Fast estimation of genetic correlation for Biobank-scale data

Yue Wu[1], Anna Yaschenko[3], Mohammadreza Hajy Heydary[4], and Sriram Sankararaman[*1,2,5]

[1]Department of Computer Science, UCLA
[2]Department of Human Genetics, UCLA
[3]Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County
[4]Department of Computer Science, California State University, Fullerton
[5]Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California

Abstract

Genetic correlation, *i.e.*, the proportion of phenotypic correlation across a pair of traits that can be explained by genetic variation, is an important parameter in efforts to understand the relationships among complex traits. The observation of substantial genetic correlation across a pair of traits, can provide insights into shared genetic pathways as well as providing a starting point to investigate causal relationships. Attempts to estimate genetic correlations among complex phenotypes attributable to genome-wide SNP variation data have motivated the analysis of large datasets as well as the development of sophisticated methods.

Bi-variate Linear Mixed Models (LMMs) have emerged as a key tool to estimate genetic correlation from datasets where individual genotypes and traits are measured. The bi-variate LMM jointly models the effect sizes of a given SNP on each of the pair of traits being analyzed. The parameters of the bi-variate LMM, *i.e.*, the variance components, are related to the heritability of each trait as well as correlation across traits attributable to genotyped SNPs. However, inference in bi-variate LMMs, typically achieved by maximizing the likelihood, poses serious computational challenges.

We propose, RG-Cor, a scalable randomized Method-of-Moments (MoM) estimator of genetic correlations in bi-variate LMMs. RG-Cor leverages the structure of genotype data to obtain runtimes that scale sub-linearly with the number of individuals in the input dataset (assuming the number of SNPs is held constant). We perform extensive simulations to validate the accuracy and scalability of RG-Cor. RG-Cor can compute the genetic correlations on the UK biobank dataset consisting of $430,000$ individuals and $460,000$ SNPs in 3 hours on a stand-alone compute machine.

1 Introduction

Understanding the underlying shared genetic structure between traits and diseases can provide insights into shared disease etiology and can form the starting point to investigate causal relationships among traits [2]. Genetic correlation *i.e.*, the proportion of phenotypic correlation across a pair of traits that can be explained by genetic variation, is an important parameter in efforts to quantifying

*Corresponding author: sriram@cs.ucla.edu

the relationships among complex traits as it can provide insights into biological pathways that are shared among the pair of traits. For example, significant genetic correlation between body mass index (BMI) and lymphocyte count have been used to conclude that lymphocytes are relevant to body weight regulation [1]. Similarly, a number of studies have reported a high genetic correlation between schizophrenia and bipolar disorder [6, 2].

While traditionally reliant on family studies, the availability of genome-wide genetic data have led to a number of approaches to estimate genetic correlation from these datasets. Bi-variate Linear Mixed Models (LMMs) have emerged as a key statistical model for this problem [13]. The bi-variate LMM jointly models the effect sizes of a given SNP on each of the pair of traits being analyzed. The parameters of the bi-variate LMMs, *i.e.*, the variance components, are related to the heritability of each trait well as the genetic correlation across the traits.

The most commonly used method for estimating genetic correlation as well as trait heritabilities in a bi-variate LMM relies on the restricted maximize likelihood method, termed genomic restricted maximum likelihood (GREML)[5, 3, 8, 7]. However, GREML poses serious computational burdens. GREML is a non-convex optimization problem that relies on an iterative optimization algorithm. While a number of methods have been proposed to improve the computational efficiency of GREML [5], current GREML methods are still computationally expensive when applied to large-scale datasets such as the UK Biobank that contains genotypes from around half a million individuals at a million SNPs [11].

Another state-of-the-art method, LD-score regression (LDSC), requires only summary statistics from genome-wide association studies (GWAS) to estimate genetic correlations [2]. LDSC is appealing as it does not require individual level data thereby mitigating concerns of privacy that arise from sharing individual-level data. Further, LDSC often has substantially reduced computational requirements (assuming that the summary statistics have been computed). Nevertheless, LDSC has some drawbacks: its estimates tend to have large standard errors and is prone to bias in settings where there is a mismatch between the samples used to estimate summary statistics and the reference datasets that are used to estimate LD [9].

1.1 Our Contribution

We propose, RG-Cor, a randomized algorithm to estimate genetic correlations of traits using individual-level genotype that can scale to the dataset sizes typical of the UK Biobank. This method for estimating genetic correlation builds upon our randomized estimator of heritability, [12]. RG-Cor is a randomized Method-of-Moment(MoM) estimator of the heritability of traits as well as the genetic correlation between pairs of traits. MoM estimators tends to be less statistically efficient comparing to GREML. Despite the statistical inefficiency, the MoM estimator leads to a closed-form solution of heritability and genetic correlation parameters. On the other hand, the main computational bottleneck of the MoM estimator in genetic correlation estimation is the computation of the $N \times N$ genetic relationship matrix, which capture the relationships between all pairs of N individuals in the dataset.

For genetic correlation estimation, our randomized MoM estimator (RG-Cor) relies on the observation that the key computation bottleneck can be replaced by multiplying the $N \times M$ (individuals $\times$ SNP) genotype matrix with a small number, B, of random vectors thereby obtaining a time complexity of $\mathcal{O}(NMB)$. We can further gain efficiency by leveraging the structure of the genotype matrix, where all the entries are in a finite set, $\{0, 1, 2\}$ so that the time complexity can be reduced to $\mathcal{O}(\frac{NMB}{max(log_3(N), log_3(M))})$.

We apply RG-Cor to pairs of traits to estimate their heritability as well as the genetic correlation, as well as computing estimates of the standard errors. We show in simulations that the RG-Cor

yields accurate estimates of genetic correlation. Compared to GREML estimators, we show that the loss in statistical inefficiency of RG-Cor is fairly modest. On the other hand, RG-Cor is several orders of magnitude faster than other methods. Finally, we applied RG-Cor to compute the genetic correlation of selected pairs of traits in $291,273$ white British individuals in the UK Biobank.

2 Methods

2.1 Model assuming complete overlap of samples across traits

For simplicity, we first assume that we observe two traits measured on the same set of N samples. We observe genotypes across these N individuals at M SNPs. The genotype vector for individual n is a length M vector denoted by $\boldsymbol{g}_n \in \{0,1,2\}^M$. The mth entry of $\boldsymbol{g}_n$ denotes the number of minor alleles carried by individual n at SNP m. Let $\boldsymbol{G}$ be the $N \times M$ matrix of genotypes. Let $\boldsymbol{X}$ denote the $N \times M$ matrix of standardized genotypes obtained by centering and scaling each column of $\boldsymbol{G}$ so that $\sum_n x_{n,m} = 0$ and $\frac{1}{N}\sum_n x_{n,m}^2 = 1$ for all $n \in \{1,\ldots,N\}$. Let $\boldsymbol{y}_1, \boldsymbol{y}_2$ denote two vectors of phenotypes of length N.

We assume the vector of phenotypes $\boldsymbol{y}_1, \boldsymbol{y}_2$ is related to the genotypes $\boldsymbol{X}$ by a bivariate linear mixed model:

$$
\begin{aligned}
\boldsymbol{y}_1|\boldsymbol{\epsilon}_1, \boldsymbol{\beta}_1 &= \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\
\boldsymbol{y}_2|\boldsymbol{\epsilon}_2, \boldsymbol{\beta}_2 &= \boldsymbol{X}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2 \\
\epsilon_{1n}, \epsilon_{2n} &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \frac{\sigma_{e1}^2}{N} & \frac{\gamma_e}{N} \\ \frac{\gamma_e}{N} & \frac{\sigma_{e2}^2}{N} \end{bmatrix}) \\
\beta_{1m}, \beta_{2m} &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \frac{\sigma_{g1}^2}{M} & \frac{\gamma_g}{M} \\ \frac{\gamma_g}{M} & \frac{\sigma_{g2}^2}{M} \end{bmatrix})
\end{aligned}
$$

Here $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ denote the M-vectors of SNP effect sizes, *i.e.*, $\beta_{1,m}$ denotes the mean change in phenotype 1 when the genotype at SNP m changes from 0 to 1 or from 1 to 2.

Here phenotypes $\boldsymbol{y}_1, \boldsymbol{y}_2$ are centered so that $\sum_n y_{1n} = 0, \sum_n y_{2n} = 0$. σ_{g1}^2 denotes the genetic variance of trait 1, *i.e.*, the variance component of phenotype 1 corresponding to the vector of genotypes across M SNPs while σ_{g2}^2 is the genetic variance for phenotype 2. $\sigma_{e1}^2, \sigma_{e2}^2$ denote the residual variance (variance not explained by genetics) for each of the two traits.

γ_g and γ_e denote the genetic and residual covariances. We define the genetic correlation as $\rho_g \equiv \frac{\gamma_g}{\sigma_{g1}\sigma_{g2}}$. Let $\boldsymbol{y} \equiv [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T]^T$, $\boldsymbol{\epsilon} \equiv [\boldsymbol{\epsilon_1}^T, \boldsymbol{\epsilon_2}^T]^T$, and $\boldsymbol{\beta} \equiv [\boldsymbol{\beta_1}^T, \boldsymbol{\beta_2}^T]^T$. Thus we have:

$$
\boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X} \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}
$$

We have $\mathbb{E}[\boldsymbol{y}] = 0$ since the phenotypes are centered. The population covariance of the two phenotypes is given by:

$$
cov(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^T = \begin{bmatrix} \sigma_{g1}^2\boldsymbol{K} & \gamma_g\boldsymbol{K} \\ \gamma_g\boldsymbol{K}^T & \sigma_{g2}^2\boldsymbol{K} \end{bmatrix} + \begin{bmatrix} \sigma_{e1}^2 I_N & \gamma_e \boldsymbol{I}_N \\ \gamma_e I_N & \sigma_{e2}^2 \boldsymbol{I}_N \end{bmatrix} \tag{2}
$$

Here $\boldsymbol{K} = \frac{\boldsymbol{X}_1\boldsymbol{X}_1^T}{M}$ is the genetic relatedness matrix (GRM).

We aim to jointly estimate the genetic and residual variance as well as covariance parameters. Our approach to estimate both the variance components and the genetic correlation relies on

a Method-of-Moments (MoM) estimator obtained by equating the population covariance to the empirical covariance. The empirical covariance of the concatenated phenotype vector $\boldsymbol{y}$ is estimated by the sample covariance: $\boldsymbol{y}\boldsymbol{y}^T$. The MoM estimator is obtained by solving the following ordinary least squares problem:

$$(\widehat{\gamma_g}, \widehat{\gamma_e}, \widehat{\sigma^2_{g1}}, \widehat{\sigma^2_{g2}}, \widehat{\sigma^2_{e1}}, \widehat{\sigma^2_{e2}}) = argmin_{\gamma_g, \gamma_e, \sigma^2_{g1}, \sigma^2_{g2}, \sigma^2_{e1}, \sigma^2_{e2}} ||\boldsymbol{y}\boldsymbol{y}^T - (\begin{bmatrix} \sigma^2_{g1}\boldsymbol{K} & \gamma_g \boldsymbol{K} \\ \gamma_g \boldsymbol{K}^T & \sigma^2_{g2}\boldsymbol{K} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1} I_N & \gamma_e I_N \\ \gamma_e I_N & \sigma^2_{e2} I_N \end{bmatrix})||^2_F \tag{3}$$

Setting the gradient of the objective function to zero gives us the normal equations (see Supplementary Material). We observe that solving for genetic and residual covariance parameters (γ_g, γ_e) is independent of solving the variance parameters: $\sigma^2_{g1}, \sigma^2_{e1}, \sigma^2_{g2}, \sigma^2_{e2}$. Thus, MoM estimates of the covariance parameters can be obtained by solving the set of normal equations:

$$\begin{bmatrix} tr(\boldsymbol{K}^2) & tr(\boldsymbol{K}) \\ tr(\boldsymbol{K}) & N \end{bmatrix} \begin{bmatrix} \widehat{\gamma_g} \\ \widehat{\gamma_e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_2^T \boldsymbol{K} \boldsymbol{y}_1 \\ \boldsymbol{y}_2^T \boldsymbol{y}_1 \end{bmatrix} \tag{4}$$

The GRM $\boldsymbol{K}$ can be computed in time $O(MN^2)$ and requires $O(N^2)$ memory. Given the GRM, computing each of the coefficients for the normal equations requires $O(N^2)$ time.

Given each of the coefficients, we can solve analytically for $\widehat{\gamma_g}$, and $\widehat{\gamma_e}$. Indeed, we can write:

$$\widehat{\gamma_g} = \frac{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_1^T \boldsymbol{y}_2}{tr[\boldsymbol{K}^2] - N}$$

Here we have used the property that $tr(\boldsymbol{K}) = N$ due to the use of a standardized genotype matrix.

Finally, we use estimates of the genetic variance parameters to obtain a plug-in estimate of the genetic correlation:

$$\widehat{\rho_g} = \frac{\widehat{\gamma_g}}{\sqrt{\widehat{\sigma^2_{g1}}}\sqrt{\widehat{\sigma^2_{g2}}}}$$

The estimators for σ^2_{g1} and σ^2_{g2} are give by $\widehat{\sigma^2_{g1}} = \frac{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_1 - \boldsymbol{y}_1^T \boldsymbol{y}_1}{tr[\boldsymbol{K}^2] - N}$ and $\widehat{\sigma^2_{g2}} = \frac{\boldsymbol{y}_2^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_2^T \boldsymbol{y}_2}{tr[\boldsymbol{K}^2] - N}$ (see Supplementary Material).

Substituting these expressions for the genetic covariance and variances gives us the following estimator of genetic correlation:

$$\widehat{\rho_g} = \frac{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_1^T \boldsymbol{y}_2}{\sqrt{\boldsymbol{y}_1^T \boldsymbol{K} \boldsymbol{y}_1 - \boldsymbol{y}_1^T \boldsymbol{y}_1}\sqrt{\boldsymbol{y}_2^T \boldsymbol{K} \boldsymbol{y}_2 - \boldsymbol{y}_2^T \boldsymbol{y}_2}} \tag{5}$$

2.2 Model assuming partial overlap of samples across traits

We now generalize our approach to the setting where the traits are no longer observed on the same samples. Assume we have N_1 samples for trait 1 and N_2 samples for trait 2 of which N samples ($N \leq N_1, N \leq N_2$) contain measurements for both the traits. $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$ denote the matrix of genotypes for the two traits separately and assume that the samples are observed on the same set of SNPs. We define $\boldsymbol{X_1}, \boldsymbol{X_2}$ to be the $N_1 \times M$ and $N_2 \times M$ matrices of standardized genotypes obtained by centering and scaling each column of $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$ so that $\sum_n x_{a,n,m} = 0$ for all $m \in \{1, \ldots, M\}, a \in \{1, 2\}$. Let $\boldsymbol{y_1}, \boldsymbol{y_2}$ denote the two vectors of phenotype with size N_1 and N_2 respectively. Additionally, we define an $N_1 \times N_2$ indicator matrix, $\boldsymbol{C}$ where $\boldsymbol{C}_{i,j} = 1$ when

individual i and j refer to the same sample and 0 otherwise. We also define $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$ to be the M-vectors of SNP effect sizes.

We assume the two phenotypes, $\boldsymbol{y_1}, \boldsymbol{y_2}$ are related to the genotypes by the following bivariate linear mixed model:

$$
\begin{aligned}
\boldsymbol{y_1}|\boldsymbol{\epsilon_1}, \boldsymbol{\beta_1} &= \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_1} \\
\boldsymbol{y_2}|\boldsymbol{\epsilon_2}, \boldsymbol{\beta_2}, &= \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon_2} \\
\epsilon_{1n_1}, \epsilon_{2n_2} &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \frac{\sigma_{e1}^2}{N_1} & \frac{\gamma_e}{N} \\ \frac{\gamma_e}{N} & \frac{\sigma_{e2}^2}{N_2} \end{bmatrix}) \\
&\quad n_1, n_2 \text{ refer to the same sample}, C_{n_1,n_2} = 1 \\
\epsilon_{1n_1} &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{\sigma_{e1}^2}{N_1}) \\
&\quad n_1 \text{ does not overlap any sample measured for trait 2}, \sum_{n_2} C_{n_1,n_2} = 0 \\
\epsilon_{2n_2} &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{\sigma_{e2}^2}{N_2}) \\
&\quad n_2 \text{ does not overlap any sample measured for trait 1}, \sum_{n_1} C_{n_1,n_2} = 0 \\
\beta_{1m}, \beta_{2m} &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \frac{\sigma_{g1}^2}{M} & \frac{\gamma_g}{M} \\ \frac{\gamma_g}{M} & \frac{\sigma_{g2}^2}{M} \end{bmatrix}))
\end{aligned}
$$

The population covariance of the phenotypes is now:

$$
cov(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^T = \begin{bmatrix} \sigma_{g1}^2\boldsymbol{K_1} & \gamma_g\boldsymbol{K_A} \\ \gamma_g\boldsymbol{K_A}^T & \sigma_{g2}^2\boldsymbol{K_2} \end{bmatrix} + \begin{bmatrix} \sigma_{e1}^2 I_{N_1} & \gamma_e\boldsymbol{C} \\ \gamma_e\boldsymbol{C}^T & \sigma_{e2}^2 I_{N_2} \end{bmatrix} \tag{6}
$$

Here $\boldsymbol{K_1} = \frac{\boldsymbol{X_1}\boldsymbol{X_1}^T}{M}$ is the GRM for the samples observed for the first trait while $\boldsymbol{K_2} = \frac{\boldsymbol{X_2}\boldsymbol{X_2}^T}{M}$ is the GRM for the samples for the second trait. $\boldsymbol{K_A}$ is the GRM for all pairs of samples: $\boldsymbol{K_A} = \frac{\boldsymbol{X_1}\boldsymbol{X_2}^T}{M}$.

Thus the MoM estimator could be obtained by equating the population covariance to the empirical covariance, estimated by $\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{T}}$. Thus the MoM estimator is obtained by solving the following ordinary least squares problem:

$$
(\widehat{\gamma_g}, \widehat{\gamma_e}, \widehat{\sigma_{g1}^2}, \widehat{\sigma_{g2}^2}, \widehat{\sigma_{e1}^2}, \widehat{\sigma_{e2}^2}) = argmin_{\gamma_g, \gamma_e, \sigma_{g1}^2, \sigma_{g2}^2, \sigma_{e1}^2, \sigma_{e2}^2} ||\boldsymbol{y}\boldsymbol{y}^T - (\begin{bmatrix} \sigma_{g1}^2\boldsymbol{K_1} & \gamma_g\boldsymbol{K_A} \\ \gamma_g\boldsymbol{K_A}^T & \sigma_{g2}^2\boldsymbol{K_2} \end{bmatrix} + \begin{bmatrix} \sigma_{e1}^2\boldsymbol{I}_{N_1} & \gamma_e\boldsymbol{C} \\ \gamma_e\boldsymbol{C}^T & \sigma_{e2}^2\boldsymbol{I}_{N_2} \end{bmatrix})||_F^2 \tag{7}
$$

Define $\boldsymbol{K_C}$ to be the GRM for samples with measurements of both phenotypes while $\boldsymbol{y}_{1C}$ and $\boldsymbol{y}_{2C}$ denote the N-vector of phenotypes for traits 1 and 2. The MoM estimator for genetic covariance satisfies the set of normal equations:

$$
\begin{bmatrix} tr(\boldsymbol{K_A}\boldsymbol{K_A}^{\mathrm{T}}) & tr(\boldsymbol{K_C}) \\ tr(\boldsymbol{K_C}) & N \end{bmatrix} \begin{bmatrix} \widehat{\gamma_g} \\ \widehat{\gamma_e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T\boldsymbol{K_A}\boldsymbol{y}_2 \\ \boldsymbol{y}_{2C}^T\boldsymbol{y}_{1C} \end{bmatrix} \tag{8}
$$

Finally, given each of the coefficients, we can solve analytically for $\widehat{\gamma_g}$, and $\widehat{\gamma_e}$.

Finally the estimate of genetic correlation is given by the plug-in estimate:

$$
\widehat{\rho_g} = \frac{\widehat{\gamma_g}}{\sqrt{\widehat{\sigma_{g1}^2}}\sqrt{\widehat{\sigma_{g2}^2}}} \tag{9}
$$

The estimators for $\sigma_{g1}^2, \sigma_{g2}^2$ require computing $tr[\boldsymbol{K}_1^2]$ and $tr[\boldsymbol{K}_2^2]$ (see Supplementary Materials).

2.3 RG-Cor: Randomized MoM estimator for Genetic Correlation

The computational bottleneck in obtaining MoM estimators for $(\widehat{\gamma_g}, \widehat{\sigma^2_{g1}}, \widehat{\sigma^2_{g2}})$ lies in computing $\boldsymbol{y}_s^{\mathrm{T}} \boldsymbol{K} \boldsymbol{y}_t, s, t \in \{1, 2\}$ for the setting of complete overlap of samples, and computing $tr(\boldsymbol{K_A} \boldsymbol{K_A}^{\mathrm{T}})$, $tr(\boldsymbol{K_1}^2)$ $tr(\boldsymbol{K_2}^2)$ for partially overlapping samples.

Naive computation of $tr(\boldsymbol{K_A} \boldsymbol{K_A}^{\mathrm{T}})$ requires $\mathcal{O}(N_1 N_2 M)$ operations, where N_1, N_2 are the sample size of each of the traits. Similarly, $tr(\boldsymbol{K_1}^2)$ and $tr(\boldsymbol{K_2}^2)$ can be computed in $\mathcal{O}(N_1^2 M)$ and $\mathcal{O}(N_2^2 M)$ time.

To overcome this computational bottleneck, we replace these quantities with randomized estimators $\widehat{tr(\boldsymbol{K_1}^2)}$, $\widehat{tr(\boldsymbol{K_2}^2)}$ and $\widehat{tr(\boldsymbol{K_A} \boldsymbol{K_A}^{\mathrm{T}})}$.

Given a $N \times N$ matrix $\boldsymbol{A}$ and a random vector $\boldsymbol{z}$ with mean zero and covariance $\boldsymbol{I}_N$, we use the following identity to construct the randomized estimators [4].

$$\mathbb{E}[\boldsymbol{z}^T \boldsymbol{A} \boldsymbol{z}] = tr[\boldsymbol{A}] \tag{10}$$

Equation 10 leads to the following unbiased estimator for the trace of $tr(\boldsymbol{K_A} \boldsymbol{K_A}^{\mathrm{T}})$, $tr(\boldsymbol{K_1}^2)$ $tr(\boldsymbol{K_2}^2)$ given B random vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$, $z_b \in \mathbb{R}^N, b \in 1 \ldots B$ drawn independently from a distribution with zero mean and identity covariance matrix $\boldsymbol{I_N}$:

$$L_{B_A} = \widehat{tr(\boldsymbol{K_A} \boldsymbol{K_A}^{\mathrm{T}})} = \frac{1}{B} \frac{1}{M^2} \sum_b ||\boldsymbol{X_A} \boldsymbol{X_A}^T \boldsymbol{z}_b||_2^2$$

$$L_{B_1} = \widehat{tr(\boldsymbol{K}_1^2)} = \frac{1}{B} \frac{1}{M^2} \sum_b ||\boldsymbol{X}_1 \boldsymbol{X}_1^T \boldsymbol{z}_b||_2^2$$

$$L_{B_2} = \widehat{tr(\boldsymbol{K}_2^2)} = \frac{1}{B} \frac{1}{M^2} \sum_b ||\boldsymbol{X}_2 \boldsymbol{X}_2^T \boldsymbol{z}_b||_2^2$$

In practice, we compute the above estimators by drawing each entry of $\boldsymbol{z}_b$ independently from a standard normal distribution.

The RG-Cor estimator $(\tilde{\gamma_g}, \tilde{\gamma_e})$ is obtained by solving Equation 8 by replacing $tr[\boldsymbol{K_A} \boldsymbol{K_A}^{\mathrm{T}}]$ with L_{B_A}.

$$\begin{bmatrix} L_{B_A} & tr(\boldsymbol{K_C}) \\ tr(\boldsymbol{K_C}) & N \end{bmatrix} \begin{bmatrix} \tilde{\gamma_g} \\ \tilde{\gamma_e} \end{bmatrix} = \begin{bmatrix} tr(\boldsymbol{y}_1^T \boldsymbol{K_A} \boldsymbol{y}_2) \\ tr(\boldsymbol{y}_{1N} \boldsymbol{y}_{2N}^T) \end{bmatrix}$$

2.4 Inclusion of covariates

In a number of settings, it is desirable to include covariates, such as age, sex, or principal components to correct for population structure, in the analysis. In the complete overlap setting, the samples share the covariates. This changes Equation 1 into:

$$\boldsymbol{y} | \boldsymbol{\epsilon}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{W} \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X} \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{11}$$

Here $\boldsymbol{W}$ is $N \times C$ matrix of covariates while $\boldsymbol{\alpha}$ is a C-vector of fixed effects. In this setting, we transform Equation 11 by multiplying both slides by the projection matrix $\begin{bmatrix} \boldsymbol{V} \\ \boldsymbol{V} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I_N} - \boldsymbol{W}(\boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{W}^T \\ \boldsymbol{I_N} - \boldsymbol{W}(\boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{W}^T \end{bmatrix}$

Similarly, for traits that partially share samples, where the covariate is $\begin{bmatrix} \boldsymbol{W_1} \\ \boldsymbol{W_2} \end{bmatrix}$, where $\boldsymbol{W_1}$ is the covariate for trait 1, and $\boldsymbol{W_2}$ is the covariate for trait 2, the projection matrix is: $\begin{bmatrix} \boldsymbol{V_1} \\ \boldsymbol{V_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I_{N_1}} - \boldsymbol{W_1}(\boldsymbol{W_1}^T\boldsymbol{W_1})^{-1}\boldsymbol{W_1}^T \\ \boldsymbol{I_{N_2}} - \boldsymbol{W_2}(\boldsymbol{W_2}^T\boldsymbol{W_2})^{-1}\boldsymbol{W_2}^T \end{bmatrix}$

3 Experiments

3.1 Simulation to assess the accuracy and computational efficiency of RG-Cor

We performed simulations to compare the performance of RG-Cor to other methods for genetic correlation estimation in terms of accuracy, running time and memory usage. Specifically, we compared the performance of RG-Cor to GREML (as implemented in the GCTA software) [5], BOLT-REML [7], and LD-score regression (LDSC) [2]. The GREML software aims to compute maximum likelihood (or restricted maximum likelihood (REML)) estimates of a bi-variate linear mixed model [5]. BOLT-REML is an approximate REML method that can scale to larger problem sizes relative to GREML. LDSC, on the other hand, is widely used to estimate genetic correlation when only summary statistics from GWAS on pairs of traits are available.

Experiments were based on real genotypes from the UK Biobank [11] and on the Northern Finland Birth Cohort (NFBC) dataset [10]. We simulated pairs of traits with known values of heritability and genetic correlation. Experiments to assess the estimation accuracy of each method used the full NFBC dataset, containing $315,529$ SNPs and 5326 individuals, so that all the methods could be run in reasonable time. While comparing computational efficiency, we compared RG-Cor to BOLT-REML and GREML in terms of running time and memory usage on subsets of UK biobank data.

3.2 RG-Cor estimator is accurate

In our first set of simulations, we compared the accuracy of RG-Cor to GREML, BOLT-REML, and LDSC. We evaluated the accuracy of estimates of RG-Cor when the number of random vectors B were set to 10 as well as 100.

For these experiments, we analyzed the Northern Finland Birth Cohort (NFBC) dataset, which contains 5326 individuals and $315,529$ SNPs after removing SNPs with minor allele frequency ≤ 0.05 and with Hardy-Weinberg Equilibrium p-value < 0.01.

Given the genotypes, we simulated a pair of phenotypes based on the complete overlap model specified in equation (1). We assume all SNPs have an effect on each trait (*i.e.*, the trait architecture is infinitesimal). We considered settings where the true heritability of phenotypes are i) both low (set to 0.1 and 0.2 respectively), and ii) one of the phenotypes has low heritability while the other has high heritability (set to 0.2 and 0.8) Fixing the true heritability of each phenotype, we vary the true genetic correlation across $\{0, 0.2, 0.5, 0.8\}$. We repeated each experiment 100 times.

Figure 1(a), we show the situation where the true heritability of both phenotypes are low and fixed to be 0.1 and 0.2. This is a typical situation since complex phenotypes tend to have low heritability in human populations. In Figure 1(b), the true heritability of each phenotype is fixed to be 0.2 and 0.8. Table 1 summarizes these results reporting the bias, standard error and mean square error (MSE) of the methods for each parameter setting of Figure 1. We observe that the the statistical efficiency of the estimates from BOLT (approximate REML) and RG-Cor are comparable. While GREML estimates tend to be the most statistically efficient, as expected, in

the low heritability setting, RG-Cor achieves a lower standard error than GREML. We also observe that in all cases LDSC attains large standard errors, consistent with previous observations [14]. We conclude that RG-Cor is comparable to BOLT-REML and is particularly useful in cases where heritability for both traits are low and their genetic correlation is high. Finally, the results are indistinguishable when RG-Cor uses $B = 10$ versus $B = 100$.

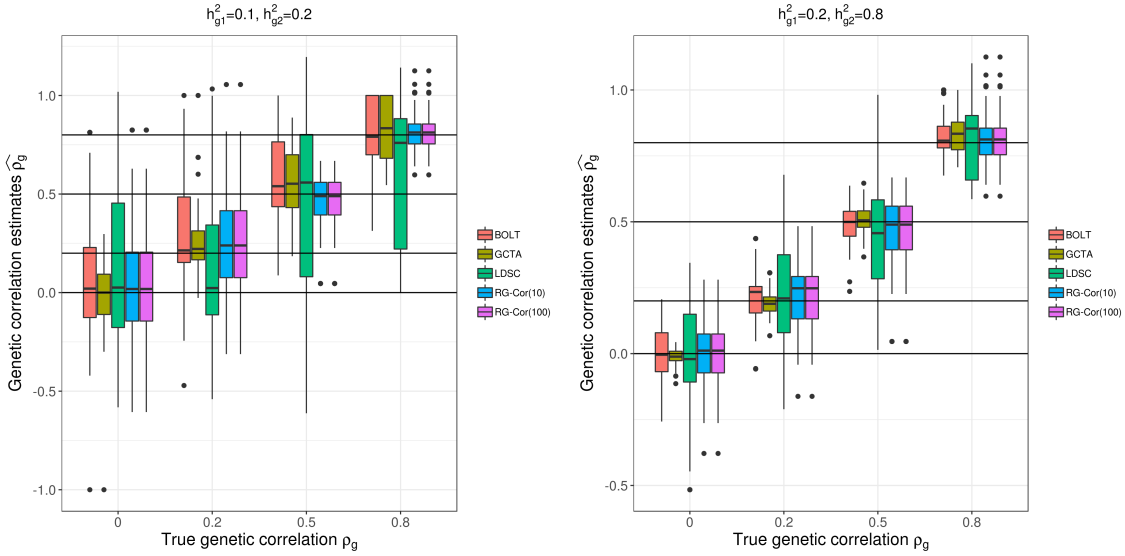

Figure 1: **RG-Cor is an accurate estimator of genetic correlation**: We compared the accuracy of methods for genetic correlation estimation using simulated phenotypes and genotypes from the North Finland Birth Cohort (NFBC). In figure 1(a), the heritability of two traits are fixed to be 0.1 and 0.2 while in figure 1(b), the heritabilities of the two traits are 0.2 and 0.8. We vary the genetic correlation to be $\{0, 0.2, 0.5, 0.8\}$. In some cases where the genetic correlation is high, RG-Cor is statistical efficient relative to . We observe that the standard error of the RG-Cor estimates is relatively insensitive when we change B from 10 to 100.

Table 1: **Estimates of bias, mean square error and standard error of genetic correlation estimation methods in simulations**

Method	Trait 1 h^2_g	Trait 2 h^2_g	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	SE	Bias	MSE	se
			$\rho_g = 0$			$\rho_g = 0.2$			$\rho_g = 0.5$			$\rho_g = 0.8$		
GREML	0.1	0.2	-0.007	0.036	0.19	0.037	0.025	0.154	0.054	0.051	0.168	0.03	0.022	0.145
BOLT-REML	0.1	0.2	0.062	0.104	0.316	0.091	0.093	0.29	0.042	0.052	0.223	0.013	0.029	0.171
LDSC	0.1	0.2	0.125	0.226	0.459	-0.093	0.14	0.362	-0.121	0.263	0.499	-0.182	0.162	0.36
RG-Cor (B=100)	0.1	0.2	0.05	0.1	0.313	0.072	0.083	0.281	-0.039	0.018	0.129	0.025	0.011	0.1
RG-Cor (B=10)	0.1	0.2	0.05	0.1	0.313	0.072	0.083	0.281	-0.039	0.018	0.129	0.025	0.011	0.1
GREML	0.2	0.8	-0.01	0.001	0.033	-0.007	0.003	0.05	0.007	0.004	0.062	0.034	0.008	0.084
BOLT-REML	0.2	0.8	0.008	0.01	0.099	0.014	0.011	0.103	-0.003	0.006	0.081	0.028	0.009	0.091
LDSC	0.2	0.8	-0.032	0.043	0.204	0.013	0.035	0.185	-0.056	0.044	0.202	0.003	0.021	0.144
RG-Cor (B=100)	0.2	0.8	-0.016	0.018	0.134	0.013	0.016	0.129	-0.04	0.018	0.129	0.025	0.01	0.1
RG-Cor (B=10)	0.2	0.8	-0.016	0.018	0.134	0.013	0.016	0.129	-0.04	0.018	0.129	0.025	0.01	0.1

3.3 RG-Cor is computationally efficient

In order to measure computational efficiency, we sub-sampled the UK Biobank genotypes to sample sizes of $1,000$, $2,000$, $5,000$, $10,000$ $50,000$, $100,000$, and $290,000$ which is approximately the sample size of the UK Biobank dataset after quality control.

Prior the sub-sampling experiment, we performed the following individual-level and SNP-level quality controls. We constrained the samples to the British white population as indicated by self-reported ethnicity. We removed $14,255$ samples with missingness > 0.1 . We restricted our analysis to SNPs that were present on the UKBiobank Axiom array used to genotype the UKBiobank. We removed SNPs with greater than 1% missingness and minor allele frequency smaller than 1%. Our final dataset contained $291,273$ individuals and $459,792$ SNPs after quality control. All experiments were performed on an AMD EPYC machine on which we restricted the run time to 6 days and memory usage to 200GB.

Figure 2 shows that both GREML and BOLT-REML do not scale to large sample sizes. GREML could not scale beyond sample sizes greater than $100,000$ due to the requirement of computing and operating on a genetic relatedness matrix (GRM)The runtime of BOLT-REML scales as $N^{1.5}$ as reported previously[7]. We observed a difficulty in convergence while running BOLT-REML on subsets of the UK Biobank data. Based on the observed runtimes, we extrapolate that BOLT-REML would require about 17 days to run on the full UK Biobank dataset with $291,273$ samples. On the other hand, RG-Cor ran in about 3 hours and used 50 GB memory on the set of $291,273$ individuals. The memory usage of RG-Cor also scales linearly.

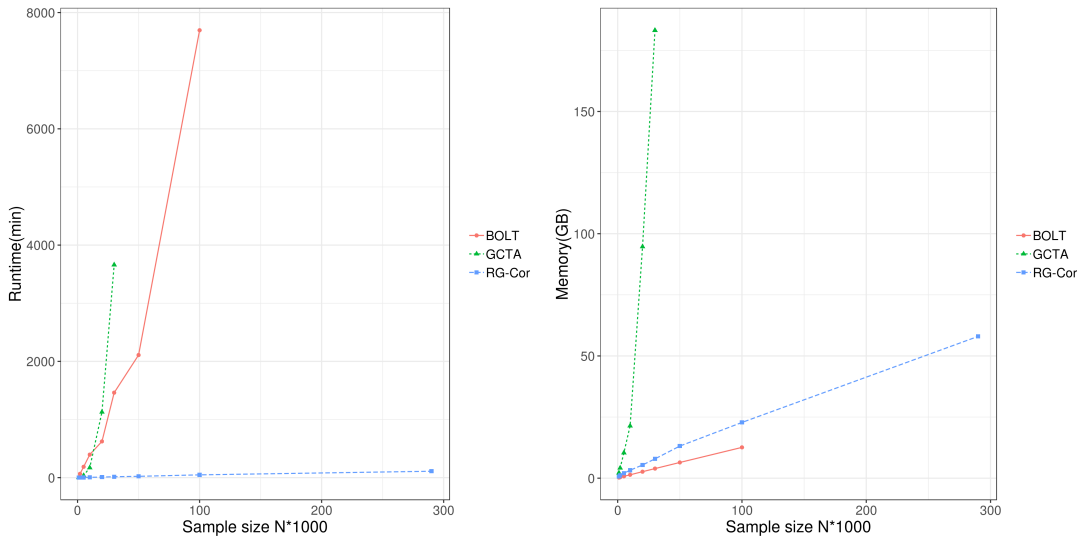

Figure 2: **RG-Cor is efficient**: We measured the run time and memory usage of methods for genetic correlation estimation as a function of the number of samples while fixing the number of SNPs to $459,792$. The samples were obtained as subsets of unrelated, white British individuals in the UK Biobank. We performed all comparisons on an AMD EPYC machine. In Figure 2(a), GREML could not finish computation on $100,000$ samples. BOLT-REML scales well but is nevertheless computationally intensive with increasing sample sizes. RG-Cor runs in a few hours on even the largest dataset. Figure 2(b) shows that both RG-Cor and BOLT-REML have scalable memory requirements.

3.4 Genetic correlation among traits in UK biobank

We applied RG-Cor to analyze phenotypes in the UK Biobank. We restricted our analysis to SNPs genotyped on the UK Biobank Axiom array, filtering out the genetic markers that had high missingness rate ($> \%1$) and low minor allele frequency ($< 1\%$). We also filtered out subjects that had high missing genotype rate ($> 1\%$), as well as samples that have genetic kinship with any other sample (samples having any relatives in the dataset using the phenotype field 22021: Genetic

kinship to other participants).After quality control, we obtained $291,273$ non-related individuals and $459,792$ SNPs.

We analyzed seven continuous phenotypes and estimated the genetic correlations on all 21 pairs of phenotypes (Figure 3). All genetic correlation estimates across traits are adjusted for gender, UK Biobank assessment center, age at recruitment, and top 10 principal components of the genotype. We estimated the standard error of the RG-Cor estimator using a computationally efficient block Jackknife (see Supplementary Material).

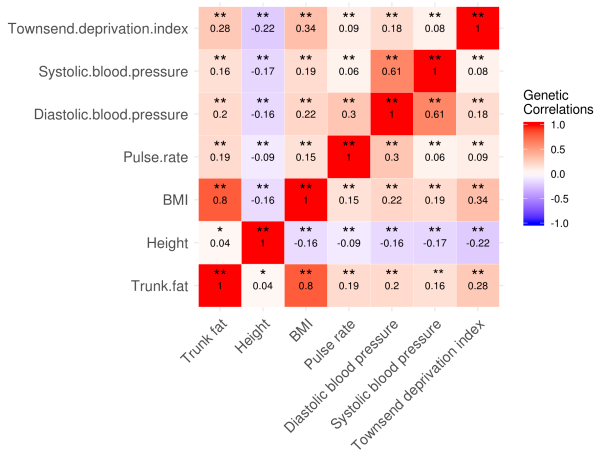

Figure 3: **Genetic correlation among seven traits in the UK Biobank analyzed by RG-Cor**. Blue indicates negative genetic correlation and red indicates positive genetic correlation. Genetic correlations that are significantly different from 0 are marked with an asterisk. Genetic correlations that are significantly different from 0 after Bonferroni correction for 21 tests in this analysis are marked with two asterisk. Genetic correlations between traits are computed after correcting for covariates: gender, UK Biobank assessment center, age at recruitment, and top 10 genotype principal components.

4 Discussion

We have described RG-Cor, a scalable estimator for genetic correlation. We show that the RG-Cor estimates for genetic correlation are accurate and achieve similar statistical efficiency while being highly scalable. We use RG-Cor to compute genetic correlations across seven continuous phenotypes in UK Biobank obtaining estimates consistent with previous results.

This genome-wide analysis of genetic correlation is the stepping stone for understanding the relationships across human traits and diseases. In future analyses, we intend to systematically scan pairs of traits in biobank datasets to obtain genetic correlation estimates. We can further partition the genetic correlation with respect to the function and minor allele frequency of the SNPs to further interpret the underlying relationships and causality.

Availability

The RHE-reg software is made freely available to the research community at:
https://github.com/sriramlab/RHE-reg

Acknowledgments

This research was conducted using the UK Biobank Resource under applications 33127. We thank the participants of UK Biobank for making this work possible. SS was supported in part by is supported in part by NIH grants R35GM125055, NSF Grant III-1705121, an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Foundation.

References

1. Akiyama, M., Okada, Y., Kanai, M., *et al.* (2017). Genome-wide association study identifies 112 new loci for body mass index in the japanese population. *Nature genetics*, **49**(10), 1458.

2. Bulik-Sullivan, B., Finucane, H. K., Anttila, V., *et al.* (2015). An atlas of genetic correlations across human diseases and traits. *Nature genetics*, **47**(11), 1236.

3. Chen, G.-B. (2014). Estimating heritability of complex traits from genome-wide association studies using ibs-based haseman–elston regression. *Frontiers in genetics*, **5**, 107.

4. Hutchinson, M. (1989). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, **18**(3), 1059–1076.

5. Lee, S., Yang, J., Goddard, M., Visscher, P., and Wray, N. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, **28**(19), 2540–2542.

6. Lee, S. H., Ripke, S., Neale, B. M., *et al.* (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature genetics*, **45**(9), 984.

7. Loh, P.-R., Bhatia, G., Gusev, A., *et al.* (2015a). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, **47**(12), 1385.

8. Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., *et al.* (2015b). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, **47**(3), 284.

9. Ni, G., Moser, G., Ripke, S., *et al.* (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *The American Journal of Human Genetics*.

10. Sabatti, C., Service, S. K., Hartikainen, A.-L., *et al.* (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, **41**(1), 35–46.

11. Sudlow, C., Gallacher, J., Allen, N., *et al.* (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, **12**(3), e1001779.

12. Wu, Y. and Sankararaman, S. (2018). A scalable estimator of snp heritability for biobank-scale data. *Bioinformatics*, **34**(13), i187–i194.

13. Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2010). GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet*.

14. Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics*, **11**(4), 2027.