UW Biostatistics Working Paper Series

11-13-2015

Meta-analysis of genome-wide association studies with correlated individuals: application to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

Tamar Sofer
University of Washington, tsofer@uw.edu

John R. Shaffer
University of Pittsburgh, jrs51@pitt.edu

Misa Graff
University of N Carolina, Chapel Hill, migraff@email.unc.edu

Qibin Qi
Albert Einstein College of Medicine, qibin.qi@einstein.yu.edu

Adrienne M. Stilp
University of Washington, amstilp@uw.edu

See next page for additional authors

Authors
Tamar Sofer, John R. Shaffer, Misa Graff, Qibin Qi, Adrienne M. Stilp, Stephanie M. Gogarten, Kari E. North, Carmen R. Isasi, Cathy C. Laurie, and Adam A. Szpiro

Meta-analysis of genome-wide association studies with correlated individuals: application to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

Tamar Sofer[1]*, John R. Shaffer[2], Misa Graff[3], Qibin Qi[4], Adrienne M. Stilp[1], Stephanie M. Gogarten[1], Kari E. North[3], Carmen R. Isasi[4], Cathy C. Laurie[1] and Adam A. Szpiro[1]

[1]Department of Biostatistics, University of Washington, Seattle, WA, United States of America

[2]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, United States of America

[3]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, United States of America

[4]Department of Epidemiology & Population Health, Department of Pediatrics, Albert Einstein College of Medicine, Bronx, NY, 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.

Abstract

Investigators often meta-analyze multiple genome-wide association studies (GWASs) to increase the power to detect associations of single nucleotide polymorphisms (SNPs) with a trait. Meta-analysis is also performed within a single cohort that is stratified by, e.g., sex or ancestry group. Having correlated individuals among the strata may complicate meta-analyses, limit power, and inflate Type 1 error. For example, in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), sources of correlation include genetic relatedness, shared household, and shared community. We propose a novel mixed-effect model for meta-analysis, "MetaCor", which accounts for correlation between stratum-specific effect estimates. Simulations show that MetaCor controls inflation better than alternatives such as ignoring the correlation between the strata or analyzing all strata together in a "pooled" GWAS, especially with different minor allele frequencies (MAF) between strata. We illustrate the benefits of MetaCor on two GWASs in the HCHS/SOL. Analysis of dental caries (tooth decay) stratified by ancestry group detected a genome-wide significant SNP (rs7791001, p-value $= 3.66 \times 10^{-8}$, compared to 4.67×10^{-7} in pooled), with different MAF between strata. Stratified analysis of BMI by ancestry group and sex reduced over-all inflation from $\lambda_{GC} = 1.050$ (pooled) to $\lambda_{GC} = 1.028$ (MetaCor). Furthermore, even after removing close relatives to obtain nearly uncorrelated strata, a naïve stratified analysis resulted in $\lambda_{GC} = 1.058$ compare to $\lambda_{GC} = 1.027$ for MetaCor.

KEY WORDS: Mixed models; Effect heterogeneity; Stratified analysis; Inflation.

1 Introduction

Investigators often analyze a GWAS according to multiple strata, defined by various covariates such as sex or genetic ancestry group. Usually, there is an interest in studying the genetic effects in each of the strata separately, and also in the combined effect on all individuals in the study. For instance, Landi et al. (2009) stratified a lung cancer GWAS by tumor histology and by smoking status; Hamza et al. (2011) conducted a stratified GWAS according to coffee-drinking habits, and Randall et al. (2013) stratified by sex. Results from stratified GWAS can be combined by meta-analysis, if the individuals within each stratum are independent of the individuals within the other strata. However, in some studies such as the HCHS/SOL (LaVange et al., 2010; Sorlie et al., 2010), various sources of correlation between individuals are present, for instance, correlations due to genetic relatedness (kinship), household sharing, and sampling block unit. In this study, it is likely that any stratification will result in correlated individuals between strata. In this paper we propose a method to test for association of Single Nucleotide Polymorphism (SNPs) with an outcome by combining information across strata when such correlations exist.

Various meta-analytic methods are used (Franke et al., 2010; Zeggini and Ioannidis, 2009; Lill et al., 2012) to combine information across several GWASs. Tests under the fixed and random effect assumptions on the SNP effect size, use a weighted sum of the effect estimate of interest. These methods use summary statistics such as the effect size estimates and their standard errors, rather than individual-level data. A key assumption made is that of independence, which is not met when individuals are correlated between strata. Lin and Sullivan (2009) proposed a meta-analysis method within the GEE framework for combining information across studies that share participants. They estimate the covariance between a single participant's contribution to a pair of studies

as the covariance between their score equations. This method does not allow for the fact that individuals are correlated between and within each stratum, with potentially more complex specification of the correlation (e.g. correlation due to genetic relatedness and household). Zhu et al. (2015) recently proposed a meta-analysis procedure for correlated traits via test statistics from multiple GWASs. The test statistics are used to calculate the correlation between traits. This method could be used to calculate the correlation between the same outcome across different strata. However, this estimated correlation will be fixed for all SNPs, while the correlations between the SNP effects may in fact vary. Further, this method cannot be used for a single SNP, since the correlation is evaluated from a large number of test statistics, e.g. from a GWAS. When all individual data are available, as in the case of a stratified analysis of a single study, it is desirable to obtain a more accurate model of the data.

In this manuscript we propose model-based tests that utilize test statistics from strata with correlated individuals between them. We specify a mixed effects model for decomposing the variance of an outcome, with random effects corresponding to multiple sources of correlation between and within strata, with stratum-specific variance components for the shared random effects. We use the effect-specific correlations and the estimated variance components to calculate covariances between all pairs of individuals in the study. We then calculate the covariances between the stratum-specific effect estimates. Wald tests are then readily obtained: a test of the weighted fixed effects meta-analysis estimator, a test for interaction, Cochrans's Q test for heterogeneity, etc.

In the presence of individual-level data, a potential alternative to stratification is a pooled analysis of the entire sample together, which includes all strata indicators, interactions between covariates and strata indicators, and a sophisticated variance model to allow for heterogeneous variance components due to errors and other factors. However, a stratified analysis is easier to communicate, individual-stratum estimates are readily

obtained, and more importantly, it is computationally simpler, both in the analyst level, and for large samples also in terms of computer memory usage and timing, as the sizes of the matrices involved in a stratified analysis are substantially smaller than those in a pooled analysis.

2 Methods

2.1 Model

Suppose that y_{ik} is the outcome, $\boldsymbol{x}_{ik}$ are the covariates, and g_{ik} is the allelic dosage of SNP g of individual i, $i = 1, \ldots, n, k = 1, \ldots, K$, a member of stratum k. There are n_k individuals in the k_{th} stratum, with $n_1 + \ldots + n_K = n$. Suppose further that there are $l = 1, \ldots, L$ sources of correlation. For instance, in the HCHS/SOL, participants were sampled from multiple block groups, some share household, and some are genetically related. Consider the model:

$$y_{ik} = \boldsymbol{x}_{ik}\boldsymbol{\beta}_k + g_{ik}\alpha_k + a_{k1}b_{1i} + \ldots + a_{kL}b_{Li} + \sigma_k\epsilon_{ik}, \tag{1}$$

where $\boldsymbol{\beta}_k, \alpha_k$ are the fixed effects in the kth stratum, $b_{1i}, \ldots, b_{Li}$ are the mean-zero random effects of individual i corresponding to the L sources of variation. We assume that $b_{li} \perp b_{l'i}, l \neq l', l, l' = 1, \ldots, L$, and $b_{li} \perp \epsilon_{ik}, k = 1, \ldots, K$, where $\perp$ denotes independence. Note that b_{li} is not stratum-specific while ϵ_{ik} is.

Our model further assumes that $\text{var}(b_{li}) = 1$ for every $l = 1, \ldots, L$, and $\text{var}(\epsilon_{ki}) = 1$. Stratum-specific variances are modeled via the variables $a_{k1}, \ldots a_{kL}$ and $\sigma_k, k = 1, \ldots, K$. Thus, for instance, if participant i is in stratum k, and participant i' is in stratum k', the following hold about their individual outcome variances and covariance:

$$\text{var}(y_{ik}) = a_{k1}^2 + \ldots + a_{kL}^2 + \sigma_k^2, \tag{2a}$$

$$\text{var}(y_{i'k'}) = a_{k'1}^2 + \ldots + a_{k'L}^2 + \sigma_{k'}^2, \quad \text{and,} \tag{2b}$$

$$\text{cov}(y_{ik}, y_{i'k'}) = a_{k1}a_{k'1}\text{cor}(b_{1i}, b_{1i'}) + \ldots + a_{kL}a_{k'L}\text{cor}(b_{Li}, b_{Li'}). \tag{2c}$$

For example, if the 1st source of correlation is household, and persons i and i' share a household, then $\text{cor}(b_{1i}, b_{1i'}) = 1$, and if the 2nd source of correlation is genetic relatedness, then $\text{cor}(b_{2i}, b_{2i'}) = 2\theta_{i.i'}$, where $\theta_{i,i'}$ is the probability that person i and person i' have a single allele identical by descent (IBD) at this SNP.

2.2 Estimating the covariance between stratum-specific SNP effects

Our main goal is to test for the effect of a SNP g on the outcome y. We can obtain estimates of $\alpha_1, \ldots, \alpha_K$, the SNP effects in each of the strata, using traditional mixed-effects models. The null hypothesis of interest is $\widehat{\boldsymbol{\alpha}} = (\alpha_1, \ldots, \alpha_K)^T$, e.g. $H_0 : \boldsymbol{\alpha} = \mathbf{0}_K$, where $\mathbf{0}_K$ is the vector of length K with all zero entries. Since the estimated effects are correlated with each other due to the correlations between the individuals, we estimate the correlations between them to obtain an estimate of the covariance matrix $\widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}})$.

2.2.1 An estimator of $\widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}})$

Let $\boldsymbol{\gamma}_k = (\boldsymbol{\beta}_k, \alpha_k)^T, k = 1, \ldots, K$, be the vector of fixed effects in stratum k. Suppose stratum-specific mixed-model were fitted and estimates of the variance components $a_{k1}, \ldots, a_{kL}, \sigma_k^2,\ k = 1, \ldots, K$ are available. For the kth stratum, let $\boldsymbol{b}_{lk}$ be the sub-vector of random effects corresponding to the lth source of correlation in the n_k individuals. Let $\mathbf{I}_{n_k}$ be the $n_k \times n_k$ dimensional identity matrix, $\mathbf{X}_k = ((\boldsymbol{x}_{k1}, g_{k1})^T, \ldots, (\boldsymbol{x}_{kn_k}, g_{kn_k})^T)^T$ the stratum design matrix, and $\boldsymbol{y}_k$ the n_k subvector of

outcomes. The stratum-specific outcome covariance matrix is estimated by

$$\widehat{\mathbf{V}}_k = \widehat{a}^2_{k1}\text{cor}(\boldsymbol{b}_{1k}) + \ldots \widehat{a}^2_{kL}\text{cor}(\boldsymbol{b}_{Lk}) + \widehat{\sigma}^2_k \mathbf{I}_{n_k},$$

and the estimator of $\boldsymbol{\gamma}_k$ is given by

$$\widehat{\boldsymbol{\gamma}}_k = \left(\mathbf{X}_k^T \widehat{\mathbf{V}}_k^{-1} \mathbf{X}_k\right)^{-1} \mathbf{X}_k^T \widehat{\mathbf{V}}_k^{-1} \boldsymbol{y}_k$$

We now incorporate the predicted covariances between any two individuals in the study, obtained via equation (2), in a formula for the covariance between $\widehat{\boldsymbol{\gamma}}_k$ and $\widehat{\boldsymbol{\gamma}}_{k'}$:

$$\widehat{\text{cov}}\left(\widehat{\boldsymbol{\gamma}}_k, \widehat{\boldsymbol{\gamma}}_{k'}\right) = \left(\mathbf{X}_k^T \widehat{\mathbf{V}}_k^{-1} \mathbf{X}_k\right)^{-1} \mathbf{X}_k^T \widehat{\mathbf{V}}_k^{-1} \widehat{\text{cov}}(\boldsymbol{y}_k, \boldsymbol{y}_{k'}) \widehat{\mathbf{V}}_{k'}^{-1} \mathbf{X}_{k'} \left(\mathbf{X}_{k'}^T \widehat{\mathbf{V}}_{k'}^{-1} \mathbf{X}_{k'}\right)^{-1}$$

For $k, k' = 1, \ldots, K$. Note that if $k = k'$, we have that $\widehat{\text{cov}}(\boldsymbol{y}_k, \boldsymbol{y}_{k'}) = \widehat{\text{cov}}(\boldsymbol{y}_k, \boldsymbol{y}_k) = \widehat{\mathbf{V}}_k$, and the usual estimator of $\widehat{\text{cov}}(\boldsymbol{\gamma}_k)$ is obtained. Finally, from the pair-wise estimators of $\text{cov}(\widehat{\boldsymbol{\gamma}}_k, \widehat{\boldsymbol{\gamma}}_{k'})$, we obtain the estimator $\widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}})$.

2.2.2 Tests

Having $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}})$, we can perform tests of $\widehat{\boldsymbol{\alpha}}$. Let $\mathbf{1}_K$ is the vector of length K with all entries equal to 1. Consider:

1. MetaCor$_1$. The standard inverse variance fixed effect estimator is given by:

$$\widehat{\alpha}_F = \frac{\sum_{k=1}^K w_k \widehat{\alpha}_k}{\sum_{j=1}^K w_j}$$

with $w_k = 1/\widehat{\text{var}}(\widehat{\alpha}_k)$. To obtain the standard error of this estimator, note that

$$\widehat{\alpha}_F = \mathbf{1}_K^T \text{diag}\left(\frac{w_1}{\sum_{j=1}^K w_j}, \ldots, \frac{w_K}{\sum_{j=1}^K w_j}\right) \widehat{\boldsymbol{\alpha}} \mathbf{1}_K, \text{ and thus,}$$

$$\widehat{\text{var}}(\widehat{\boldsymbol{\alpha}}_F) = \mathbf{1}_K^T \text{diag}\left(\frac{w_1}{\sum_{j=1}^K w_j}, \ldots, \frac{w_K}{\sum_{j=1}^K w_j}\right) \widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}}) \text{diag}\left(\frac{w_1}{\sum_{j=1}^K w_j}, \ldots, \frac{w_K}{\sum_{j=1}^K w_j}\right) \mathbf{1}_K.$$

Under the null ,$\widehat{\alpha}^2_F/\text{var}(\widehat{\alpha}_F)$ is distributed as a 1 degree-of-freedom (df) χ^2 variable.

2. MetaCor$_2$, that is based on the generalized least squares estimator, utilizes the correlations between the strata more efficiently:

$$\begin{aligned} \widehat{\alpha}_{gls} &= \frac{\mathbf{1}_K^T \left[\widehat{\mathrm{cov}}(\widehat{\alpha})\right]^{-1} \widehat{\boldsymbol{\alpha}}}{\mathbf{1}_K^T \left[\widehat{\mathrm{cov}}(\widehat{\alpha})\right]^{-1} \mathbf{1}_K} \qquad \text{with,} \\ \widehat{\mathrm{var}}(\widehat{\alpha}_{gls}) &= \left(\mathbf{1}_K^T \left[\widehat{\mathrm{cov}}(\widehat{\alpha})\right]^{-1} \mathbf{1}_K\right)^{-1}. \end{aligned}$$

3. The Cochran's Q test of heterogeneity is adapted to account for the covariances between the stratum-specific effects. The test statistic is given by

$$Q = \sum_{k=1}^{K} w_k (\hat{\alpha}_k - \hat{\alpha}_F)^2,$$

and it can be expressed as a quadratic form:

$$\begin{aligned} Q &= \sum_{k=1}^{K} w_k (\hat{\alpha}_k - \hat{\alpha}_F)^2 = \sum_{k=1}^{K} w_k \left(\hat{\alpha}_k - \frac{\sum_{i=1}^{K} w_i \widehat{\alpha}_i}{\sum_{j=1}^{K} w_j}\right)^2 \\ &= \sum_{k=1}^{K} w_k \alpha_i^2 - \frac{\left(\sum_{i=1}^{K} w_i \widehat{\alpha}_i\right)^2}{\sum_{j=1}^{K} w_j} = \widehat{\boldsymbol{\alpha}} \mathbf{A} \widehat{\boldsymbol{\alpha}}, \text{ with} \\ \mathbf{A} &= \left[\begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_K \end{pmatrix} - \frac{1}{K \sum_{k=1}^{K} w_k} \begin{pmatrix} w_1 & \dots & w_K \\ \vdots & \vdots & \vdots \\ w_1 & \dots & w_K \end{pmatrix} \begin{pmatrix} w_1 & \dots & w_1 \\ \vdots & \vdots & \vdots \\ w_K & \dots & w_K \end{pmatrix} \right]. \end{aligned}$$

Under the null of equal SNP effects across strata, it is distributed as the weighted sum $\sum_{k=1}^{K} \lambda_k \chi^2_{(1),k}$, with the weights $\lambda_1, \dots, \lambda_K$ being the eigenvalues of the matrix $\mathbf{A}\mathrm{cov}(\widehat{\boldsymbol{\alpha}})$ (Imhof, 1961) and the $\chi^2_{(1),k}$ being independent of each other.

It is also simple to obtain a test of interaction. For instance, if there are $P = K/2$ pairs of strata (males and females of a few ethnicities, say), the interaction effect may be a weighted sum of $(\widehat{\alpha}_{11} - \widehat{\alpha}_{12}), \dots, (\widehat{\alpha}_{P1} - \widehat{\alpha}_{P2})$. Other linear tests of the form $\mathbf{1}\mathbf{A}\widehat{\boldsymbol{\alpha}}$ for some $\mathbf{A}$ matrix and $\mathbf{1}$ vector of ones of an appropriate dimension, with variance $\mathbf{1}^T \mathbf{A} \widehat{\mathrm{cov}}(\widehat{\boldsymbol{\alpha}}) \mathbf{A} \mathbf{1}$ could be easily obtained.

2.2.3 Relationship with existing tests

Here we compare MetaCor with the following three tests.

4. Pooled, the estimator that does not stratify the analysis at all, i.e. estimates a single α parameter. This estimator usually has a mispecified model, unless all interaction terms between the covariates and strata indicators are specified, and a complex variance model is incorporated via stratum-specific variance components.

5. StraInd. The inverse variance fixed effect estimator given in (1), implemented on a reduced data set in which study participants are removed to create independent strata, and $\text{cov}(\boldsymbol{\alpha})$ is assumed to the $K \times K$ identity matrix.

6. The inverse variance fixed effect estimator can also be erroneously implemented under the (wrong) assumption that the strata are independent. We refer to this test as MetaNaive and it is identical to StratInd, but is implemented on a different sample set.

2.3 Computation

Computation of any test statistic begins with estimating the parameters $\boldsymbol{\alpha}$ and $\text{cov}(\widehat{\boldsymbol{\alpha}})$. We first describe their computation, and then refer to the test statistics.

2.3.1 Estimating SNP effects and their covariances

As is common in the mixed-effects based GWAS practice (Kang et al., 2010), we first estimate the variance components using only the "null model", i.e. a model with all covariates, principal components, and matrices modeling the correlations between the random effects such as genetic relatedness matrix (GRM), but without individual genotypes. These models are estimated separately in each stratum, and result in the estimators $\widehat{a}^2_{lk}$ for $l = 1, \ldots, L, k = 1, \ldots, K$, and $\hat{\sigma}^2_1, \ldots, \hat{\sigma}^2_K$. The estimated covariance matrices $\widehat{\mathbf{V}}_k$, their inverses, and $\widehat{\text{cov}}(\boldsymbol{y}_k, \boldsymbol{y}_{k'})$, for $k, k' = 1, \ldots, K$, are obtained by substituting the appropriate variance estimators. Then, multiple GWAS by stratum are

conducted jointly (not in parallel). Stratum-specific projection matrices are calculated:

$$\mathbf{P}_k = \left(\mathbf{X}_k^T \widehat{\mathbf{V}}_k^{-1} \mathbf{X}_k\right)^{-1} \mathbf{X}_k^T \widehat{\mathbf{V}}_k^{-1}$$

and are used to obtain the individual-stratum fixed effects $\widehat{\boldsymbol{\gamma}}_k = \mathbf{P}_k \boldsymbol{y}_k$, within-stratum covariance matrices $\widehat{\mathbf{V}}_k = \mathbf{P}_k \widehat{\text{var}}(\boldsymbol{y}_k) \mathbf{P}_k^T$, and between-strata covariance matrices $\mathbf{P}_k \widehat{\text{cov}}(\boldsymbol{y}_k, \boldsymbol{y}_{k'}) \mathbf{P}_{k'}$. Thus, the estimation procedure is equivalent to running the GWAS in each stratum separately, with the added computation of between-strata covariance matrices. Note that $\widehat{\text{cov}}(\boldsymbol{y}_k, \boldsymbol{y}_{k'})$ is calculated once and the computation time of $\mathbf{P}_k \widehat{\text{cov}}(\boldsymbol{y}_k, \boldsymbol{y}_{k'}) \mathbf{P}_{k'}$ is $O(n_k^2 n_{k'}^2)$. This time is in practice quite small. For instance, with a total of about 12,000 individuals, and over 2 million SNPs, in a stratified analysis of 6 strata we calculated effect estimates and standard errors across all strata, and all 21 covariances between the effect estimates in only 6 hours and 10 minutes, and in a stratified analysis (with the same individuals) of 2 strata we calculated effect estimates, standard errors and a single covariance between each of the two effect estimates in 6 hours and 50 minutes. Both analyses were performed on a single Intel® Xeon® E5-2630 CPU (2.40 GHz) core. The later analysis took slightly longer even though it had only two strata (while the first had six), since the computation time is quadratic in an individual stratum sample size, so that larger strata increase the computation time significantly.

2.3.2 Calculation of test statistics

The quickest test statistic to compute is MetaCor_1, since tens of millions of tests could be calculated at the same time in a matter of seconds using simple matrix operations. MetaCor_2 requires either inverting the matrix $\widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}})$, or computing the quantities $[\widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}})]^{-1}\widehat{\boldsymbol{\alpha}}$, $[\widehat{\text{cov}}(\widehat{\boldsymbol{\alpha}})]^{-1}\widehat{\mathbf{1}}$. We found that in analyzing up to six strata, we could compute MetaCor_2 for 100,000 SNPs together in a few seconds, by applying the recursive method for computing matrix inverses (using cofactors) on the estimated covariances

of the batch of SNPs jointly. However, this method is computer memory intensive, and for more strata it is better to compute MetaCor_2 for each SNP individually.

2.4 The HCHS/SOL data set

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). Households were randomly sampled from sampled block groups (two stage sampling), and of the sampled individuals, almost 13,000 people were genotyped. Some of these individuals are from the same block group, some live in the same house, and some are genetically related. Thus, there are three sources of correlation corresponding to block group, household, and relatedness.

2.5 Genotyping, kinship estimation and definition of genetic analysis groups

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 and were informative (unique and polymorphic). Pairwise kinship coefficients and principal components reflecting ancestry were estimated in an iterative procedure that accounts for admixture (Conomos, 2014).

Individuals in HCHS/SOL were classified into six "genetic analysis groups" (Cuban,

Dominican, Puerto Rican, Mexican, Central American, or South American) based on their self-reported background and position in the n-dimensional space defined by the first 5 genetic principal components (M.P.C, C.A.L et al., unpublished).

We first investigate in simulations the effects of heterogeneity of MAFs and phenotypic variances on the various estimators. We then investigate the effect of stratification and the different estimators on BMI and dental caries (tooth decay) analyses.

2.6 Simulation study

We investigated the properties of MetaCor_1 and MetaCor_2 under a few simulation settings in which individuals within strata are more similar to each other than individuals between strata. We compared them to the alternatives StratInd, MetaNaive, and Pooled. Here, Pooled misspecifies the variance model, but it correctly specifies the mean model, by including all interaction terms between covariates and strata indicators. We compared power for the test of the marginal SNP effect on the total population. We simulated two-strata scenarios. We generated correlation between strata due to genetic relatedness and due to environment (to be described henceforth), and investigated the effect of differences in the SNP effect and MAF between the strata. To assess power and estimation accuracy, we generated 5,000 simulations, each with 10 SNPs, from various combinations of the simulation parameters, and 200,000 simulations for type 1 error, also each with 10 (null) SNPs. In both power and type 1 error simulations, we use a p-value threshold of 0.001, balancing the very low significance threshold employed in actual GWASs with the need to limit computational burden in a simulation study.

2.6.1 Setting the strata and correlation structure

In all simulations we generated a population of 6,000 individuals from 2,000 simulated "families" of three members in each, about half of which were called females and the

rest were called males, for simplicity. "Sex" (strata) indicator was sampled at random, and each of the simulations had the same breakdown of family membership: of the 2,000 families, 729 had members from only a single stratum, and 1,271 families had 2 members from one stratum, and one member from the second. Therefore, to create independent strata, we removed 1,271 individuals from the analysis.

2.6.2 Other simulation parameters

In each simulation, we generated 10 SNPs (independent of each other) with either null effect ($\alpha_m = \alpha_f = 0$) to study type 1 error, or non-null effect $\alpha_m = 0.4$, to study power. Here α_m is the SNP effect in the males. The SNP effect in the females was either $\alpha_f = \alpha_m$, or $\alpha_f = 1.2 \times \alpha_m$. SNPs were sampled from a binomial distribution according to a set MAF. The baseline MAF was 0.3, while in some settings we changed the MAF between strata so that in one stratum the MAF was 0.2, and 0.4 in the other. Note that although it may be uncommon to have the MAF differ between males and females, we consider this scenario since in some cases of stratified analysis, e.g. by ethnicity, individuals from different ethnicities may live together (be from the same family), while their different ethnicities are likely to have different MAF in many SNPs. The two strata have different error and random effect variances: let $\sigma^2_{err,st}, a^2_{fam,st}$ be the variances of the error and family variance components in strata $st \in \{m, f\}$. In general, males had larger variance in our simulations, with $a^2_{err,m} = 50, \sigma^2_{fam,m} = 31$ and $a^2_{err,f} = 16, \sigma^2_{fam,f} = 10$. The errors were generated from normal distributions with mean of zero and variance of one independently for each of the individuals, and were then multiplied by the strata-specific standard deviations. The random effects were also generated from a mean zero, unit variance, normal distribution, but they were not independent, but rather entire families had the same random effects. The random effect associated with each individual was multiplied by the strata-specific standard error of

this individual. In what we described so far there is no genetic similarity between members of the same family, but rather only environmental similarity, modeled via the random effects alone. In some simulations we also generated genetic similarity, in a rather simplistic way: the three members of the same family had the same allelic dosages.

Finally, the outcome of a female indexed by i from family l was:

$$y_i = 2 + x_{i1} + x_{i2} + \alpha_f g_i + a_{fam,f} b_l + \sigma_{err,f} \epsilon_i$$

and of a male indexed by j from family l:

$$y_j = 2.5 + 1.5 x_{j1} + x_{j1}^2 + 0.7 x_2 + \alpha_m g_j + a_{fam,m} b_l + \sigma_{err,m} \epsilon_j$$

Note that the Pooled had additional interaction terms between the strata indicator and the covariates $I_{(\text{sex}_i=\text{male})} \times X_1, I_{(\text{sex}_i=\text{male})} \times X_1^2$, and $I_{(\text{sex}_i=\text{male})} \times X_2$.

3 Results

3.1 Simulation studies

Table I presents type 1 error and power estimates for the various simulations scenarios for all compared estimators. As expected, stratification is in general beneficial when there are differences between strata (here different phenotypic variances between strata in all scenarios). The stratified estimators MetaCor_1, MetaCor_2 and StratInd all protected type 1 error, as expected. However, MetaNaive was inflated when there was genetic similarity within families, while it was not inflated otherwise. Pooled had correct type 1 error when the MAF was the same in both strata, but otherwise its type 1 error was either inflated or deflated. Of the stratified estimators, StratInd was the least powerful, not surprisingly, as it uses a smaller number of the study participants to obtain independent strata. MetaCor_1, MetaCor_2, and MetaNaive all performed almost

identically when correlation between individuals was solely environmental. However, when individuals from the same family had the same genotypes, MetaCor_2 was slightly more powerful than MetaCor_1, and MetaNaive was inflated.

3.2 Data analysis of stratified GWAS in the HCHS/SOL

3.2.1 Analysis of BMI

There are 12,705 HCHS/SOL individuals available for BMI analysis. We compared the Pooled analysis that did not stratify to MetaCor_2 under various stratification schemes: by sex, by genetic analysis group, and by both sex and genetic analysis group. In all analyses, the outcome was log-transformed to approximate a normal distribution. We adjusted for age via linear and quadratic terms, the first five principal components estimated from the combined data set, and also sex (for analyses that were not stratified by sex), and genetic analysis group (for analyses that were not stratified by genetic analysis group).

Figure 1 presents the estimated variance components associated with the error variance, household, and kinship, for each of the genetic analysis groups, sex strata, and the pooled analysis that estimated the variances for all participants jointly. The number of participants in each of the presented groups is also provided. The top panel provides the absolute values of the estimated variance components, together with a 95% confidence intervals, based on normal asymptotic distribution of the estimates. The bottom panel provides the estimated proportion of the total variance, attributed to each of the variance components. Note that the proportion of variance due to kinship could be interpreted as narrow-sense heritability, if close relatives are excluded when variance components are estimated (Yang et al., 2010). The absolute values of variances differed somewhat between both genetic analysis and sex groups, with the largest differences observed in the error variance.

We studied the control of inflation via the inflation factor λ_{GC} (Yang et al., 2011). Throughout, all inflation factors were calculated over the autosomal SNPs with more than 30 counts of the minor allele (MAC) across all participants. Part (a) of Table II compares the inflation factors obtained from the pooled and stratified analyses. Indeed, for the Pooled BMI analysis, which analyzes all individuals together, assuming common fixed effects and variance components had moderate inflation ($\lambda_{GC} = 1.05$). λ_{GC} decreased with stratification, with the largest reduction seen upon stratifying by genetic analysis group. This is probably due to differences in MAF between genetic analysis groups in SNPs associated with BMI.

We also considered the estimators compared in the simulations in analyses stratified by genetic analysis group. Our goals here were: to see that the data analysis was consistent with the simulation results; to check whether MetaCor_2 was beneficial compared to the computationally simpler MetaCor_1; and to study the feasibility of generating six independent strata of genetic analysis groups by removing individuals, and seeing if their analysis using StratInd yielded similar results to MetaCor_2. Part (b) of Table II provides the inflation factors for both sex and genetic stratified analysis for MetaCor_2, MetaCor_1, MetaNaive, and StratInd. To implement StratInd, we generated 12 genetic strata with low correlations, by restricting the data set to 9,029 individuals such that any genetic group did not have a person living in the same household with someone, or a relative of up to 3^{rd} degree, from another genetic group. We called this reduced data set "Distant". MetaCor_1 and MetaCor_2 produced very similar λ_{GC}s. Indeed, they had very similar results overall. As expected, MetaNaive, which assumes that the strata are independent, was highly inflated with $\lambda_{GC} = 1.088$. Surprisingly, applying StratInd on the Distand data set that has only low correlations between the strata, i.e. only due to shared community (city block unit) and distant relatedness, also gave inflated results ($\lambda_{GC} = 1.058$). We hypothesized that distant relatives, of degree 4^{th} and higher, are

responsible for this inflation, and applied MetaCor_2 to the reduced data set, to account for relatedness between individuals in the different strata. For MetaCor_2, inflation was reduced ($\lambda_{GC} = 1.027$).

There was only a single locus of genome-wide significant SNPs associated with BMI, of SNPs in the well-known FTO gene (Speliotes et al., 2010). This association remained significant in all analyses. Manhattan and q-q plots comparing these four analyses are found in the supplementary material.

3.2.2 Analysis of dental caries

We analyzed the commonly used index of dental caries, DMFS, which corresponds to the count of the number of Decayed, Missing, and Filled (i.e., restored) tooth Surfaces across the permanent dentition. Analyses were adjusted for age, the first five principal components, sex, genetic analysis group, and smoking status (past, current, or former smoker). For this trait, the inflation factor in the pooled analysis was relatively low, with $\lambda_{GC} = 1.018$ over all genotyped SNPs with minor allele count across all participants being at least 30. Still, we considered stratification by smoking status (ever versus never smoker), by genetic analysis group, and by both. Part (a) in Table II provides the inflation factors from all analyses. While stratification by smoking status alone did not result in reduction in λ_{GC}, stratification by genetic analysis group, as well as by both genetic analysis group and smoking status, reduced λ_{GC}. The low inflation factor in the analysis stratified by both genetic analysis group and sex, may indicate potential over-adjustment. We also compared the various meta-analytic estimators MetaCor_2, MetaCor_1, MetaNaive, and StratInd on the genetic analysis group-stratified analysis, with the complete and Distant data sets. The conclusions were similar to those in the BMI analysis, though in general the inflation was much lower. Interestingly, the inflation was a bit lower when applying MetaCor_2 on the complete data set ($\lambda_{GC} = 0.992$),

compared to the Distant data set ($\lambda_{GC} = 0.997$). Since the difference was very small, it may be just a random variation from having a slightly different data set.

For this trait, however, it is more interesting to focus on the top, and only genome-wide significant association, that was detected in the stratified analysis, but was not genome-wide significant in the pooled analysis. Figure 2 provides the forest plot comparing the results for SNP rs7791001 for the pooled analysis, and the analyses stratified by smoking status and by genetic analysis group. The SNP effect was genome-wide significant only when the analysis was stratified by genetic analysis group (p-value= 3.66×10^{-8}, MetaCor_2, while p-value= 4.67×10^{-7} in Pooled). The effect allele frequencies (EAF) vary somewhat between genetic analysis groups, and the EAF is especially smaller (larger MAF) among Dominicans, who also have the largest estimated effect size. This is consistent with the simulation results wherein the most dramatic improvement in power for MetaCor compare to Pooled occurred when the MAF was larger in the group with the larger effect size. We omit the analysis stratified by both genetic analysis group and smoking status from this figure for clarity. (The p-value for rs7791001 was below genome-wide significance (p-values= 8.98×10^{-7}), possibly due to the over-adjustment (consider the deflated λ_{GC} value observed in Table II) or random variation). Manhattan plots and q-q plots for all the different analyses, as well as figures comparing the variance components and the fixed effects across strata of genetic group, smoking status, and in the pooled analysis are provided in the supplementary material.

4 Discussion

In this manuscript, we propose estimators to meta-analyze multiple GWASs with correlated individuals. The proposed test statistics MetaCor_1 and MetaCor_2 account for correlations between individuals within- and between-studies or strata, they control type 1 error and they are more powerful than existing approaches that try to simplify

the data by either removing individuals or ignoring some of the existing correlations between study individuals. Our simulation studies demonstrate that stratification is useful when the main regression model, including phenotypic variances, and MAFs differ between strata defined by a specific variable, such as genetic analysis group. Specifically, when MAF differed between strata, the pooled estimator that analyses all participants together had sometimes inflated and sometimes deflated type 1 error, and usually lower power. For example, in the analysis of dental caries, the analysis that stratified by genetic analysis group detected a genome-wide significant SNP (p-value= 3.66×10^{-8}, MetaCor_2) while the pooled test p-value did not pass the established threshold (p-value= 4.67×10^{-7}). For this SNP, one of the strata had higher MAF and lower residual variance than the rest of the strata. In the analysis of BMI, a trait that is well-known to have different distributions among different ethnicities and sex, a pooled analysis had $\lambda_{GC} = 1.05$, while an analysis that stratified by both genetic analysis group and sex and accounted for correlation between the strata had $\lambda_{GC} = 1.028$. Such stratified analyses could not be achieved without MetaCor. A naïve stratified analysis, that ignored the correlation between the strata, had $\lambda_{GC} = 1.08$.

We provide two estimators and tests for the effect of SNP on the outcome combining multiple strata: MetaCor_1 and MetaCor_2. Although MetaCor_2 is theoretically more efficient, as it uses the correlation between the strata, down-weighting contributions of highly correlated strata, in practice it was almost identical to MetaCor_1, which is computationally simpler. In studies with higher degree of relatedness between the participants MetaCor_2 will be advantageous so we recommend its use.

We simulated environmental relatedness via correlated residuals, and genetic relatedness in a rather simplistic manner: in our simulated families of three individuals, all members had the same allele count. This simplified scenario helped us gain insight into the advantage of MetaCor_2 compared to MetaCor_1, and the cause of inflation in

meta-analyzing strata with related individuals, while ignoring relatedness (MetaNaive). Note that MetaNaive was not inflated when there was only environmental association between the strata, rather, only when there was genetic relatedness. This is because the environmental association was independent of the simulated genotypes. Our analysis demonstrated that for some traits (for example, BMI), distant relatedness of 4^{th} degree may contribute to inflation when meta-analyzing multiple studies, and accounting for this correlation reduces inflation. This may indicate that large meta-analyses performed in the past were inflated due to distant relatedness. However, it is not easy to account for relatedness between two studies when individual-level data are not available.

Our model assumes that the correlation structures of the random effects (e.g. correlations due to kinship across all individuals, etc.) are independent of strata, i.e. they depend only on the relationship between the individuals. The covariances between the random effects of any pair of individuals do vary by strata assignments, as they depend on stratum-specific variance components. One can argue that the model should allow for a more general correlation model, in which the correlations differ between the strata, for instance, setting the correlation between the random effects of two females living in the same household to be different than the correlation between the random effects of a male and a female living in the same household. Such a model will include additional parameters and will be more computationally intensive; however as was seen in simulations, misspecification of the variance in the mixed model did not dramatically inflate the type 1 error, if at all. Therefore, we believe that our model well balances model simplicity and computational demands. Furthermore, in an era of increasing sample sizes, stratifying studies to smaller sets and combining the results using MetaCor would be a computationally convenient alternative to a pooled analysis, as large matrices (e.g. a squared matrix of 20,000 rows and columns) will be difficult to compute for all but the most powerful hardware.

This work can be extended in a few ways. First, the proposed tests are under the fixed-effects framework, and it will be useful to develop the random-effects model for the SNP effect on the outcome for these settings. Second, the presented model applies to continuous outcomes. It is a topic of future work to extend this model to generalized linear models, and especially binary traits, which are commonly investigated in GWAS.

Acknowledgements

The authors thank the staff and participants of HCHS/SOL for their important contributions. The authors are thankful to Bruce Weir for reviewing the manuscript and providing helpful comments. 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 Institute 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

Conomos, M. P. (2014). *Inferring, estimating and accounting for population and pedigree structure in genetic analyses*. Ph.D. thesis, University of Washington, Seattle.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R. et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, **42** 1118–1125.

Hamza, T. H., Chen, H., Hill-Burns, E. M., Rhodes, S. L., Montimurro, J.,

Kay, D. M., Tenesa, A., Kusel, V. I., Sheehan, P., Eaaswarkhanth, M. et al. (2011). Genome-wide gene-environment study identifies glutamate receptor gene grin2a as a parkinson's disease modifier gene via interaction with coffee. *PLoS genetics*, **7** e1002237.

Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* 419–426.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C. and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, **42** 348–354.

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M. et al. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *The American Journal of Human Genetics*, **85** 679–691.

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.

Lill, C. M., Roehr, J. T., McQueen, M. B., Kavvoura, F. K., Bagade, S., Schjeide, B.-M. M., Schjeide, L. M., Meissner, E., Zauft, U., Allen, N. C. et al. (2012). Comprehensive research synopsis and systematic meta-analyses in parkinson's disease genetics: The pdgene database. *PLoS genetics*, **8** e1002548.

Lin, D.-Y. and Sullivan, P. F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. *The American Journal of Human Genetics*, **85** 862–872.

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., Kilpeläinen, T. O., Esko, T., Mägi, R., Li, S. et al. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS genetics*, **9** e1003500.

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.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R. et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, **42** 937–948.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, **42** 565–569.

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'Connell, J. R., Mangino, M., Magi, R., Madden, P. A., Heath, A. C., Nyholt, D. R., Martin, N. G., Montgomery, G. W., Frayling, T. M., Hirschhorn, J. N., McCarthy, M. I., Goddard, M. E. and Visscher, P. M. (2011). Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet*, **19** 807–812.

Zeggini, E. and Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, **10** 191–201.

Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., Smith, J. A., Yanek, L. R., Sun, Y. V., Edwards, T. L. et al. (2015). Meta-analysis of correlated traits via summary statistics from gwass with an application in hypertension. *The American Journal of Human Genetics*, **96** 21–36.

	simulation parameters							
genetic similarity	Yes	Yes	No	No	No	No	No	No
MAF-females	0.3	0.3	0.3	0.3	0.4	0.4	0.2	0.2
MAF-males	0.3	0.3	0.3	0.3	0.2	0.2	0.4	0.4
α_f	α_m	$1.2 \times \alpha_m$	α_m	$1.2 \times \alpha_m$	α_m	$1.2 \times \alpha_m$	α_m	$1.2 \times \alpha_m$
Test	Type 1 error							
MetaCor_1	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
MetaCor_2	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
MetaNaive	0.0036	0.0036	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
StratInd	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
Pooled	0.0010	0.0010	0.0010	0.0010	0.0005	0.0005	0.0017	0.0017
	Power							
MetaCor_1	0.18	0.28	0.56	0.75	0.60	0.79	0.46	0.63
MetaCor_2	0.19	0.32	0.56	0.75	0.59	0.79	0.46	0.63
MetaNaive	—	—	0.56	0.75	0.59	0.79	0.46	0.63
StratInd	0.15	0.25	0.43	0.62	0.47	0.66	0.34	0.49
Pooled	0.12	0.16	0.4	0.52	0.37	0.52	—	—

Table I: Simulation results. Type 1 error and power simulation results averaged over 5,000 (power) and 200,000 (type 1 error) simulations, comparing the estimators MetaCor_1, MetaCor_2, MetaNaive, StratInd and Pooled. A SNP passed testing if its p-value was lower than the threshold (here 0.001). Powers were omitted in the instances where the respective test did not protect type 1 error.

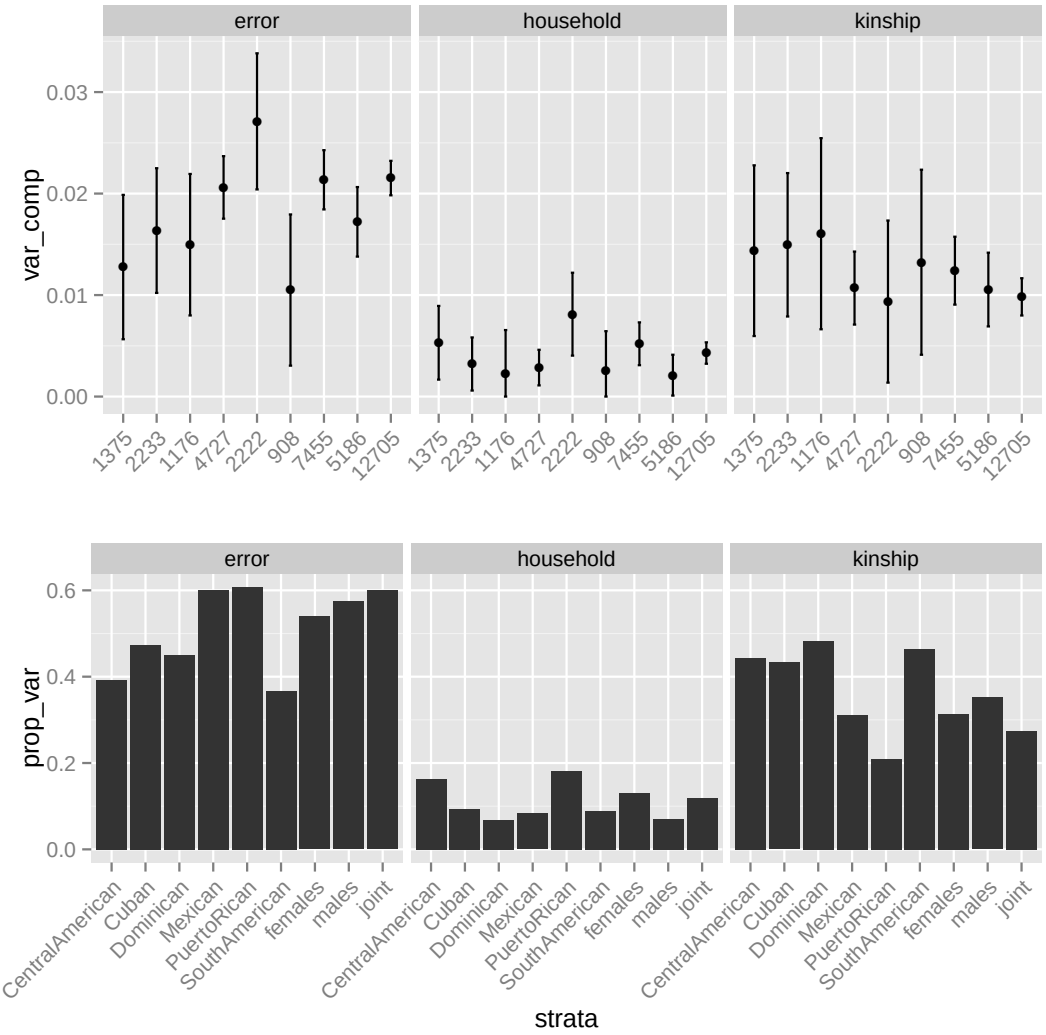

Figure 1: The top panel provides estimated variance component, along with 95% confidence intervals based on normal approximation, estimated for the genetic groups, males, females, and jointly over all participants in the pooled analysis ('joint'). The X-labels provide the sample sizes. The bottom panel provides the proportion of the estimated variances out of the total variances. The presented variance components correspond to the error variances, and variances due to household and kinship. Estimated variances due to block group are not presented, since they were always relatively small.

Part	Data set	BMI			NDMF		
		Stratification	Analysis	λ_{GC}	Stratification	Analysis	λ_{GC}
(a) Compare stratification schemes	Complete	none	Pooled	1.050	none	Pooled	1.018
	Complete	sex	MetaCor_2	1.048	smoke	MetaCor_2	1.019
	Complete	gengrp	MetaCor_2	1.034	gengrp	MetaCor_2	0.992
	Complete	sex gengrp	MetaCor_2	1.028	smoke gengrp	MetaCor_2	0.983
(b) Compare meta-analytic tests	Complete	sex gengrp	MetaCor_2	1.028	gengrp	MetaCor_2	0.992
	Complete	sex gengrp	MetaCor_1	1.028	gengrp	MetaCor_1	0.992
	Complete	sex gengrp	MetaNaive	1.088	gengrp	MetaNaive	1.018
	Distant	sex gengrp	StratInd	1.058	gengrp	StratInd	1.020
	Distant	sex gengrp	MetaCor_2	1.027	gengrp	MetaCor_2	0.997

Table II: Observed inflation factors in the analyses of log(BMI) and dental caries. In part (a), the tables provides the inflation factor λ_{GC} under various stratification schemes: when stratification is not performed (Pooled), by sex, genetic analysis group, and both (BMI), and by smoking status (ever vs. never smoker), genetic analysis group, and both (dental caries). The test used here is the recommended test statistic MetaCor_2. Part (b) compares a few potential methods to meta-analyze a stratified analysis. Comparisons of meta-analytic tests were performed on the stratification schemes that were determined as most appropriate in part (a), i.e. that their inflation factor was closest to 1. Both MetaNaive and StratInd assume that there are no correlations between the strata. However, MetaNaive is applied on the complete data set, and StratInd was applied on a reduced data set, here called "Distant", that removed about 1,000 individuals to obtain nearly-independent strata. In all instances, λ_{GC} was calculated over the genotyped autosomal SNPs with minor allele count larger than 30 in the pooled data set.

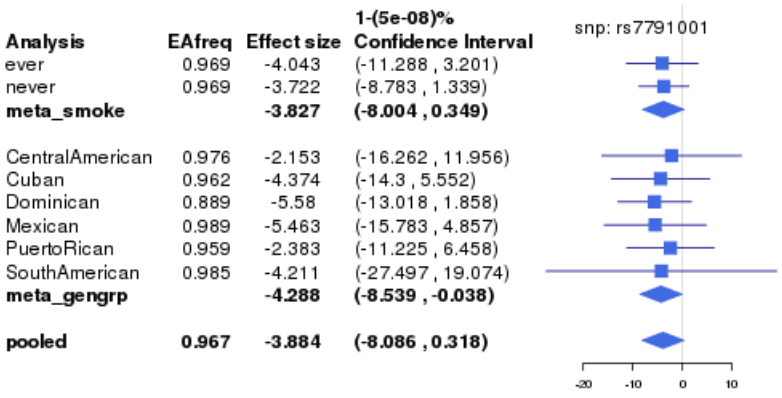

Figure 2: Forest plot comparing the top genotyped SNP from the genome-wide significant locus bound in the dental caries (DMFS index) analysis. P-values for this SNP from the various estimators are 4.67×10^{-7} (pooled), 3.66×10^{-8} (stratified by genetic group, MetaCor_2), and 8.98×10^{-7} (stratified by smoking status, MetaCor_2).