 5 first principal components, study center, age, sex, and genetic analysis group (in the pooled models). For some traits we used additional covariates.

To study the use of our method for meta-analysis when there are some related individuals across studies, we first generated a restricted data set of 7,848 individuals that none of them lived in the same household. We then treated each of the genetic analysis groups as a separate study, and used the proposed procedure for calculating heritability in each of the genetic analysis group and in meta-analysis. We also compared these analysis to the pooled analysis that modeled all 7,848 individuals together. Note that for this exercise we neglected block unit correlation, i.e. assumed that it does not contribute to the phenotypic variance.

Simulation studies

We study the accuracy of the proposed method for calculating confidence intervals in simulations. We used correlation matrices from the HCHS/SOL corresponding to kinship, household, and community block unit, to generate quantitative outcomes with realistic correlation structures. In any given simulation, data were sampled by first generating an error vector $\boldsymbol{e}^{ind}$ from a standard normal distribution. We simulated the covariance

structure

$$\text{cov}(\boldsymbol{e}) = \sigma_e^2\mathbf{I} + \sigma_k^2\mathbf{K} + \sigma_h^2\mathbf{H} + \sigma_c^2\mathbf{C} = \boldsymbol{\Sigma}$$

by taking $\boldsymbol{e} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{e}^{ind}$. The matrices $\mathbf{K}, \mathbf{H}$, and $\mathbf{C}$ were the kinship, household, and community matrices in the HCHS/SOL. The outcomes were simulated by

$$\boldsymbol{y} = 2 + 3\text{PC}_1 + \boldsymbol{e},$$

where PC_1 is the first principal component of the HCHS/SOL data. All simulations were performed 1,000 times.

In the first simulation study we set $\boldsymbol{\sigma} = (\sigma_e^2, \sigma_k^2, \sigma_h^2, \sigma_c^2) = (100, 40, 15, 2)$, and studied our method in settings of small sample size ($n = 1,000$) and large sample size ($n = 12,784$). In a second simulation study we set $\boldsymbol{\sigma} = (\sigma_e^2, \sigma_k^2, \sigma_h^2, \sigma_c^2) = (100, \sigma_k^2, 0, 0)$, with $\sigma_k^2 \in \{0, 40\}$. Here we also considered settings of small and large sample sizes, and in addition, we randomly divided the large dataset into 5 subgroups, to generate data mimicking five different studies with possible genetic relatedness between participants of different studies, and studied the meta-analysis approach in this setting. We randomly partitioned the data to subgroups four times, to make sure that results did not depend on a specific partition.

Results

Simulation studies

Figure 1 provides the results from simulations with three modeled sources of variation: kinship, household, and community block unit, mimicking the HCHS/SOL study. The coverages of the confidence intervals obtained in the simulations with large sample size

were very close to the nominal value, but a bit lower (92%) for the variance component corresponding to community block unit (which was simulated to account for 1% proportion of variance). The confidence intervals obtained from the simulations with small sample size have larger coverage than nominal value (98%-99%), and are much wider. However, they are not trivially large (i.e. they are not of the form $[0, 1]$).

Figure 2 provides the results from simulations in which only the kinship matrix was modeled as a single source of phenotypic variation between individuals. The coverages were good, with the tightest confidence intervals estimated in the large sample size, when all individuals were pooled together as in a single large study, and the estimated genetic relatedness between all individuals were used. The confidence intervals from the meta-analysis were wider, and the confidence intervals from the small sample size simulations were quite wide, as expected.

Heritability estimation in the HCHS/SOL

For each of the investigated traits, Figure 3 provides estimated proportion of variance (heritability and environmental variance) and 95% confidence intervals from the Full data set, that included environmentally correlated individuals, and from the restricted data set that did not include environmentally correlated individuals. We used the restricted data set to compare a pooled analysis, genetic analysis group specific analyses, and meta-analysis that ignores the correlation between individuals from different genetic analysis groups, to mimic meta-analysis across different studies. For all analyses, numerical estimates are provided Table 1 in the appendix.

In the Full dataset, for most traits (height, SBP, Dental caries, and FEV1), the heritability was larger than the proportion of variance attributed to the modeled environmental components. However, the proportion of variance of the depression score due to environmental effects was higher than the heritability, and was also statistically significant (p-value $= 0.002$), while the heritability was not (p-value $= 0.1$). The heritability of height was estimated as almost 60%, consistent with other estimates from GWAS, but the 95% confidence intervals was $(0.47, 0.69)$.

Considering the restricted data set, the analyses of specific genetic analysis groups yielded wide confidence intervals, which often included zero. This is expected due to low power. In addition, the meta-analyses that did not account for the correlations between the genetic analysis groups had wider confidence intervals than the corresponding pooled analyses.

Discussion

In this manuscript we investigate the properties of Haseman-Elston regression estimators of variance components. We get a closed-form expression for the variance estimators, and use them to characterize the distribution of the estimated variance components and ratios of variance, and to compute confidence intervals. Our confidence intervals require normality of the residuals from the trait regression model after adjusting for covariates. We further show how to obtain unbiased estimates of the variance components and proportions of variance by meta-analyzing information from multiple studies. In this case, the heritability estimates are unbiased even if individuals are related between studies, but the asymptotic distribution

of the estimators depends on the unknown (and non-estimated) kinship coefficients of cross-study individuals.

We show in simulations based on the HCHS/SOL correlation structure that the coverage of our confidence intervals is good both in pooled analysis, and in meta-analysis (even when individuals are related between studies) while being quite conservative when the sample size is small. More work is needed to study the analytic properties of the confidence intervals in meta-analysis when individuals are related between studies.

Software

An R code for estimating heritability (or proportion of variances due to other modeled factors), and their confidence intervals, together with an example script, can be found at `https://github.com/tamartsi/Heritability_CIs`.

Acknowledgements

The author thanks Dr. Bruce Weir and Dr. Bill Hill for reviewing earlier versions of the manuscripts, and the staff and participants of HCHS/SOL for their important contributions. This work was supported in part by NHLBI HHSN268201300005C. The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National In-

stitute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.

References

Burch, B. D. (2011). Assessing the performance of normal-based and REML-based confidence intervals for the intraclass correlation coefficient. *Computational Statistics & Data Analysis*, **55** 1018–1028.

Conomos, M., Laurie, C., Stilp, A., Gogarten, S., McHugh, C., Nelson, S., Sofer, T., Fernandez-Rhodes, L., Justice, A., Graff, M., Young, K., Seyerle, A., Avery, C., Taylor, K., Rotter, J., Talavera, G., Daviglus, M., Wassertheil-Smoller, S., Schneiderman, N., Heiss, G., Kaplan, R., Franceschini, N., Reiner, A., Shaffer, J., Barr, R., Kerr, K., Browning, S., Browning, B., Weir, B., Avilés-Santa, M., Papanicolaou, G., Lumley, T., Szpiro, A., North, K., Rice, K., Thornton, T. and Laurie, C. (2016). Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*, **98** 165 – 184.

Duchesne, P. and de Micheaux, P. L. (2010). Computing the distribution of quadratic

forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, **54** 858–862.

Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., Keurentjes, J. J. and van Eeuwijk, F. A. (2015). Marker-based estimation of heritability in immortal populations. *Genetics*, **199** 379–398.

Laurie, C. et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, **34** 591–602.

Laurie, C. C., Laurie, C. A. et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, **44** 642–650.

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J. et al. (2010). Sample design and cohort selection in the hispanic community health study/study of latinos. *Annals of epidemiology*, **20** 642–649.

Lieberman, O. (1994). Saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables. *Journal of the American Statistical Association*, **89** 924–928.

Schweiger, R., Kaufman, S., Laaksonen, R., Kleber, M. E., März, W., Eskin, E., Rosset, S. and Halperin, E. (2016). Fast and Accurate Construction of Confidence Intervals for Heritability. *The American Journal of Human Genetics*, **98** 1181–1192. URL http://dx.doi.org/10.1016/j.ajhg.2016.04.016.

Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L., Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., LaVange, L., Chambless, L. E. and Heiss, G. (2010). Design and implementation of the hispanic community health study/study of latinos. *Annals of epidemiology*, **20** 629–641.

Zaitlen, N. and Kraft, P. (2012). Heritability in the genome-wide association era. *Human genetics*, **131** 1655–1664.

A Mathematical derivation

Suppose that a quantitative trait Y, measured on n individuals, follows the regression model

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n$$

where

$$E[\boldsymbol{\epsilon}] = \mathbf{0} \tag{1}$$

$$\text{var}[\boldsymbol{\epsilon}] = \sigma_e^2 \mathbf{I}_{n\times n} + \sigma_a^2 \mathbf{A}_1 + \ldots + \sigma_k^2 \mathbf{K} = \boldsymbol{\Sigma} \tag{2}$$

and $\mathbf{A}, \ldots, \mathbf{K}$ are $n \times n$ matrices modeling correlations between individuals. Let $a_{i,j}, \ldots, k_{i,j}$ denote the i, j entries of the matrices $\mathbf{A}, \ldots, \mathbf{K}$. Assuming that random effects due to $\mathbf{A}, \ldots, \mathbf{K}$ are independent, we have that:

$$E[\epsilon_i \epsilon_j] = \text{cov}(\epsilon_i, \epsilon_j) = \sigma_e^2 \mathcal{I}_{(i=j)} + \sigma_a^2 a_{i,j} + \ldots + \sigma_k^2 k_{i,j}.$$

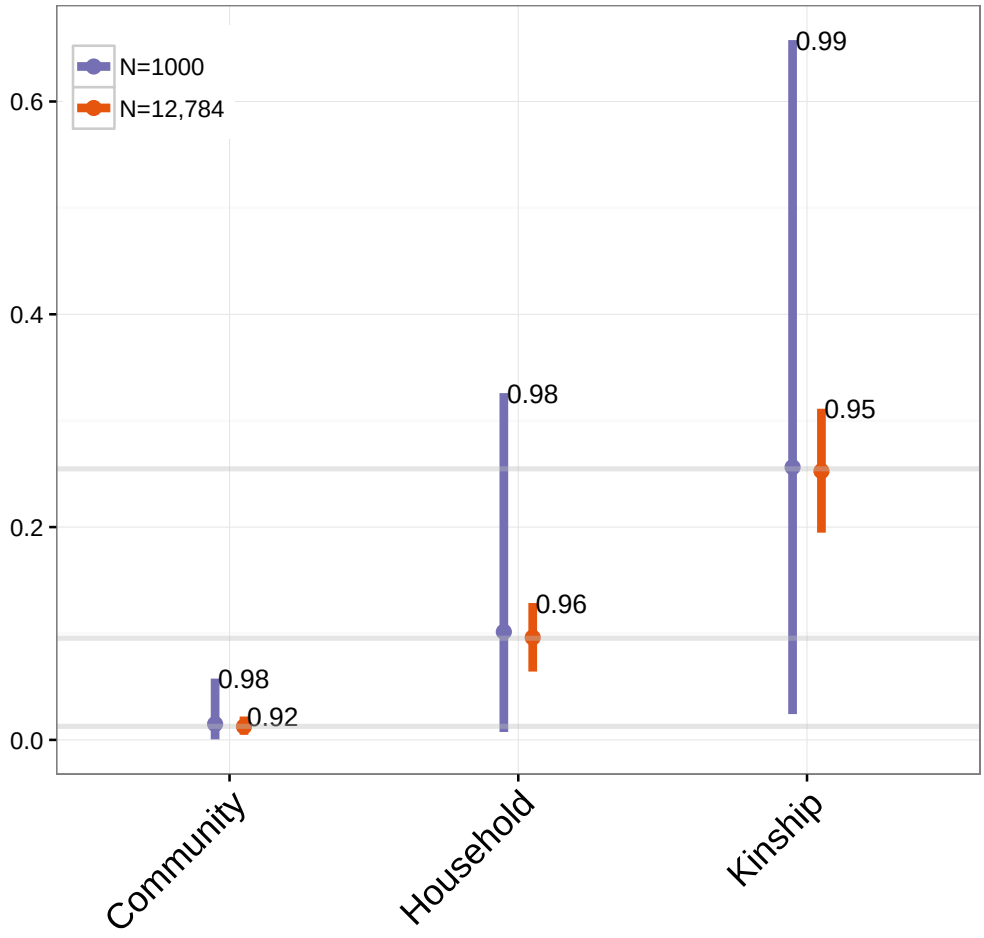

Figure 1: Results from 1,000 simulations estimating the proportion of variation attributed to community block unit ($\sigma_c^2 = 2$), household ($\sigma_h^2 = 15$), and kinship ($\sigma_k^2 = 40$) out of the total variance ($\sigma_T^2 = \sigma_c^2 + \sigma_h^2 + \sigma_k^2 + \sigma_e^2$, with $\sigma_e^2 = 100$), in scenarios with small and large sample sizes. The points represent the means of the estimated proportions of variance, and the low and high end points of the intervals represent the means of the low and high end points of the estimated confidence intervals. The coverages of the estimated confidence intervals, defined as the proportion of simulations in which the true proportion of variance was in the estimated confidence interval, are written by each line.

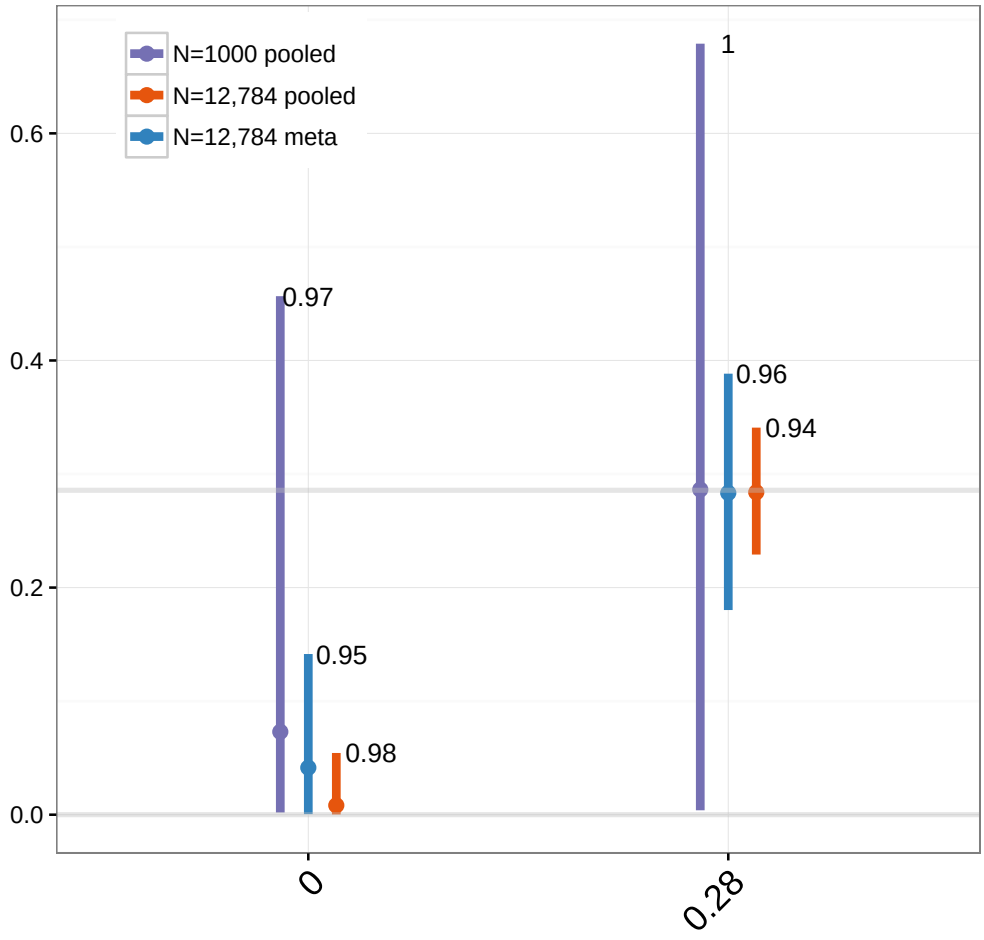

Figure 2: Results from 1,000 simulations estimating the proportion of variation attributed to kinship ($\sigma_k^2 \in \{0, 40\}$) out of the total variance ($\sigma_T^2 = \sigma_k^2 + \sigma_e^2$, with $\sigma_e^2 = 100$), in scenarios with small and large sample sizes, and when randomly dividing the large data set to 5 studies and meta-analyzing without accounting for relatedness between individuals in different studies. The points represent the means of the estimated proportions of variance, and the low and high end points of the intervals represent the means of the low and high end points of the estimated confidence intervals. The coverages of the estimated confidence intervals, defined as the proportion of simulations in which the true proportion of variance was in the estimated confidence interval, are written by each line.

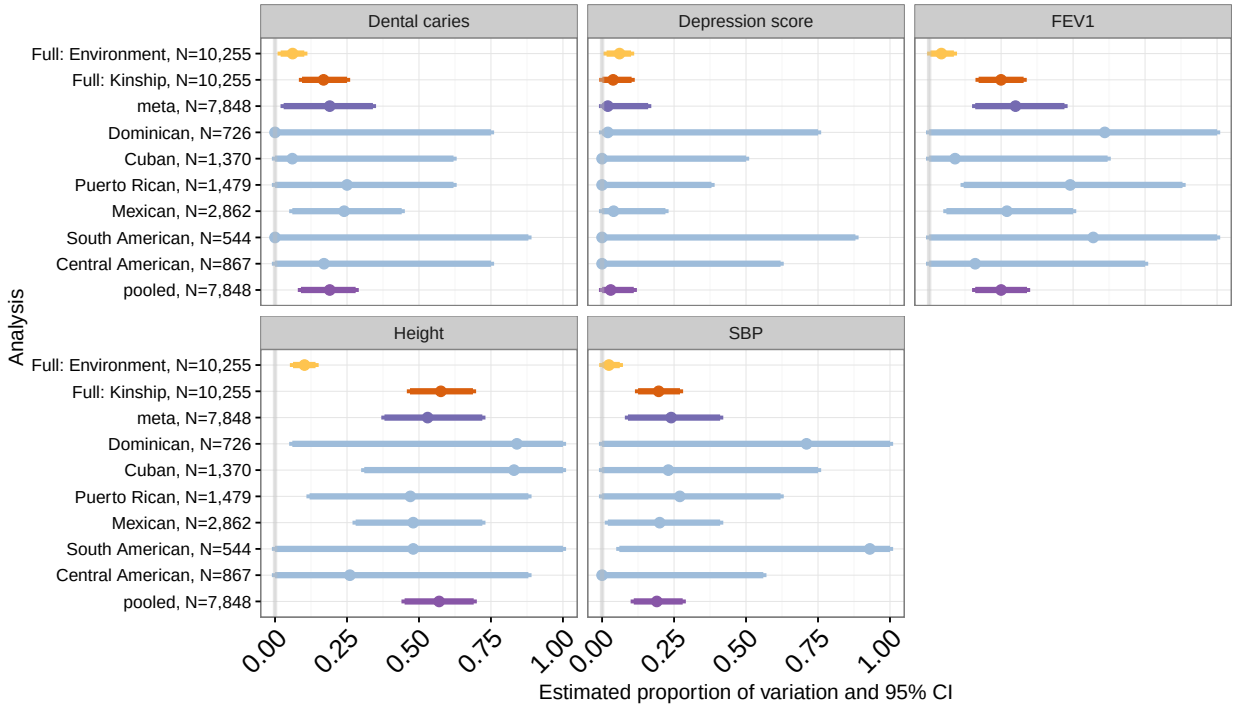

Figure 3: Estimated proportions of variance from the various subsets of the HCHS/SOL data. The Full dataset included 10,255 individuals with mutual kinship coefficient smaller than 2^{-11}. Using Full, we estimated both heritability and the proportion of variance that is due to modeled environmental effects: the sum of the variance components corresponding to household and community block unit sharing. A restricted data set included 7,848 individuals from separate households and was used to compare meta and pooled analysis heritability estimates, where the meta-analysis used information from each of the genetic analysis groups. Dental caries is a measure of teeth damage. Depression score is a summation of responses to questions related to depressive behavior or feelings in the week prior to a participant's clinic visit. FEV1 is a measure of lung function. SBP is systolic blood pressure.

Let $\widehat{\boldsymbol{\beta}}$ be an unbiased estimator of $\boldsymbol{\beta}$, and let $\hat{\epsilon}_i = y_i - \boldsymbol{w}_i^T\widehat{\boldsymbol{\beta}}$ be an estimator $\epsilon_i, i = 1, \ldots, n$. We estimate the variance components in a residual regression, i.e. by taking the vector all unique pairs of residuals $\hat{\epsilon}_i\hat{\epsilon}_j, i \leq j$ (we can do it by taking the upper diagonal sub-matrix of $\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$ that includes the diagonal), denoted by $\tilde{\boldsymbol{\epsilon}}^d$ and regressing it according to the above model. The regression design matrix is given by:

$$\mathbf{X} = \begin{pmatrix} 1 & a_{1,1} & \ldots & k_{1,1} \\ 0 & a_{1,2} & \ldots & k_{1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{1,n} & \ldots & k_{1,n} \\ 1 & a_{2,2} & \ldots & k_{2,2} \\ 0 & a_{2,3} & \ldots & k_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{2,n} & \ldots & k_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{n-1,n-1} & \ldots & k_{n-1,n-1} \\ 0 & a_{n-1,n} & \ldots & k_{n-1,n} \\ 1 & a_{n,n} & \ldots & k_{n,n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & a_{1,2} & \ldots & k_{1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{1,n} & \ldots & k_{1,n} \\ 1 & 1 & 1 & 1 \\ 0 & a_{2,3} & \ldots & k_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{2,n} & \ldots & k_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \\ 0 & a_{n-1,n} & \ldots & k_{n-1,n} \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

(because $a_{i,i}, \ldots, k_{i,i} = 1$ for all i). Denote, for simplicity of presentation, the vector of off-diagonal elements of $\mathbf{A}, \ldots, \mathbf{K}$ by $\boldsymbol{l} = (l_{1,2}, l_{1,3}, \ldots, l_{1,n}, l_{2,3}, \ldots, l_{n-1,n})^T, l = 1, \ldots, k$, and the vector of off-diagonal elements of $\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$ by $\tilde{\boldsymbol{\epsilon}}$. Then the least squares estimator of

$(\sigma_e^2, \sigma_a^2, \ldots, \sigma_k^2)$ is given by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}^d$. Clearly, we have that

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n & n & n & \ldots & n \\ n & n+\boldsymbol{a}^T\boldsymbol{a} & n+\boldsymbol{a}^T\boldsymbol{b} & \ldots & n+\boldsymbol{a}^T\boldsymbol{k} \\ \vdots & & & & \vdots \\ n & n+\boldsymbol{k}^T\boldsymbol{a} & n+\boldsymbol{k}^T\boldsymbol{b} & \ldots & n+\boldsymbol{k}^T\boldsymbol{k} \end{pmatrix}.$$

This is most likely a positive definite matrix as (we assume that) the matrices $\mathbf{A}, \ldots, \mathbf{K}$ are not highly correlated. In addition, we have that

$$\mathbf{X}^T\tilde{\boldsymbol{\epsilon}} = \begin{pmatrix} \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \sum_{i=1}^n \hat{\epsilon}_i^2 + \boldsymbol{a}^T\tilde{\boldsymbol{\epsilon}} \\ \vdots \\ \sum_{i=1}^n \hat{\epsilon}_i^2 + \boldsymbol{k}^T\tilde{\boldsymbol{\epsilon}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \vdots \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \end{pmatrix} + \begin{pmatrix} 0 \\ \boldsymbol{a}^T\tilde{\boldsymbol{\epsilon}} \\ \vdots \\ \boldsymbol{k}^T\tilde{\boldsymbol{\epsilon}} \end{pmatrix}.$$

Lemma 1:

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Leftrightarrow \left(\mathbf{X}^T\mathbf{X}\right)\begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Proof: Because $\left(\mathbf{X}^T\mathbf{X}\right)$is non-singular, and from the properties of the inverse matrix, we have that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{v} = \boldsymbol{v}$ for every $\boldsymbol{v}$. ∎

Lemma 2: *Variance component estimators corresponding to the matrices* $\mathbf{A}, \ldots, \mathbf{K}$ *depend only on the between-observation residuals of the form* $\epsilon_i\epsilon_j$ *for* $i \neq j$ *and do not depend on*

$\epsilon_i^2, i = 1, \ldots, n$. **Proof:** By noting that

$$\left(\mathbf{X}^T\mathbf{X}\right)\begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

we get from Lemma 1 that

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\begin{pmatrix} \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \vdots \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

which proves that the term $\sum_{i=1}^n \hat{\epsilon}_i^2$ contributes only to the estimator $\widehat{\sigma}_e^2$. ∎

Lemma 3: *Denote by* $\sigma_T^2 = \sigma_e^2 + \sigma_a^2 + \ldots + \sigma_k^2$. *Then* $\widehat{\sigma}_T^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$.

Proof: We show that $\widehat{\sigma}_e^2 + \widehat{\sigma}_a^2 + \ldots + \widehat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$. In the proof of Lemma 2 we saw that

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Since this $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ is symmetric, it follows that

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 & \ldots & 0 \end{pmatrix}$$

Therefore

$$\begin{aligned}\widehat{\sigma}_e^2 + \widehat{\sigma}_a^2 + \ldots + \widehat{\sigma}_k^2 &\equiv \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \left(\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}\right) \\ &= \begin{pmatrix} \frac{1}{n} & 0 & \ldots & 0 \end{pmatrix} \left(\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2,\end{aligned}$$

which completes the proof. ∎

Lemma 4: *An estimator of the ratio between any variance component (or sum of variance components) and the total variance is a ratio between two quadratic forms.*

Proof: For $\boldsymbol{L} = \mathbf{A}, \ldots, \mathbf{K}$, a quantity of the form $\boldsymbol{l}^T\tilde{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}^T\boldsymbol{L}^-\hat{\boldsymbol{\epsilon}}/2$, where the matrix $\boldsymbol{L}^{-1}$ is the matrix $\boldsymbol{L}$ with all diagonal values set to 0. An estimator a variance component σ_l^2 is a linear sum of the quadratic forms $\hat{\boldsymbol{\epsilon}}^T\mathbf{A}^-\hat{\boldsymbol{\epsilon}}, \ldots, \hat{\boldsymbol{\epsilon}}^T\mathbf{K}^-\hat{\boldsymbol{\epsilon}}$, with coefficients the entries of the corresponding row of $(\mathbf{X}^T\mathbf{X})^{-1}$. Since a weighted sum of quadratic forms is a quadratic form, any variance component (and a sum of variance components) is also a quadratic form. Similarly, the total variance estimators is the quadratic form $\frac{1}{n}\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}$. ∎

Theorem: *We say that two matrices* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *are orthogonal in the trace inner product, or "trace orthogonal" if* $tr(\mathbf{C}_1\mathbf{C}_2) = 0$. *If a matrix* $\boldsymbol{L}^-$ *is trace orthogonal to all other matrices in the set* $\{\mathbf{A}^-, \ldots, \mathbf{K}^-\}$, *then*

$$\widehat{\sigma}_l^2 = \frac{1}{\sum_{j>i} l_{i,j}^2} \sum_{j>i} l_{i,j}\hat{\epsilon}_i\hat{\epsilon}_j.$$

Proof: Without loss of generality, assume that $\mathbf{A}$ is trace orthogonal to $\boldsymbol{B}, \ldots, \mathbf{K}$. First note that for the symmetric matrices with diagonal values set to zero $\mathbf{A}^-, \ldots, \mathbf{K}^-$,

$\text{tr}\left(\mathbf{A}^{-}\boldsymbol{L}^{-}\right) = 0$ if and only if $\boldsymbol{a}^T\boldsymbol{l} = 0$. Then

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n & n & n & \dots & n \\ n & n+\boldsymbol{a}^T\boldsymbol{a} & n & \dots & n \\ n & n & n+\boldsymbol{b}^T\boldsymbol{b} & \dots & n+\boldsymbol{b}^T\boldsymbol{k} \\ \vdots & \vdots & \vdots & & \vdots \\ n & n & n+\boldsymbol{k}^T\boldsymbol{b} & \dots & n+\boldsymbol{k}^T\boldsymbol{k}. \end{pmatrix}$$

Denote by $(\mathbf{X}^T\mathbf{X})^{-1}_{[i,j]}$ the i,j element in the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$. First, we notice that the entries $(\mathbf{X}^T\mathbf{X})^{-1}_{[2,j]}, j = 3,\dots,k+1$ are all 0, because the $(\mathbf{X}^T\mathbf{X})^{-1}_{[i,j]}$ entry is a constant times the i,j minor of $(\mathbf{X}^T\mathbf{X})$, which has two identical columns (corresponding to the $1st$ and $2nd$ columns of $(\mathbf{X}^T\mathbf{X})$ when removing its $2nd$ row). Since the sum of the $2nd$ row of $(\mathbf{X}^T\mathbf{X})^{-1}$ is equal to 0, as we saw before, we get that $(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} = -(\mathbf{X}^T\mathbf{X})^{-1}_{[2,2]}$.

We now argue that

$$\begin{aligned} \widehat{\sigma}^2_a &\equiv (\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]}\sum_{i=1}^n \epsilon_i^2 + (\mathbf{X}^T\mathbf{X})^{-1}_{[2,2]}\left(\sum_{i=1}^n \epsilon_i^2 + \boldsymbol{a}^T\tilde{\boldsymbol{\epsilon}}\right) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]}\sum_{i=1}^n \epsilon_i^2 - (\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]}\left(\sum_{i=1}^n \epsilon_i^2 + \boldsymbol{a}^T\tilde{\boldsymbol{\epsilon}}\right) \\ &= -(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]}\boldsymbol{a}^T\tilde{\boldsymbol{\epsilon}} \overset{(4)}{=} \frac{1}{\sum_{j>i} a_{i,j}^2}\sum_{j>i} a_{i,j}\epsilon_i\epsilon_j, \end{aligned}$$

where we need to show that equality (4) holds to complete the proof. We need to show that $-(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} = \frac{1}{\sum_{j>i} a_{i,j}^2}$. Consider now the matrix $\mathbf{X}^T\mathbf{X}$. One can derive its determinant

from its second row, as:

$$\begin{aligned}
|\mathbf{X}^T\mathbf{X}| &= (\mathbf{X}^T\mathbf{X})_{[2,1]}M_{2,1} - (\mathbf{X}^T\mathbf{X})_{[2,2]}M_{2,2} + \ldots + (-1)^{k+1}(\mathbf{X}^T\mathbf{X})_{[2,1k]}M_{2,k} \\
&= nM_{2,1} - (n + \boldsymbol{a}^T\boldsymbol{a})M_{2,2} + 0 + \ldots 0 \\
&= n|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} - (n + \boldsymbol{a}^T\boldsymbol{a})|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,2]} \\
&= n|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} - (n + \boldsymbol{a}^T\boldsymbol{a})|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} \\
&= -\boldsymbol{a}^T\boldsymbol{a}|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]}
\end{aligned}$$

Therefore, we get that $-(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} = \frac{1}{\boldsymbol{a}^T\boldsymbol{a}}$, which completes the proof. ∎

B Computation

B.1 Variance component estimators

While any unbiased estimator of $\widehat{\boldsymbol{\beta}}$ suffices to generate residuals $\hat{\boldsymbol{\epsilon}}$ and use them to obtain variance component estimators as $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}^d$, a more efficient estimator iterates between estimating $\boldsymbol{\beta}$ and the variance component estimator as follows:

1. Initialization step: set $\widehat{\boldsymbol{\beta}}^{(0)} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y}$.

2. Iteration step:

   (a) Given the kth estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}^{(k)}$, set $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \mathbf{W}\boldsymbol{\beta}^{(k)}$ and $\tilde{\boldsymbol{\epsilon}}$ is the vector of uppder diagonal matrix (including the diagonal) of $\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$. Set $\widehat{\boldsymbol{\sigma}}^{2,(k)} = (\widehat{\sigma}_e^{2,(k)}, \widehat{\sigma}_a^{2,(k)}, \ldots, \widehat{\sigma}_k^{2,(k)}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}^d$.

   (b) Given the kth estimator of $\boldsymbol{\sigma}^2$, $\widehat{\boldsymbol{\sigma}}^{2,(k)}$, let $\widehat{\boldsymbol{\Sigma}}^{(k)} = \widehat{\sigma}_e^{2,(k)}\mathbf{I}_{n\times n} + \widehat{\sigma}_a^{2,(k)}\mathbf{A} + \ldots + \widehat{\sigma}_k^{2,(k)}\mathbf{K}$ with inverse $\widehat{\boldsymbol{\Sigma}}^{-1,(k)}$. Set $\widehat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{W}^T\widehat{\boldsymbol{\Sigma}}^{-1,(k)}\mathbf{W})^{-1}\mathbf{W}^T\widehat{\boldsymbol{\Sigma}}^{-1,(k)}\boldsymbol{y}$

The iteration step repeats until convergence.

B.2 Confidence intervals for the variance components

From Lemma 4, any variance components (or sum of variance components) is given as a quadratic form. Let $\mathbf{Q}$ be the quadratic form corresponding to a variance component estimate $\widehat{\sigma}_l^2$, such that $\widehat{\sigma}_l^2 = \hat{\boldsymbol{\epsilon}}^T \mathbf{Q} \hat{\boldsymbol{\epsilon}}$. Then this $\widehat{\sigma}_l^2$ is distributed as the sum of independent $\chi^2_{(1)}$ variables in $\sum_{i=1}^n \lambda_i \chi^2_{(1)}$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{Q}\text{cov}(\hat{\boldsymbol{\epsilon}})$. In practice, for $\text{cov}(\hat{\boldsymbol{\epsilon}})$ we use used the estimated $\widehat{\Sigma}(\widehat{\sigma}_e^2, \ldots, \widehat{\sigma}_k^2)$. Functions in the package CompQuadFrom calculate the probability function (or survival function) of this quadratic form based on $\lambda_1, \ldots, \lambda_n$. While it takes times to compute the eigenvalues, once they are computed, a calculating the probabilities associated with the quadratic form over a grid is simple and quick. We can test the hypothesis $H_0 : \sigma_l^2 = 0$ by calculating the probability

$$\Pr\left(\hat{\boldsymbol{\epsilon}}^T \mathbf{Q} \hat{\boldsymbol{\epsilon}} = 0\right) = 1 - \Pr\left(\hat{\boldsymbol{\epsilon}}^T \mathbf{Q} \hat{\boldsymbol{\epsilon}} > 0\right),$$

and calculate two-sided confidence intervals for $\widehat{\sigma}_l^2$ by calculating the survival probabilities over a grid, and taking the appropriate quantiles. For example, for a 95% confidence interval we take the values (c_1, c_2) for which

$$\begin{aligned} c_1 &= u : \Pr\left(\boldsymbol{\epsilon}^T \mathbf{Q} \boldsymbol{\epsilon} > u\right) = 0.025 \\ c_2 &= u : \Pr\left(\boldsymbol{\epsilon}^T \mathbf{Q} \boldsymbol{\epsilon} > u\right) = 0.975. \end{aligned}$$

We find these values using a binary search on the interval $[0, \widehat{\sigma}_T^2]$.

B.3 Computing heritability estimates and their confidence intervals

Suppose that the variance component corresponding to the kinship matrix is σ_k^2, which quadratic form denoted by $\mathbf{Q}_k$. We estimate heritability as $\widehat{h}_k = \widehat{\sigma}_k^2/\widehat{\sigma}_T^2$. However, we cannot use the confidence intervals for σ_k^2 to construct confidence intervals for h_k. Instead, we note that the point estimate is $\widehat{h}_k$ is given by:

$$\widehat{h}_k = \frac{\hat{\boldsymbol{\epsilon}}^T \mathbf{Q}_k \hat{\boldsymbol{\epsilon}}}{\frac{1}{n}\hat{\boldsymbol{\epsilon}}^T \mathbf{I} \hat{\boldsymbol{\epsilon}}} = \frac{\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{1/2} \mathbf{Q}_k \widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{x}}{\frac{1}{n}\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{x}} = \frac{\boldsymbol{x}^T \mathbf{F} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{G} \boldsymbol{x}}$$

where $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, for $\mathbf{F} = \widehat{\boldsymbol{\Sigma}}^{1/2} \mathbf{Q}_k \widehat{\boldsymbol{\Sigma}}^{1/2}$ and $\boldsymbol{G} = \widehat{\boldsymbol{\Sigma}}/n$. Thus, it is a ratio between two quadratic forms in (what we assume are) normal variables. For the squared root $\widehat{\boldsymbol{\Sigma}}^{1/2}$, we use the Cholesky decomposition of $\widehat{\boldsymbol{\Sigma}}$.

Now, we use the saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables, proposed by Lieberman (1994). For a given potential value of h_k, say h_k^*, we can calculate the survival probability

$$Pr\left(h_k \geq h_k^*\right) \cong 1 - \Phi(\hat{\xi}) + \phi(\hat{\xi}) \left[\frac{1}{\hat{z}} - \frac{1}{\hat{\xi}}\right]$$

where Φ and ϕ are the standard normal cdf and pdf, and

$$\begin{aligned} \hat{z} &= \hat{\omega} \left\{ 2 \sum_{i=1}^{n} \frac{d_i^{*2}}{(1 - 2\hat{\omega} d_i^*)^2} \right\}^{1/2} \\ \hat{\xi} &= \left\{ \sum_{i=1}^{n} \ln(1 - 2\hat{\omega} d_i^*) \right\}^{1/2} \operatorname{sgn}(\hat{\omega}) \end{aligned}$$

and $d_1^*, \ldots, d_n^*$ are the eigenvalues of the matrix $\mathbf{D}^* = \mathbf{F} - h_k^* \boldsymbol{G}$, and $\hat{\omega}$ is the corresponding saddlepoint satisfying

$$\sum_{i=1}^{n} \frac{d_i^*}{1 - 2\hat{\omega} d_i^*} = 0.$$

Confidence intervals are then built, as before, using a binary search to find the values satisfying the required probabilities at the tails.

B.3.1 A faster algorithm when the kinship matrix is the only source of correlation in the model

Computing the eigenvalues $d_1^*(h_k^*), \ldots, d_n^*(h_k^*)$ takes time. However, in the case where we only have a single kinship matrix, denoted by $\mathbf{K}$ we can compute the eigen decomposition of the matrix $\mathbf{K}^-$ once to obtain eigenvalues $\lambda_1, \ldots, \lambda_n$, and then transform these eigenvalues to obtain the eigenvalues $d_1^*(h_k^*), \ldots, d_n^*(h_k^*)$ for each value h_k^*. To see this, suppose that $\boldsymbol{u}$ is an eigenvector of $\mathbf{K}^-$ with eigenvalues λ. Then, by definition:

$$\mathbf{K}^- \boldsymbol{u} = \lambda \boldsymbol{u}.$$

Since $\boldsymbol{\Sigma} = \sigma_k^2(\mathbf{K}^- + \mathbf{I}) + \sigma_e^2\mathbf{I}$, it is straightforward to see that $\boldsymbol{u}$ is also an eigenvector of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}\boldsymbol{u} = \left[\sigma_k^2(\mathbf{K}^- + \mathbf{I}) + \sigma_e^2\mathbf{I}\right] \boldsymbol{u} = \left(\sigma_k^2\lambda + \sigma_k^2 + \sigma_e^2\right) \boldsymbol{u}.$$

Similarly, $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{\Sigma}^{1/2}$ with eigenvalue $\sqrt{\sigma_k^2\lambda + \sigma_k^2 + \sigma_e^2}$, which finally leads us to the transformation between an eigenvalue λ of $\boldsymbol{\Lambda}$ to an eigenvalue of $\mathbf{D}^* = \mathbf{F} - h_k^*\boldsymbol{G}$ given by:

$$d_i^*(h_k^*, \lambda_i) = \frac{1}{2\sum_{i<j} v_{ij}^2}\lambda_i(\lambda_i\sigma_k^2 + \sigma_k^2 + \sigma_e^2) - h_k^*(\lambda_i\sigma_k^2 + \sigma_k^2 + \sigma_e^2)/n.$$

As before, we use the estimated $\widehat{\sigma}_k^2, \widehat{\sigma}_e^2$ instead of the true unknown quantities.

B.3.2 Meta-analysis of across studies when kinship is the only source of correlation

A meta-analytic estimator. Suppose that there are S studies that we wanted to combined in meta-analysis. We assume that kinship is the only source of correlation. Each study has a vector of residuals $\hat{\boldsymbol{\epsilon}}_s = (\hat{\epsilon}_{s,1}, \ldots, \hat{\epsilon}_{s,n_s})^T, s = 1, \ldots, S$. Consider the Haseman-Elston regression, but incomplete, so that only the pairs of multiplied residuals within study are used (i.e. only $\hat{\epsilon}_{s,i}\hat{\epsilon}_{s,j}$ are regressed against entries of the kinship covariance matrix, but not $\hat{\epsilon}_{s,i}\hat{\epsilon}_{t,j}$). Therefore, cross-study kinship estimates are not used in the regression, however no assumption is made on them. In other words, we do not need to assume that participant in one study is genetically independent (no alleles shared IBD) of a participant in another study. It is straightforward to show that the meta-analytic estimator of σ_2^e is given by $\widehat{\sigma}_e^2 = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\hat{\epsilon}_{s,i}^2$. Let $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}}_1^T, \ldots, \hat{\boldsymbol{\epsilon}}_S)^T$. Then the meta-analysis kinship variance component estimator is given by

$$\widehat{\sigma}_k^2 = \frac{1}{\text{tr}\left(\mathbf{K}_s^- \mathbf{K}_s^-\right)} \hat{\boldsymbol{\epsilon}}^T \mathbf{K}_s^- \hat{\boldsymbol{\epsilon}}$$

where $\mathbf{K}_s^-$ is the block diagonal matrix that have all the study-specific kinship matrix (without their diagonal values) arranged diagonally, as

$$\mathbf{K}^s = \begin{pmatrix} \mathbf{K}_1^- & \mathbf{0} & \ldots & \ldots & \\ \mathbf{0} & \mathbf{K}_2^- & & & \mathbf{0} \\ \vdots & & \ddots & & \\ \vdots & \mathbf{0} & \mathbf{0} & & \mathbf{K}_S^- \end{pmatrix}$$

To see that this meta-analytic estimator of σ_k^2 is unbiased, note first that $\text{cov}(\hat{\boldsymbol{\epsilon}}) = (\sigma_e^2 + \sigma_k^2)\mathbf{I} + \sigma_k^2\mathbf{K}^-$, where now $\mathbf{K}^-$ is the kinship matrix with kinship coefficients between the

individuals across studies. Now, from characteristics of quadratic forms, we have that

$$\begin{aligned} E\left[\widehat{\sigma}_k^2\right] &= E\left[\frac{1}{\text{tr}\left(\mathbf{K}_s^- \mathbf{K}_s^-\right)} \hat{\boldsymbol{\epsilon}}^T \mathbf{K}_s^- \hat{\boldsymbol{\epsilon}}\right] = \frac{1}{\text{tr}\left(\mathbf{K}_s^- \mathbf{K}_s^-\right)} \text{tr}\left(\mathbf{K}_s^- \text{cov}(\hat{\boldsymbol{\epsilon}})\right) \\ &= \frac{1}{\text{tr}\left(\mathbf{K}_s^- \mathbf{K}_s^-\right)} \text{tr}\left(\mathbf{K}_s^- (\sigma_e^2 + \sigma_k^2)\mathbf{I} + \sigma_k^2 \mathbf{K}^-\right) \\ &= \frac{1}{\text{tr}\left(\mathbf{K}_s^- \mathbf{K}_s^-\right)} \text{tr}\left(\mathbf{K}_s^- \sigma_k^2 \mathbf{K}_s^-\right) = \sigma_k^2. \end{aligned}$$

Computing the meta-analytics heritability estimator and confidence intervals. The eigenvalues result shows that all we need to calculate heritability estimates and confidence intervals are eigenvalues of the matrix $\mathbf{K}^-$ (the kinship matrix without the diagonal), estimated σ_e^2, σ_k^2, and the sum of the entries of $\mathbf{K}^-$ ($2\sum_{i<j} k_{ij}^2$). This result could be used to extend our methods to meta-analysis of information from multiple studies. Suppose that each of m independent studies calculated the residuals from a "null model" (i.e. a regression model without genetic fixed effects other than PCs). Then, each study s reports:

1. $\mathcal{K}^s = 2\sum_{i<j} k_{ij}^2$,
2. $\widehat{\sigma}_{k,s}^2$,
3. $\widehat{\sigma}_{e,s}^2$,
4. The number of participants in the study n_s,
5. The eigenvalues $\lambda_1^s, \ldots, \lambda_{n_s}^s$ of the matrix $\mathbf{K}_s^-$.

Then, the meta-analysis estimates of the kinship and error variance components, and $\mathcal{K}^S$

are given by:

$$\begin{aligned}
\widehat{\sigma}^2_k &= \frac{\sum_{s=1}^{S} \mathcal{K}^s \widehat{\sigma}^2_{k,s}}{\sum_{s=1}^{S} \mathcal{K}^s} \\
\widehat{\sigma}^2_e &= \frac{\sum_{s=1}^{S} n_s \widehat{\sigma}^2_{e,s}}{\sum_{s=1}^{S} n_s}, \\
\mathcal{K}^S &= \sum_{s=1}^{S} \mathcal{K}^s,
\end{aligned}$$

and the eigenvalues of the cross-study $\mathbf{K}^-$ matrix are taken to be $\lambda^1_1, \ldots, \lambda^1_{n_1}, \ldots, \lambda^S_1, \ldots, \lambda^S_{n_S}$. Using these, the central location that can calculate heritability estimates and confidence intervals.

C The Hispanic Community Health Study/Study of Latinos

The HCHS/SOL, (LaVange et al., 2010; Sorlie et al., 2010)) is a community based cohort study, following self-identified Hispanic individuals from four field centers (Chicago, IL; Miami, FL; Bronx, NY; and San Diego, CA). Individuals were sampled via a two-stage sampling scheme, in which households were randomly sampled from sampled community block units. Almost 13,000 study participants consented for genotyping. HCHS/SOL individuals are classified into 'genetic analysis groups', classes that are based on self reported ethnicities and genetic similarity (Conomos et al., 2016). The genetic analysis groups are Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American. This study was approved by the institutional review boards at each field center, where all subjects gave written informed consent.

C.1 Genotyping, imputation and quality control

Blood samples from HCHS/SOL individuals were genotyped on a custom array consisting of Illumina Omni 2.5M content plus ∼150,000 custom markers selected to include ancestry-informative markers, variants characteristic of Amerindian populations, known GWAS hits and other candidate gene polymorphisms. Quality control was similar to the procedure described in Laurie et al. (2010) and included checks for sample identity, batch effects, missing call rate, chromosomal anomalies (Laurie et al., 2012), deviation from Hardy-Weinberg equilibrium, Mendelian errors, and duplicate sample discordance. 12,803 samples passed quality control, and 2,232,944 SNPs passed quality filters. Pairwise kinship coefficients and principal components reflecting ancestry were estimated in an iterative procedure which accounts for admixture (Conomos et al., 2016). All common variants were used to estimate kinship coefficients.

C.2 Heritability estimation in the HCHS/SOL

In each group of interest, including all individuals ('pooled' analysis), or specific genetic analysis groups, we randomly removed related individuals, to generate a set of individuals without any pair having kinship coefficient higher than 2^{-11}. Due to the sampling structure of the HCHS/SOL, the correlation between individuals is modeled in a kinship matrix, and also via matrices corresponding to community block units, and households. We estimated variance components via the procedure described here, with the three correlation matrices. We utilized the availability of environmental correlation to also estimate the contribution of modeled environmental factors (block unit and household) to the phenotypic variance.

Finally, we also demonstrate the use of our method for meta-analysis by removing individuals from shared household to generate a restricted set in which none of the individuals live in the same house, and used the proposed procedure for calculating heritability in meta-analysis. Note that for this purpose we neglected block unit correlation and assume that there is no correlation due to block unit sharing.

We estimated heritability for the FEV1 (a measure of lung function), standing height, depression score (CESD10, a sum of ten questionnaire items related to depression in the past few weeks of filling the form), SBP (systolic blood pressure), and dental caries, a count of tooth decays and cavities across all teeth of a participant. Finally, all regression models were adjusted (via the design matrix $\mathbf{W}$) to the 5 first principal components, study center, age, sex, and genetic analysis group (in the pooled models). For some traits we used additional covariates. Table 1 provides the various estimates and confidence intervals.

Analysis	n	Height	Depression score	SBP	Dental caries	FEV1
Full: Including environmentally correlated individuals						
Heritability	10,255	0.58 (0.47,0.69)	0.04 (0.00,0.10)	0.20 (0.12,0.27)	0.17 (0.09,0.25)	0.25 (0.17,0.33)
Environmental variance	10,255	0.10 (0.06,0.14)	0.06 (0.02,0.10)	0.02 (0.00,0.06)	0.06 (0.02,0.10)	0.04 (0.00,0.09)
Restricted: without environmentally correlated individuals (heritability only)						
Pooled	7,848	0.57 (0.45,0.69)	0.03 (0.00,0.11)	0.19 (0.11,0.28)	0.19 (0.09,0.28)	0.25 (0.16,0.34)
Central American	867	0.26 (0.00,0.88)	0.00 (0.00,0.62)	0.00 (0.00,0.56)	0.17 (0.00,0.75)	0.16 (0.00,0.75)
South American	544	0.48 (0.00,1.00)	0.00 (0.00,0.88)	0.93 (0.06,1.00)	0.00 (0.00,0.88)	0.57 (0.00,1.00)
Mexican	2,862	0.48 (0.28,0.72)	0.04 (0.00,0.22)	0.20 (0.02,0.41)	0.24 (0.06,0.44)	0.27 (0.06,0.50)
Puerto Rican	1,479	0.47 (0.12,0.88)	0.00 (0.00,0.38)	0.27 (0.00,0.62)	0.25 (0.00,0.62)	0.49 (0.12,0.88)
Cuban	1,370	0.83 (0.31,1.00)	0.00 (0.00,0.50)	0.23 (0.00,0.75)	0.06 (0.00,0.62)	0.09 (0.00,0.62)
Dominican	726	0.84 (0.06,1.00)	0.02 (0.00,0.75)	0.71 (0.00,1.00)	0.00 (0.00,0.75)	0.61 (0.00,1.00)
Meta-analysis	7,848	0.53 (0.38,0.72)	0.02 (0.00,0.16)	0.24 (0.09,0.41)	0.19 (0.03,0.34)	0.30 (0.16,0.47)

Table 1: Comparison of heritability estimates obtained in analysis of various subgroups of the HCHS/SOL. The top and bottom parts consider data sets with and without environmentally correlated individuals. When environmental correlation was present, the analysis included all participants. The heritability is $\widehat{\sigma}_k^2/\widehat{\sigma}_T^2$, and the environmental variance proportion is $(\widehat{\sigma}_h^2+\widehat{\sigma}_c^2)/\widehat{\sigma}_T^2$, for σ_h^2, σ_c^2 variance component corresponding to household and community sharing matrices. The models without environmentally correlated individuals had smaller sample sizes (because individuals who shared household were removed), and we compared heritability estimates from the pooled analysis of all individuals, the various ethnic subgroups, and their meta-analysis. SBP is systolic blood pressure. Dental caries is a measure of teeth damage. FEV1 is a measure of lung function. The distribution of the ratio between the appropriate quadratic forms was approximation using the saddlepoint approximation of Lieberman (1994).