## Title
Transethnic Genetic-Correlation Estimates from Summary Statistics



## Journal
American Journal of Human Genetics, 99(1)

## ISSN
0002-9297

## Authors
Brown, Brielin C
Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes
Ye, Chun Jimmie
et al.

## Publication Date
2016-07-01

## DOI
10.1016/j.ajhg.2016.05.001

Peer reviewed

# Transethnic Genetic-Correlation Estimates from Summary Statistics

Brielin C. Brown,[1,*] Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Chun Jimmie Ye,[2] Alkes L. Price,[3] and Noah Zaitlen[4]

The increasing number of genetic association studies conducted in multiple populations provides an unprecedented opportunity to study how the genetic architecture of complex phenotypes varies between populations, a problem important for both medical and population genetics. Here, we have developed a method for estimating the transethnic genetic correlation: the correlation of causal-variant effect sizes at SNPs common in populations. This methods takes advantage of the entire spectrum of SNP associations and uses only summary-level data from genome-wide association studies. This avoids the computational costs and privacy concerns associated with genotype-level information while remaining scalable to hundreds of thousands of individuals and millions of SNPs. We applied our method to data on gene expression, rheumatoid arthritis, and type 2 diabetes and overwhelmingly found that the genetic correlation was significantly less than 1. Our method is implemented in a Python package called Popcorn.

## Introduction

Many complex human phenotypes vary dramatically in their distributions between populations as a result of a combination of genetic and environmental differences. For example, northern Europeans are on average taller than southern Europeans,[1] and African Americans have an higher rate of hypertension than European Americans.[2] Differences in allele frequencies, effect sizes, and genetic architectures drive the genetic contribution to population phenotypic differentiation. Understanding the root causes of phenotypic differences worldwide has profound implications for biomedical and clinical practice in diverse populations, the transferability of epidemiological results, aiding multi-ethnic disease mapping,[3,4] assessing the contribution of non-additive and rare-variant effects, and modeling the genetic architecture of complex traits. In this work, we consider a central question in the global study of phenotype: do genetic variants have the same phenotypic effects in different populations?

Although the vast majority of genome-wide association studies (GWASs) have been conducted in European populations,[5,6] the growing number of non-European and multi-ethnic studies[4,7,8] provide an opportunity to study distributions of genetic effects across populations. For example, one recent study used mixed-model-based methods to show that the genome-wide genetic correlation of schizophrenia between European and African Americans is nonzero.[9] Although powerful, computational costs and privacy concerns limit the utility of genotype-based methods. In this work, we make two significant contributions to studies of transethnic genetic correlation. First, we expand the definition of genetic correlation to better account for a transethnic context. Second, we develop an approach that uses only summary-level GWAS data to estimate genetic correlation across populations. Like other recent methods based on summary statistics,[10–21] our approach supplements summary association data with linkage disequilibrium (LD) information from external reference panels, avoids privacy concerns, and is scalable to hundreds of thousands of individuals and millions of markers. Unlike traditional approaches that focus on the similarity of GWAS results,[22–26] we use the entire spectrum of GWAS associations while accounting for LD to avoid filtering correlated SNPs.

In a single population, the genetic correlation of two phenotypes is defined as the correlation coefficient of SNP effect sizes.[19,27] In multiple populations, differences in allele frequency motivate multiple possible definitions of genetic correlation. Because a variant can have a higher effect size but lower frequency in one population, we consider both the correlation of allele effect sizes and the correlation of allelic impact. We define the transethnic genetic-effect correlation ($\rho_{ge}$, previously defined by Lee et al.[27] and implemented in Genome-wide Complex Trait Analysis [GCTA]) as the correlation coefficient of the per-allele SNP effect sizes. Similarly, we define the transethnic genetic-impact correlation ($\rho_{gi}$) as the correlation coefficient of the population-specific allele-variance-normalized SNP effect sizes.

Intuitively, the genetic-effect correlation measures the extent to which the same variant has the same phenotypic change, whereas the genetic-impact correlation gives more weight to common alleles than to rare ones separately in each population. Consider the case of a SNP that is rare in population 1 but common in population 2 and has an identical effect size in both populations. In this case, the

[1]Department of Computer Science, University of California, Berkeley, Berkeley, CA 94720, USA; [2]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94117, USA; [3]Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA; [4]Department of Medicine, University of California, San Francisco, San Francisco, CA 94158, USA
*Correspondence: brielin@berkeley.edu
http://dx.doi.org/10.1016/j.ajhg.2016.05.001.

correlation of effect sizes (genetic-effect correlation, $\rho_{ge}$) is 1. This, however, provides an incomplete picture of the relationship between the two populations, given that the allele has a much bigger impact on the distribution of the phenotype in population 2. Therefore, we define the genetic-impact correlation, $\rho_{gi}$, as the correlation of effect sizes after genotypes are normalized to have mean 0 and variance 1. In our hypothetical case, $\rho_{gi} < \rho_{ge}$, but the opposite can also be true. Consider again the case of a SNP rare in population 1 but common in population 2. If the effect size is large in the first population but small in the second, then $\rho_{ge}$ might be much less than 1, but the impact of the allele in the two populations will be similar. Therefore, $\rho_{gi}$ will be close to 1. Although other definitions of the genetic correlation are possible (see Discussion), these quantities capture two important questions about the study of disease in multiple populations: to what extent do the same mutations in multiple populations differ in their phenotypic effects, and to what extent are these differences mitigated or exacerbated by differences in allele frequency?

To estimate genetic correlation, we take a Bayesian approach wherein we assume genotypes are drawn separately from within each population and effect sizes have a normal prior (the infinitesimal model[28]). Although this model is unlikely to represent reality, it has been used successfully in practice.[9,17,18,29,30] The infinitesimal assumption yields a multivariate normal distribution on the observed test statistics ($Z$ scores), where the covariance matrix is a function of the heritability and genetic correlation. Rather than pruning SNPs in LD,[11,31,32] this allows us to explicitly model the resulting inflation of $Z$ scores. We then maximize an approximate weighted likelihood function to find the heritability and genetic correlation. This method is implemented in a Python package called Popcorn. Although it is derived for quantitative phenotypes, Popcorn extends easily to binary phenotypes under the liability threshold model. We show via extensive simulation that Popcorn produces unbiased estimates of the genetic correlation and the population-specific heritabilities with a SE that decreases as the number of SNPs and individuals in the studies increases. Furthermore, we show that our approach is robust to violations of the infinitesimal assumption.

We applied Popcorn to European and Yoruban gene-expression data,[33] as well as GWAS summary statistics from European and East Asian cohorts affected by rheumatoid arthritis (RA) and type 2 diabetes (T2D).[34,35] Our analysis of GEUVADIS (Genetic European Variation in Health and Disease) data showed that our summary-statistic-based estimator is concordant with the mixed-model-based estimator. We found that the mean transethnic genetic correlation across all genes was low ($\rho_{ge} = 0.320$ [0.009]) but increased substantially when the gene was highly heritable in both populations ($\rho_{ge} = 0.772$ [0.017]). In RA and T2D, we found $\rho_{ge}$ to be 0.463 (0.058) and 0.621 (0.088), respectively.

Across all phenotypes considered, we overwhelmingly found that the transethnic genetic correlation is significantly less than 1. This observation highlights the need to study phenotypes in multiple populations because it implies that, up to the effects of unobserved variants, effect sizes at common SNPs tend to differ between populations. This indicates that results might not transfer between populations, and therefore predicting disease risk in non-Europeans on the basis of current GWAS results could be problematic. Our results provide further evidence that gaining insight into the genetic architecture of complex traits will require a multi-population approach.

## Material and Methods

Our method takes as input summary association statistics from two studies of a phenotype in two different populations, along with two sets of reference genotypes each matching one of the populations in the study. Our method has two steps: first, we estimate the diagonal elements of the LD-matrix products $\Sigma_1^2$, $\Sigma_2^2$, and $\Sigma_1\Sigma_2$; second, using these estimates, we find the maximum-likelihood values and estimate SEs of the parameters of interest: $h_1^2$ or $h_2^2$ and $\rho_{ge}$ or $\rho_{gi}$. The details follow.

Consider two GWASs conducted on the same phenotype in different populations. Assume we have $N_1$ individuals genotyped on $M$ SNPs in study 1 and $N_2$ individuals genotyped on the same SNPs in study 2. Let $X_1$ and $X_2$ be the matrices of mean-centered genotypes in studies 1 and 2, respectively, and let $Y_1$ and $Y_2$ be their normalized phenotypes. Let $f_1$ and $f_2$ be vectors of the allele frequencies of the $M$ SNPs common to both populations. If we assume Hardy-Weinberg equilibrium within each population separately, the allele variances are $\sigma_1^2 = 2f_1(1-f_1)$ and $\sigma_2^2 = 2f_2(1-f_2)$. Let $\beta_1$ and $\beta_2$ be the (unobserved) per-allele effect sizes for each SNP in studies 1 and 2, respectively. The heritability in study 1 is then $h_1^2 = \Sigma_i \sigma_{1i}^2 \beta_{1i}^2$ (and likewise for study 2). The objective of this work is to estimate transethnic genetic correlation from summary statistics of common variants, $Z_1 = [(X_1/\sigma_1)^T Y_1]/\sqrt{N_1}$ (and likewise for study 2), and estimates of population LD matrices ($\Sigma_1$ and $\Sigma_2$) from external reference panels. Define the genetic-effect correlation as $\rho_{ge} = \mathrm{Cor}(\beta_1, \beta_2)$ and the genetic-impact correlation as $\rho_{gi} = \mathrm{Cor}(\sigma_1\beta_1, \sigma_2\beta_2)$.

We assume that the genotypes are drawn randomly from each population and that phenotypes are generated by the linear model $Y_1 = X_1\beta_1 + \varepsilon_1$ (likewise for phenotype 2). When effect sizes $\beta$ are assumed to be inversely proportional to allele frequency, as is commonly done,[17,30] we show (Appendix A) that under the linear infinitesimal genetic architecture, the joint distribution of the $Z$ scores from each study is asymptotically multivariate normal with mean $\vec{0}$ and variance

$$\mathrm{Var}(Z) = \begin{bmatrix} \Sigma_1 + \frac{N_1+1}{M} h_1^2 \Sigma_1^2 & \rho_{gi}\sqrt{h_1^2 h_2^2}\,\frac{\sqrt{N_1 N_2}}{M}\Sigma_1\Sigma_2 \\ \rho_{gi}\sqrt{h_1^2 h_2^2}\,\frac{\sqrt{N_1 N_2}}{M}\Sigma_2\Sigma_1 & \Sigma_2 + \frac{N_2+1}{M} h_2^2 \Sigma_2^2 \end{bmatrix}. \quad \text{(Equation 1)}$$

However, when effect sizes are assumed to be independent of allele frequency, we show

$$\mathrm{Var}(Z) = \begin{bmatrix} \Sigma_1 + \frac{N_1+1}{\|\sigma_1^2\|_1} h_1^2 \Sigma_1 \sigma_1^2 \Sigma_1 & \rho_{ge}\sqrt{h_1^2 h_2^2}\ \frac{\sqrt{N_1 N_2}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}} \Sigma_1 \sqrt{\sigma_1^2 \sigma_2^2} \Sigma_2 \\ \rho_{ge}\sqrt{h_1^2 h_2^2}\ \frac{\sqrt{N_1 N_2}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}} \Sigma_2 \sqrt{\sigma_2^2 \sigma_1^2} \Sigma_1 & \Sigma_2 + \frac{N_2+1}{\|\sigma_2^2\|_1} h_2^2 \Sigma_2 \sigma_2^2 \Sigma_2 \end{bmatrix}. \quad \text{(Equation 2)}$$

Given these equations for variance, we can estimate the quantities $\rho_{gi}$ or $\rho_{ge}$ and $h_1^2$ or $h_2^2$ by maximizing the multivariate normal likelihood, $l(\rho_{g\{i,e\}}, h_1^2, h_2^2 | Z, \Sigma, \sigma) \propto -\ln(|C|) - Z^T C^{-1} Z$, where $C$ is either of the above covariance matrices in Equation 1 or 2. Because $\Sigma_1$ and $\Sigma_2$ are estimated from finite external reference panels, estimating the maximum likelihood of the above multivariate normal distribution leads to over-fitting. We employ two optimizations to avoid this problem. First, we maximize an approximate weighted likelihood that uses only the diagonal elements of each block of Var($Z$). This allows us to account for the LD-induced inflation of tests statistics, but it discards covariance information between pairs of $Z$ scores and therefore leads to over-counting $Z$ scores of SNPs in high LD. To compensate for this, we downweight $Z$ scores of SNPs in proportion to their LD. Second, rather than compute the full products $\Sigma_1^2$, $\Sigma_2^2$, and $\Sigma_1\Sigma_2$ over all $M$ SNPs in the genome, we choose a window size $W$ and approximate the product by $(\Sigma_a \Sigma_b)_{ii} = \sum_{w=i-W}^{w=i+W} r_{aiw} r_{biw}$. These optimizations are similar to those employed by LD-score regression.[17] The full details of the derivation and optimization are provided in Appendix A.

## Results

### Simulated Genotypes and Simulated Phenotypes

Using HAPGEN2,[34] we simulated 50,000 European (EUR)-like and 50,000 East Asian (EAS)-like individuals at 248,953 chromosome 1–3 SNPs with an allele frequency above 1% in both EUR and EAS HapMap 3 populations. HAPGEN2 implements a model that combines haplotype recombination with mutation and results in excess local relatedness among the simulated individuals. To account for this local structure, we used PLINK 2[35] to filter individuals with genetic relatedness above 0.05, resulting in 4,499 EUR-like individuals and 4,837 EAS-like individuals. From these simulated individuals, 500 per population were chosen uniformly at random to serve as an external reference panel for estimating $\Sigma_1$ and $\Sigma_2$.

In each simulation, effect sizes were drawn from a "spike and slab" model, where $\beta_{1i}, \beta_{2i} \sim \mathcal{N}\left(0, \begin{bmatrix} h_1^2 & \rho_{ge}\sqrt{h_1^2 h_2^2} \\ \rho_{ge}\sqrt{h_1^2 h_2^2} & h_2^2 \end{bmatrix}\right)$ with probability $p$ and $\beta_{1i}, \beta_{2i} = (0,0)$ with probability $1-p$. $\rho_{gi}$ were analytically computed from the simulated effect sizes and allele frequencies in the simulated reference genotypes. Quantitative phenotypes were generated under a linear model with independent and identically distributed noise and normalized to have mean 0 and variance 1, whereas binary phenotypes were generated under a liability threshold model where individuals are labeled as case subjects when their liability exceeds a threshold $\tau = \Phi^{-1}(1-K)$, in which $K$ is the population disease prevalence.[36]

We varied $h_1^2$, $h_2^2$, $\rho_{ge}$, and $\rho_{gi}$, as well as the number of individuals in each study ($N_1$ and $N_2$), the number of SNPs ($M$), the population prevalence ($K$), and proportion of causal variants ($p$) in the simulated GWASs, and generated summary statistics for each study. The results shown in Figure 1 and Figure S1 demonstrate that the estimators are nearly unbiased as the genetic correlation and heritabilities vary. Furthermore, by varying the proportion of causal variants $p$, we show that our estimator is robust to violations of the infinitesimal assumption (Figure S2). In Figure S3, we show that the SE of the estimator decreases as the number of SNPs and individuals in the study increases. In Figure S4, we simulate data for 4,499 EUR-like individuals and 15,101 EAS-like individuals for a range of genetic correlations to show that our estimators remain nearly unbiased when the sample sizes of the two populations are very different. Finally, we show in Table S1 that our estimates of the heritability of liability in case-control studies are nearly unbiased.

### Simulations with Nonstandard Disease Models

Our approach, as well as genotype-based methods such as GCTA, makes assumptions about the genetic architecture of complex traits. Previous work has shown that violations of these assumptions can lead to bias in heritability estimation;[37] therefore, we sought to quantify the extent to which this bias might affect our estimates. We simulated phenotypes under six different disease models: (1) independent, where the effect size is independent of allele frequency; (2) inverse, where the effect size is inversely proportional to allele frequency; (3) rare, where only SNPs with an allele frequency under 10% affect the trait; (4) common, where only SNPs with an allele frequency between 40% and 50% affect the trait; (5) difference, where effect size is proportional to the difference in allele frequency; and (6) adversarial, a difference model where the sign of $\beta$ is set to increase the phenotype in the population where the allele is most common. Additional genetic architectures, including ones where effect sizes are not a direct function of MAF,[38] are possible.

We simulated phenotypes by using genotypes with an allele frequency above 1% or 5% and compared the true

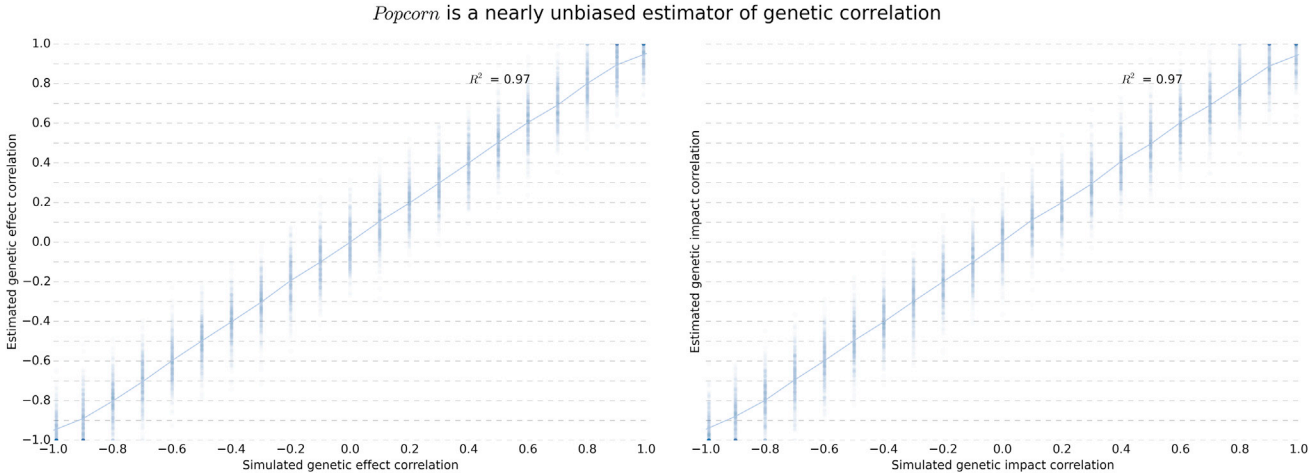


**Figure 1. True and Estimated Genetic-Impact and Genetic-Effect Correlations**
All simulations were conducted with a simulated EUR and EAS heritability of 0.5 with 4,499 simulated EUR and 4,836 simulated EAS individuals at 248,953 SNPs.

and estimated genetic-impact and genetic-effect correlations among all models (Table 1). We found that when only SNPs with a frequency above 5% in both populations were used, the difference in $\rho_{ge}$ and $\rho_{gi}$ was minimal except in the most adversarial cases. Even in the adversarial model, the true difference was only 7%. Although they are unlikely to represent reality, the four nonstandard disease models result in substantial bias in our estimators. When SNPs with an allele frequency above 1% in both populations are included, the differences are more pronounced. This is because the normalizing constant $1/\sigma$ rapidly increases as the SNP becomes more rare. Indeed, as SNPs become more rare, having an accurate disease model becomes increasingly important. Therefore, we proceeded with a 5% MAF cutoff in our analysis of real data and used the notation $h_c^2$ to refer to the heritability of SNPs with an allele frequency above 5% in both populations (the common-SNP heritability). Note, however, that one of the advantages of maximum-likelihood estimation in general is that the likelihood can be reformulated to mimic the disease model of interest.

## Validating Popcorn by Using Gene Expression in GEUVADIS

We compared the common-SNP heritability ($h_c^2$) and genetic-correlation estimates of Popcorn to those of GCTA in the GEUVADIS dataset, for which raw genotypes are publicly available. GEUVADIS consists of RNA-sequencing (RNA-seq) data for 464 lymphoblastoid cell line (LCL) samples from five populations in the 1000 Genomes Project. Of these, 375 are of European ancestry (CEU [Utah residents with ancestry from northern and western Europe from the CEPH collection], FIN [Finnish in Finland], GBR [British in England and Scotland], and TSI [Toscani in Italia]), and 89 are of African ancestry (YRI [Yoruba in Ibadan, Nigeria]). Raw RNA-seq reads obtained from the European Nucleotide Archive (accession number ENA: ERP001942) were aligned to the transcriptome with hg19 coordinates from the UCSC Genome Browser. RSEM[39] was used for estimating the abundances of each annotated isoform, and total gene abundance was calculated as the sum of all isoform abundances normalized to one million total counts or transcripts per million (TPM). For mapping of expression quantitative trait loci (eQTLs), European and Yoruban samples were analyzed separately. For each population, we median normalized TPMs to account for differences in sequencing depth in each sample and standardized to mean 0 and variance 1. Of the 29,763 total genes, 9,350 with TPM $> 2$ in both populations were chosen for this analysis.

For each gene and using 30 principal components as covariates, we conducted a *cis*-eQTL association study at all SNPs within 1 Mb of the gene body and with an allele frequency above 5% in both populations. We found that GCTA and Popcorn agreed on the global distribution of heritability (Figure S5) and that GCTA's estimates of genetic correlation had a similar distribution to Popcorn's estimates of genetic-effect and genetic-impact correlation (Figure 2). Although the number of SNPs and individuals included in each gene analysis is too small for obtaining accurate point estimates of the genetic correlation on a per-gene basis ($N = 464$, $M = 4279.5$), the large number of genes enables accurate estimation of the global mean heritability and genetic correlation.

## Common-SNP Heritability and Genetic Correlation of Gene Expression in GEUVADIS

We found that the average *cis*-$h_c^2$ of the expression of the genes we analyzed was 0.093 (0.002) in EUR and 0.088 (0.002) in YRI. Our estimates are higher than previously reported average *cis*-heritability estimates of 0.055 in whole blood and 0.057 in adipose,[40] which could have arisen for several reasons. First, we removed 68% of the transcripts that are lowly expressed in either YRI or EUR data.

**Table 1. True and Estimated Values of Genetic-Impact and Genetic-Effect Correlations in Simulated EUR-like and EAS-like Genotypes**

| Model | MAF > 0.01 | | | | MAF > 0.05 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho_{ge}$ | $\rho_{gi}$ | $\widehat{\rho}_{ge}$ | $\widehat{\rho}_{gi}$ | $\rho_{ge}$ | $\rho_{gi}$ | $\widehat{\rho}_{ge}$ | $\widehat{\rho}_{gi}$ |
| Independent | 0.500 | 0.478 | 0.500 | 0.460 | 0.500 | 0.488 | 0.509 | 0.469 |
| Inverse | 0.431 | 0.500 | 0.567 | 0.496 | 0.479 | 0.500 | 0.555 | 0.482 |
| Rare | 0.500 | 0.467 | 0.382 | 0.863 | 0.500 | 0.496 | 0.998 | 0.756 |
| Common | 0.500 | 0.500 | 0.522 | 0.493 | 0.500 | 0.500 | 0.502 | 0.496 |
| Difference | 0.500 | 0.416 | 0.354 | 0.435 | 0.500 | 0.461 | 0.410 | 0.412 |
| Adversarial | 0.710 | 0.604 | 0.525 | 0.651 | 0.714 | 0.667 | 0.601 | 0.675 |

Results are the average of 100 simulations with a phenotype heritability of 0.5 in each population.

Second, estimates from RNA-seq analysis of cell lines might not be directly comparable to microarray data from tissue.

The average genetic-effect correlation was 0.320 (0.010), whereas the average genetic-impact correlation was 0.313 (0.010). Notably, the genetic correlation increased as the *cis*-$h^2_c$ of expression in both populations increased (Figure 3). In particular, when the *cis*-$h^2_c$ of the gene was at least 0.2 in both populations, the genetic-effect correlation was 0.772 (0.017), whereas the genetic-impact correlation was 0.753 (0.018).

In order to verify that our analysis did not contain any small-sample-size or conditioning biases, we analyzed the genetic correlation of simulated phenotypes over the GEUVADIS genotypes. We sampled pairs of heritabilities from the distribution of estimated expression heritability and simulated pairs of phenotypes to have the given heritability and a genetic-effect correlation of 0.0 over randomly chosen 4,000 bp regions from chromosome 1 of the GEUVADIS genotypes. Without conditioning, the average estimated genetic-effect correlation was −0.002 (0.003), indicating that the estimator remained unbiased. Furthermore, with conditioning on the heritability estimates above a certain threshold, the average estimated genetic-effect correlation was not significantly different from 0.0 (Figure S6).

We found that although the average genetic correlation was low, the genetic correlation increased with the *cis*-$h^2_c$ of the gene, indicating that as *cis*-genetic regulation of gene expression increases, it does so similarly in both YRI and EUR populations. This could help interpret the recent observation that although the global genetic correlation of gene expression across tissues is low,[40] *cis*-eQTLs tend to replicate across tissues.[41] Because the presence of a *cis*-eQTL indicates substantial *cis*-genetic regulation, an analysis of eQTL replication across tissues implicitly conditions on a high heritability of gene expression and therefore might indicate a much higher genetic correlation than the average.

### Summary Statistics of RA and T2D

Finally, we sought to examine the transethnic $\rho_{gi}$ and $\rho_{ge}$ in RA and T2D cohorts for which raw genotypes are not available. We obtained summary statistics from a RA GWAS of 58,284 individuals of European descent and 22,515 individuals of East Asian descent[8] and from T2D GWASs of 69,033 individuals of European descent (DIAGRAM stage 1[42]) and 18,817 individuals of East Asian descent.[43] We used genotypes from 504 East Asian and 503 European individuals sequenced as part of the 1000 Genomes Project as population-specific external reference panels for our EAS and EUR summary statistics, respectively. We removed the major histocompatibility region (chromosome 6, 25–35 Mb) from the RA summary statistics. We estimated the common-SNP heritability and genetic correlation by using 2,539,629 strand-unambiguous SNPs genotyped or imputed in both RA studies and 1,054,079 strand-unambiguous SNPs genotyped or imputed in both T2D studies; all SNPs had an allele frequency above 5% in 1000 Genomes EUR and EAS populations. The $h^2_c$ and genetic-correlation estimates are presented in Table 2. Our RA $h^2_c$ estimates of 0.177 (0.015) and 0.221 (0.026) for EUR and EAS, respectively, are lower than a previously reported mixed-model-based heritability estimate of 0.32 (0.037) in Europeans.[45] Similarly, our T2D $h^2_c$ estimates of 0.242 (0.013) and 0.105 (0.021) for EUR and EAS, respectively, are lower than a previously reported mixed-model-based estimate of 0.51 (0.065) in Europeans.[45] We stress that this discrepancy is most likely due to the correction of genomic control in summary association data (such correction does not affect genetic-correlation estimates[19]) and the difference between common-SNP heritability $h^2_c$ and total narrow-sense heritability $h^2$. Furthermore, estimates of the heritability of T2D from family studies can vary significantly.[46,47]

We found the genetic-effect correlation in RA and T2D to be 0.463 (0.058) and 0.621 (0.088), respectively, and the genetic-impact correlation was not significantly different at 0.455 (0.056) and 0.606 (0.083), respectively. The transethnic genetic-impact and genetic-effect correlations for both phenotypes were significantly different from both 1 and 0 (Table 2), showing that although the phenotypes have clear genetic overlap, the per-allele effect sizes differ significantly between the two populations.

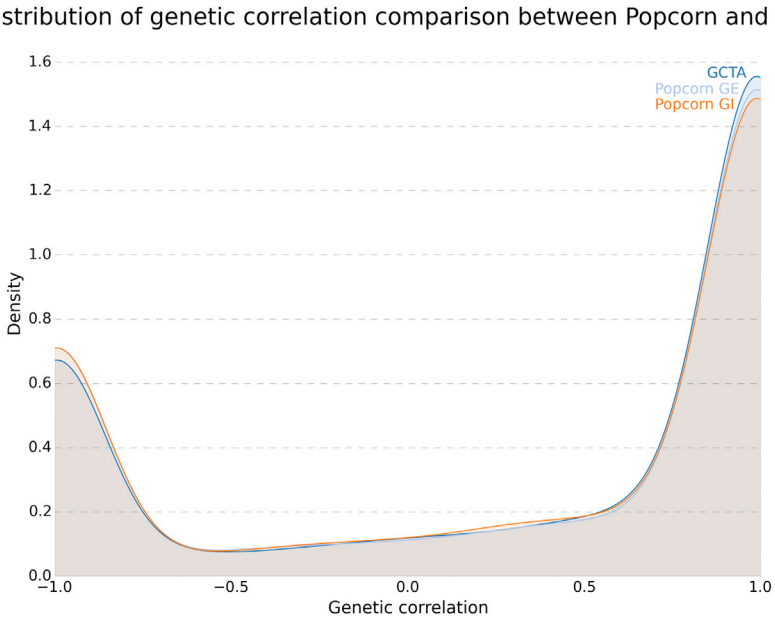


**Figure 2. The Distributions of the Estimates of Genetic Correlation Computed with Popcorn and GCTA Are Compared**
The distribution was computed via Gaussian kernel density estimation on the set of genetic-correlation estimates.

### Summary Statistics of Height and BMI

To further validate that our observations were not a statistical artifact, we used Popcorn to estimate the genetic correlation of one trait in one population across studies and compared it with those of GCTA and LDSC (LD Score). We obtained sex-stratified summary statistics of height and BMI from the GIANT consortium[48] and used Popcorn and LDSC to estimate the genetic correlation of height and BMI. Values for GCTA were taken from Yang al.[49] Scores for Popcorn and LDSC were computed from variants with an allele frequency above 5% in 1000 Genomes European-descent individuals, and genetic correlation was computed with all strand-unambiguous variants with an allele frequency above 5% in HapMap 3 (these are supplied with the summary statistics). Popcorn's sex-stratified genetic correlations of height and BMI were not significantly different from 1.0 or from those of LDSC or GCTA (Table S2).

## Discussion

We have developed transethnic genetic-effect and genetic-impact correlations and provided a method for estimating these quantities on the basis of only summary-level GWAS information and suitable reference panels. We have applied our estimator to several phenotypes: RA, T2D, and gene expression. Although the GEUVADIS dataset lacks enough power for inferring the genetic correlation of single or small subsets of genes, we can make inferences about the global structure of genetic correlation of gene expression. We have found that the global mean genetic correlation is low but that it increases substantially when the heritability is high in both populations. In all phenotypes analyzed, the genetic correlation was significantly different from both 0 and 1. Our results show that global differences in SNP effect size of complex traits can be large. In contrast, effect sizes of gene expression appear to be more conserved where there is strong genetic regulation.

It is not possible to draw conclusions about polygenic selection from estimates of transethnic genetic correlation. The effect sizes can be identical ($\rho_{ge} = 1$) while polygenic selection acts to change only the allele frequencies. Similarly, the effect sizes can be different ($\rho_{ge} < 1$) without selection. Differences in effect sizes at common SNPs can result from many phenomena. We expect that untyped and unimputed variants that are differentially linked to observed SNPs, along with rare or population-specific variants differentially linked to observed SNPs, will contribute significantly. If a gene-gene or gene-environment interaction exists but only marginal effects are tested, the observed marginal effects could be different in each population as a result of differences in allele frequency, even if the interaction effect is the same in both populations, and this will result in decreased genetic correlation. Although within-locus (dominance) interactions might also play a role,[50] the magnitude of this effect has been debated.[51] Statistical noise could also be to blame, given that within-population meta-analyses of the same trait do not always show identical effects; however, we estimated the sex-stratified genetic correlation of height and BMI and found that neither was significantly different from 1, which agrees with previous results.[49] We emphasize that we cannot differentiate between these effects on the basis of this analysis alone, and establishing how much each of these effects contributes to inter-population differences in effect size will require further research.

Estimates of the transethnic genetic correlation are important for several reasons. They might help inform best practices for transethnic meta-analysis and potentially offer improvements over current methods that use $F_{ST}$ to cluster populations for analysis.[4] Further, the transethnic genetic correlation constrains the limit of out-of-sample phenotype predictive power. If the maximum within-population correlation of predicted phenotype $P$ to true phenotype $Y$ is $\rho_{YP}^{\max} = \sqrt{h_1^2}$, then the maximum out-of-population correlation is $\rho_{YP}^{\max} = \rho_{ge}\sqrt{h_1^2}$ (Appendix A). Our observation that the genetic correlation is low for

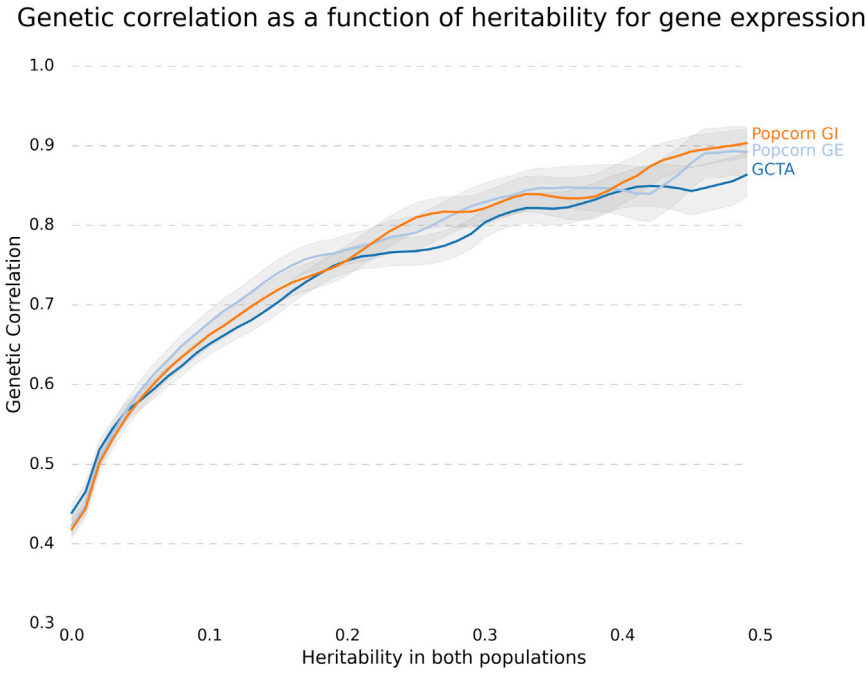


**Figure 3. Genetic Correlation as a Function of Heritability for Gene Expression**
The mean and SE of the genetic correlation of the set of genes with $h_1^2$ and $h_2^2$ exceeding threshold $h$ in each analysis (y axis) are plotted against $h$ (x axis).

RA, T2D, and gene expression shows that out-of-population phenotypic predictive power is quite low. Similarly, it implies that assessing disease risk in non-Europeans on the basis of current GWAS results might be problematic; gaining insight into differences in genetic architecture and improving risk assessment will necessitate increased study of disease in many populations.

Although the genetic correlation of multiple phenotypes in one population has a relatively straightforward definition, extending this to multiple populations motivates multiple possible extensions. In this work, we have provided estimators for the correlation of genetic effect and genetic impact, but other quantities related to the shared genetics of complex traits between populations include the correlation of explained variance, $\rho_{ge} = \mathrm{Cor}(\sigma_1^2\beta_1^2, \sigma_2^2\beta_2^2)$, and the proportion of causal variants shared by the two populations. Interestingly, although our goal was to construct an estimator that determines the extent of genetic sharing independently of allele frequency, we have observed that the correlations of genetic effect and genetic impact are similar. Furthermore, our simulations show that under a random-effects model utilizing only SNPs with an allele frequency above 5% in both populations, the true genetic-effect and genetic-impact correlations are similar. We conclude that at variants common in both populations, differences in effect size and not allele frequency are driving the transethnic phenotypic differences in these traits.

Our approach to estimating genetic correlation has two major advantages over mixed-model-based approaches. First, utilizing summary statistics allows application of the method without data-sharing and privacy concerns that come with raw genotypes. Second, our approach is linear in the number of SNPs and thus avoids the computational bottleneck required for estimating the genetic relationship matrix. Conceptually, our approach is very similar to that taken by LD-score regression. Indeed, the diagonal of the LD-matrix product in one population is exactly the LD scores ($\sum_{1ii}^2 = l_i$). One could ignore our likelihood-based approach and define cross-population scores as $c_i = \sum_m r_{1im} r_{2im}$ in order to exploit the linear relationship $\mathbb{E}[Z_{1i}Z_{2i}] = (\sqrt{N_1N_2}/M)\rho_{gi}\sqrt{h_1^2h_2^2}c_i$ (a similar approach can be taken for the genetic-effect correlation). Given that LD-score regression has been successfully used for computing the genetic correlation of two phenotypes in a single population, this derivation can be viewed as an extension of LD-score regression to one phenotype in two different populations. The main difference in our approach is choosing maximum likelihood rather than regression in order to fit the model. A comparison of our method to the LDSC software shows that they perform similarly as heritability estimators (Figure S7).

Of course, our method is not without drawbacks. First, it requires a large sample size and large number of SNPs to achieve SEs low enough for accurate estimation. Until recently, large-sample GWASs have been rare in non-European populations, although they are becoming more common. Similarly, the quality of reference panels could suffer in non-European populations, which could affect downstream analysis.[52] Second, our method is limited to analyzing relatively common SNPs, both because having an accurate disease model is important for the analysis of rare variants and because estimates of effect size and correlation coefficients have a high SE at rare SNPs.[17] Third, our analysis is currently limited to SNPs that are present in both populations. Indeed, it is currently unclear how best to handle population-specific variants in this framework. Fourth, our estimator of $\rho$ is bounded between $-1$ and 1. This could induce bias when the true value is close to the boundary and the sample size is small. Fifth and finally, admixed populations induce very long-range LD that is not accounted for in our approach, and we are therefore limited to unadmixed populations.[17]

Our analysis leaves open several avenues for future work. We approximately maximize the likelihood of an $M \times M$

**Table 2. Heritability and Genetic Correlation of RA and T2D between EUR and EAS Populations**

| | | $h_{EUR}^2$ Liability | $h_{EAS}^2$ Liability | $\rho_{ge}$ | $\rho_{gi}$ |
|---|---|---|---|---|---|
| RA | estimate (SE) | 0.18 (0.02) | 0.22 (0.03) | 0.46 (0.06) | 0.46 (0.06) |
| | 95% CI | [0.15, 0.21] | [0.16, 0.28] | [0.34, 0.58] | [0.34, 0.58] |
| | $p > 0$ | 3.90e−32 | 1.89e−17 | 1.37e−15 | 8.16e−16 |
| | $p < 1$ | 0.0 | 3.1e−197 | 2.53e−20 | 4.87e−22 |
| T2D | estimate (SE) | 0.24 (0.01) | 0.11 (0.02) | 0.62 (0.09) | 0.61 (0.08) |
| | 95% CI | [0.22, 0.26] | [0.07, 0.15] | [0.44, 0.80] | [0.45, 0.77] |
| | $p > 0$ | 2.41e−77 | 5.73e−7 | 1.70e−12 | 2.85e−13 |
| | $p < 1$ | 0.0 | 0.0 | 1.066e−5 | 2.06e−6 |

EUR RA data contained 8,875 case and 29,367 control subjects for a study prevalence of 0.23. EAS RA data contained 4,873 case and 17,642 control subjects for a study prevalence of 0.22. RA prevalence was assumed to be 0.5% in both populations.[8] T2D EUR data contained 12,171 case and 56,862 control subjects for a study prevalence of 0.18. T2D EAS data contained 6,952 case and 11,865 control subjects for a study prevalence of 0.37. T2D EUR prevalence was assumed to be 8%,[42] whereas T2D EAS prevalence was assumed to be 9%.[44] CI, confidence interval.

multivariate normal distribution via a method that uses only the diagonal elements of each block and discards covariance information between $Z$ scores. A better approximation might lower the SE of the estimator, facilitating an analysis of the genetic correlation of functional categories, pathways, and genetic regions. We would also like to extend our analysis to include population-specific variants and variants with frequencies from 1% to 5% or lower than 1%. Our simulations indicate that having an accurate disease model is important for determining the difference between genetic-effect and genetic-impact correlations when rare variants are included. Maximum-likelihood approaches are well suited to different genetic architectures. For example, one could estimate both the global relationship between allele frequency and effect size and the global relationship between per-SNP $F_{ST}$ and genetic correlation by incorporating parameters $\alpha$ and $\Upsilon$ into the prior distribution of the effect sizes:

$$\beta_{1i}, \beta_{2i} \sim \mathcal{N}\left(0, \begin{bmatrix} h_1^2 \sigma_{1i}^{\alpha} & \rho_{ge}\sqrt{h_1^2 h_2^2} F_{STi}^{\gamma} \\ \rho_{ge}\sqrt{h_1^2 h_2^2} F_{STi}^{\gamma} & h_2^2 \sigma_{1i}^{\alpha} \end{bmatrix}\right).$$

We expect that incorporating these parameters will improve estimates of heritability and genetic correlation while revealing important biological insights.

## Appendix A

Consider two GWASs of a phenotype conducted in different populations. Assume we have $N_1$ individuals genotyped or imputed to $M$ SNPs in study 1 and $N_2$ individuals genotyped or imputed to $M$ SNPs in study 2. Let $X_1$ and $X_2$ and $Y_1$ and $Y_2$ be the matrices of mean-centered genotypes and phenotypes, respectively, of the individuals in studies 1 and 2, respectively. Let $f_1$ and $f_2$ be the allele frequencies of the $M$ SNPs common to both populations. If we assume Hardy-Weinberg equilibrium, the allele variances are $\sigma_1^2 = 2f_1(1-f_1)$ and $\sigma_2^2 = 2f_2(1-f_2)$. Let $\beta_1$ and $\beta_2$ be the (unobserved) per-allele effect size for each SNP in studies 1 and 2, respectively. Define the genetic-impact correlation as $\rho_{gi} = \mathrm{Cor}(\sqrt{\sigma_1^2}\beta_1, \sqrt{\sigma_2^2}\beta_2)$ and the genetic-effect correlation as $\rho_{ge} = \mathrm{Cor}(\beta_1, \beta_2)$. We present a maximum-likelihood framework for estimating the heritability of the phenotype in study 1 and its SE, the heritability of the phenotype in study 2 and its SE, and the genetic-effect and genetic-impact correlations of the phenotype between the studies and their SEs given only the summary statistics $Z_1$ and $Z_2$ and reference genotypes $G_1$ and $G_2$ representing the populations in the studies. We assume that genotypes are drawn randomly from populations with expected correlation matrices $\Sigma_1$ and $\Sigma_2$ and that every SNP is causal with a normally distributed effect size (although this assumption is not necessary in practice; see Figure S1).

### Genetic-Impact Correlation

Let $X_1' = X_1/\sqrt{\sigma_1^2}$ (and similarly for study 2) be normalized genotype matrices. We consider the standard linear model for the generation of phenotypes, where $Y_1 = X_1'\beta_1 + \varepsilon_1$ and $Y_2 = X_2'\beta_2 + \varepsilon_2$.

For convenience of notation, let $h_{ix} = \rho_{gi}\sqrt{h_1^2 h_2^2}$. We assume that the SNP effects follow the infinitesimal model, where every SNP has an effect size drawn from the normal distribution, and that the residuals are independent for each individual and normally distributed:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{M}\begin{bmatrix} h_1^2 \mathbb{I}_M & h_{ix}\mathbb{I}_M \\ h_{ix}\mathbb{I}_M & h_2^2 \mathbb{I}_M \end{bmatrix}\right) \quad \text{(Equation A1)}$$

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \left(1-h_1^2\right)\mathbb{I}_M & 0 \\ 0 & \left(1-h_2^2\right)\mathbb{I}_M \end{bmatrix}\right), \quad \text{(Equation A2)}$$

where $h_1^2$ and $h_2^2$ are the heritability of the disease in studies 1 and 2, respectively, and $\rho_{gi}$ is the genetic-impact correlation.

Using the above model, we compute the distribution of the observed $Z$ scores as a function of the reference-panel correlations and the model parameters ($h_1^2$, $h_2^2$, and $\rho_{gi}$). Given a distribution for $Z$ and an observation of $Z$, we can then choose parameters that give the highest probability of observing $Z$. First, we compute the distribution of $Z$. It is well known that the $Z$ scores of a linear regression are normally distributed given $\beta$ when the sample size is large enough. Because $\mathbb{P}(Z) \propto \mathbb{P}(Z \mid \beta)\mathbb{P}(\beta)$ and the product of normal distributions is normal, we need to compute only the unconditional mean and variance of $Z$ to know its distribution. Specifically, let $Z = [Z_1^{\mathrm{T}}, Z_2^{\mathrm{T}}]^{\mathrm{T}}$. Then, its mean is

$$\mathbb{E}[Z] = \mathbb{E}\begin{bmatrix} \frac{X_1'^{\mathrm{T}} Y_1}{\sqrt{N_1}} \\ \frac{X_2'^{\mathrm{T}} Y_2}{\sqrt{N_2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{N_1}}\left(\mathbb{E}\left[X_1'^{\mathrm{T}} X_1'\right]\mathbb{E}[\beta_1] + \mathbb{E}\left[X_1'^{\mathrm{T}}\right]\mathbb{E}[\varepsilon_1]\right) \\ \frac{1}{\sqrt{N_2}}\left(\mathbb{E}\left[X_2'^{\mathrm{T}} X_2'\right]\mathbb{E}[\beta_2] + \mathbb{E}\left[X_2'^{\mathrm{T}}\right]\mathbb{E}[\varepsilon_2]\right) \end{bmatrix} = 0.$$

The within-population variance is

$$\mathrm{Cov}\left[Z_{1i}, Z_{1j}\right] = \mathbb{E}\left[Z_{1i} Z_{1j}\right] = \mathbb{E}_{X,\beta,\varepsilon}\left[\mathbb{E}\left[Z_{1i} Z_{1j} \mid X, \beta, \varepsilon\right]\right]$$

$$= \frac{1}{N_1}\mathbb{E}_{X,\beta,\varepsilon}\left[X_{1i}'^{\mathrm{T}}\left(X_1'\beta_1 + \varepsilon_1\right)\left(X_1'\beta_1 + \varepsilon_1\right)^{\mathrm{T}} X_{1j}'\right]$$

$$= \frac{1}{N_1}\mathbb{E}_{X,\beta}\left[X_{1i}'^{\mathrm{T}} X_1' \beta_1 \beta_1^{\mathrm{T}} X_1^{\mathrm{T}} X_{2j}\right] + \frac{1}{N_1}\mathbb{E}_{X,\varepsilon}\left[X_{1i}'^{\mathrm{T}} \varepsilon_1 \varepsilon_1^{\mathrm{T}} X_{1j}'\right]$$

$$= \frac{h_1^2}{MN_1}\mathbb{E}_X\left[X_{1i}'^{\mathrm{T}} X_1' X_1'^{\mathrm{T}} X_{1j}'\right] + \frac{1 - h_1^2}{N_1}\mathbb{E}_X\left[X_{1i}'^{\mathrm{T}} X_{1j}'\right]$$

$$= \frac{h_1^2}{MN_1}\left(N_1 M r_{1ij} + N_1 \sum_{m=1}^{M} r_{1im} r_{1jm} + N_1^2 \sum_{m=1}^{M} r_{1im} r_{1jm}\right) + \frac{1 - h_1^2}{N_1} r_{1ij}$$

$$= r_{1ij} + \frac{N_1 + 1}{M} h_1^2 \Sigma_{1(i)} \Sigma_1^{(j)},$$

where $r_{pij} = \Sigma_{pij}$ is the correlation coefficient of SNP $i$ and $j$ in population $p$. Similarly, the between-population variance is

$$\mathrm{Cov}\left[Z_{1i}, Z_{2j}\right] = \frac{1}{\sqrt{N_1 N_2}}\mathbb{E}_{X,\beta}\left[X_{1i}'^{\mathrm{T}} X_1' \beta_1 \beta_2^{\top} X_2'^{\mathrm{T}} X_{2j}'\right] + \frac{1}{\sqrt{N_1 N_2}}\mathbb{E}_{X,\varepsilon}\left[X_{1i}'^{\top} \varepsilon_1 \varepsilon_2^{\top} X\prime_{2j}\right]$$

$$= \frac{h_{\mathrm{ix}}}{M\sqrt{N_1 N_2}}\mathbb{E}_X\left[X_{1i}'^{\top} X_1' X_2'^{\mathrm{T}} X_{2j}'\right]$$

$$= \frac{h_{\mathrm{ix}}}{M\sqrt{N_1 N_2}}\left(N_1 N_2 \sum_{m=1}^{M} r_{1im} r_{2jm}\right)$$

$$= \frac{\sqrt{N_1 N_2}}{M} h_{\mathrm{ix}} \Sigma_{1(i)} \Sigma_2^{(j)},$$

where $\Sigma_{(i)}$ denotes the $i^{\mathrm{th}}$ row of $\Sigma$, and $\Sigma^{(j)}$ denotes the $j^{\mathrm{th}}$ column. The covariance of the $Z$ scores is thus

$$C = \mathrm{Var}(Z) = \begin{bmatrix} \Sigma_1 + \frac{N_1 + 1}{M} h_1^2 \Sigma_1^2 & h_{\mathrm{ix}} \frac{\sqrt{N_1 N_2}}{M} \Sigma_1 \Sigma_2 \\ h_{\mathrm{ix}} \frac{\sqrt{N_1 N_2}}{M} \Sigma_2 \Sigma_1 & \Sigma_2 + \frac{N_2 + 1}{M} h_2^2 \Sigma_2^2 \end{bmatrix} \quad \text{(Equation A3)}$$

and $Z \sim \mathcal{N}(0, C)$.

**Genetic-Effect Correlation**

Let $h_{\mathrm{ex}} = \rho_{\mathrm{ge}}\sqrt{h_1^2 h_2^2}$. We modify the procedure above to use mean-centered instead of normalized genotype matrices and model the distribution of the effect sizes as

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{h_1^2}{\|\sigma_1^2\|_1}\mathbb{I}_M & \frac{h_{\mathrm{ex}}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}}\mathbb{I}_M \\ \frac{h_{\mathrm{ex}}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}}\mathbb{I}_M & \frac{h_2^2}{\|\sigma_2^2\|_1}\mathbb{I}_M \end{bmatrix}\right). \quad \text{(Equation A4)}$$

Notice that a linear model with effect sizes acting on unnormalized genotypes is the same as a linear model with effect sizes acting on normalized genotypes under the substitution $\beta_{1,2} \rightarrow \sqrt{\sigma_{1,2}^2}\beta_{1,2}$. Therefore, the covariance of $Z$ scores on the per-allele scale can be immediately inferred from the prior derivation:

$$C = \mathrm{Var}(Z) = \begin{bmatrix} \Sigma_1 + \frac{N_1 + 1}{\|\sigma_1^2\|_1} h_1^2 \Sigma_1 \sigma_1^2 \Sigma_1 & h_{\mathrm{ex}} \frac{\sqrt{N_1 N_2}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}} \Sigma_1 \sqrt{\sigma_1^2 \sigma_2^2} \Sigma_2 \\ h_{\mathrm{ex}} \frac{\sqrt{N_1 N_2}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}} \Sigma_2 \sqrt{\sigma_2^2 \sigma_1^2} \Sigma_1 & \Sigma_2 + \frac{N_2 + 1}{\|\sigma_2^2\|_1} h_2^2 \Sigma_2 \sigma_2^2 \Sigma_2 \end{bmatrix}.$$

### Approximate Maximum-Likelihood Estimation

Let $C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ be either of the above covariance matrices written in block form. We approximately optimize the above likelihood as follows: first, we find $h_1^2$ and $h_2^2$ by maximizing the likelihood corresponding to $C_{11}$ and $C_{22}$, and then we find $\rho_{\mathrm{gi}}$ or $\rho_{\mathrm{ge}}$ by maximizing the likelihood corresponding to $C_{12}$:

$$l\left(h_1^2 \mid Z_1, \Sigma, \sigma\right) \approx -\sum_{i=1}^{M} w_{11i}\left(\ln(C_{11ii}) + \frac{Z_{1i}^2}{C_{11ii}}\right)$$

$$l\left(h_2^2 \mid Z_2, \Sigma, \sigma\right) \approx -\sum_{i=1}^{M} w_{22i}\left(\ln(C_{22ii}) + \frac{Z_{2i}^2}{C_{22ii}}\right)$$

$$l\left(\rho_{\mathrm{g\{i,e\}}} \mid Z, \widehat{h_1^2}, \widehat{h_2^2}, \Sigma, \sigma\right) \approx -\sum_{i=1}^{M} w_{12i}\left(\ln(C_{12ii}) + \frac{Z_{1i}Z_{2i}}{C_{12ii}}\right).$$

Because we are discarding between-SNP covariance information ($\mathrm{Cov}(Z_{1i}, Z_{1j})$), highly correlated SNPs will be over-counted in our approximate likelihood. As a simple example, notice that two SNPs in perfect LD will each contribute identical terms to the approximate likelihood and therefore should be downweighted by a factor of 1/2. The extent to which SNP $i$ is over-counted is exactly the $i^{\mathrm{th}}$ entry in its corresponding LD-matrix product. Therefore, we let $w_{jki}^{\mathrm{gi}} = 1/(\Sigma_j\Sigma_k)_{ii}$ and $w_{jki}^{\mathrm{ge}} = 1/(\Sigma_j\sqrt{\sigma_j^2\sigma_k^2}\,\Sigma_k)_{ii}$ to reduce the variance in our estimates of the parameters $h_1^2$, $h_2^2$, $\rho_{\mathrm{gi}}$, and $\rho_{\mathrm{ge}}$.

Furthermore, rather than compute the full products $\Sigma_1^2$, $\Sigma_2^2$, and $\Sigma_1\Sigma_2$ over all $M$ SNPs in the genome, we choose a window size $W$ and approximate the product by $(\Sigma_a\Sigma_b)_{ii} = \sum_{w=i-W}^{w=i+W} r_{aiw}r_{biw}$. Although maximum-likelihood estimation admits a straightforward estimate of the SE via the fisher information, we found these estimates to be inaccurate in practice. Instead, we use a block jackknife with a block size equal to $\min(100, (M/200))$ SNPs to ensure that blocks are large enough for the removal of residual correlations.

### Out-of-Population Prediction of Phenotypic Values

Consider using the results of a GWAS with perfect power in population 2 to predict the phenotypic values of a set of individuals from population 1. This defines the upper limit of the correlation of true and predicted phenotypic values. Let the true values of the effect sizes in population 2 be $\beta_2$. Let the true phenotypes in population 1 be $Y = X_1\beta_1 + \varepsilon_1$ and the predicted phenotypes be $P = X_1\beta_2$. We are interested in the correlation of the predicted and true phenotypes $\rho_{YP}^{\mathrm{MAX}} = \mathrm{Cor}(Y, P)$. Notice that given $X$, the true and predicted phenotype of each individual is an affine transformation of a multivariate normal random variable ($\beta$):

$$\begin{bmatrix} Y_i \\ P_i \end{bmatrix} = \begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_i \\ 0 \end{bmatrix}.$$

Therefore, $(Y_i, P_i)$ for individual $i$ is multivariate normal with the expected covariance matrix

$$\mathbb{E}_X[\mathrm{Cov}(Y_i, P_i)] = \mathbb{E}_X\begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix} \times \begin{bmatrix} \frac{1}{\|\sigma_1^2\|_1}\mathbb{I}_M & \frac{h_{\mathrm{ex}}}{\sqrt{\|\sigma_1^2\|_1\|\sigma_2^2\|_1}}\mathbb{I}_M \\ \frac{h_{\mathrm{ex}}}{\sqrt{\|\sigma_1^2\|_1\|\sigma_2^2\|_1}}\mathbb{I}_M & \frac{h_2^2}{\|\sigma_2^2\|_1}\mathbb{I}_M \end{bmatrix}\begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix}^{\mathrm{T}}$$

$$= \mathbb{E}_X\begin{bmatrix} \frac{\sum_m X_{im}^2}{\|\sigma_1^2\|_1} & \frac{h_{\mathrm{ex}}\sum_m X_{im}^2}{\sqrt{\|\sigma_1^2\|_1\|\sigma_2^2\|_1}} \\ \frac{h_{\mathrm{ex}}\sum_m X_{im}^2}{\sqrt{\|\sigma_1^2\|_1\|\sigma_2^2\|_1}} & \frac{h_2^2\sum_m X_{im}^2}{\|\sigma_2^2\|_1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & h_{\mathrm{ex}}\sqrt{\frac{\|\sigma_1^2\|_1}{\|\sigma_2^2\|_1}} \\ h_{\mathrm{ex}}\sqrt{\frac{\|\sigma_2^2\|_1}{\|\sigma_1^2\|_1}} & h_2^2\frac{\|\sigma_1^2\|_1}{\|\sigma_2^2\|_1} \end{bmatrix}.$$

Therefore, the expected correlation $\mathbb{E}[\mathrm{Cor}(Y_i, P_i)]$ is

$$\frac{h_{\mathrm{ex}}}{\sqrt{h_2^2}}\sqrt{\frac{\|\sigma_1^2\|_1}{\|\sigma_2^2\|_1}\frac{\|\sigma_2^2\|_1}{\|\sigma_1^2\|_1}} = \rho_{\mathrm{ge}}\sqrt{h_1^2}.$$

The expected population correlation tends to the sample correlation as the number of samples increases; therefore,

$$\rho_{YP}^{\mathrm{MAX}} = \mathrm{Cor}(Y, P) \to \rho_{\mathrm{ge}}\sqrt{h_1^2} \qquad \text{(Equation A5)}$$

as $N \to \infty$.

### Supplemental Data

Supplemental Data include six figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.05.001.


### Acknowledgments

The authors would like to acknowledge Lior Pachter, Hilary Finucane, and Yukinori Okada for insightful discussion about the problem. B.C.B. is supported by the National Science Foundation Graduate Research Fellowship Program. A.L.P. is supported by NIH grant R01 HG006399. N.Z. is supported by NIH grant K25HL121295.

Received: December 19, 2015
Accepted: May 3, 2016
Published: June 16, 2016

## Web Resources

DIAGRAM stage 1 data, http://diagram-consortium.org/downloads.html

Popcorn, https://github.com/brielin/popcorn

Summary statistics of height and BMI, https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

Summary statistics of RA, http://plaza.umin.ac.jp/~yokada/datasource/software.htm

## References


1. Robinson, M.R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J.E., Vinkhuyzen, A., Berndt, S.I., Gustafsson, S., et al. (2015). Population genetic differentiation of height and body mass index across Europe. Nat. Genet. *47*, 1357–1362.
2. Burt, V.L., Whelton, P., Roccella, E.J., Brown, C., Cutler, J.A., Higgins, M., Horan, M.J., and Labarthe, D. (1995). Prevalence of hypertension in the US adult population. Results from the Third National Health and Nutrition Examination Survey, 1988-1991. Hypertension *25*, 305–313.
3. Coram, M.A., Candille, S.I., Duan, Q., Chan, K.H.K., Li, Y., Kooperberg, C., Reiner, A.P., and Tang, H. (2015). Leveraging Multi-ethnic Evidence for Mapping Complex Traits in Minority Populations: An Empirical Bayes Approach. Am. J. Hum. Genet. *96*, 740–752.
4. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.
5. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.
6. Oh, S.S., Galanter, J., Thakur, N., Pino-Yanes, M., Barcelo, N.E., White, M.J., de Bruin, D.M., Greenblatt, R.M., Bibbins-Domingo, K., Wu, A.H.B., et al. (2015). Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. PLoS Med. *12*, e1001918.
7. Coronary Artery Disease (C4D) Genetics Consortium (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat. Genet. *43*, 339–344.
8. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al.; RACI consortium; GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
9. de Candia, T.R., Lee, S.H., Yang, J., Browning, B.L., Gejman, P.V., Levinson, D.F., Mowry, B.J., Hewitt, J.K., Goddard, M.E., O'Donovan, M.C., et al.; International Schizophrenia Consortium; Molecular Genetics of Schizophrenia Collaboration (2013). Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. Am. J. Hum. Genet. *93*, 463–470.
10. Lee, S., Teslovich, T.M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet. *93*, 42–53.
11. Palla, L., and Dudbridge, F. (2015). A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. Am. J. Hum. Genet. *97*, 250–259.
12. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
13. Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D.P., Patterson, N., and Price, A.L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics *30*, 2906–2914.
14. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.
15. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. Bioinformatics *31*, i206–i213.
16. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.
17. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
18. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
19. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
20. Park, D.S., Brown, B., Eng, C., Huntsman, S., Hu, D., Torgerson, D.G., Burchard, E.G., and Zaitlen, N. (2015). AdaptMix: learning local genetic correlation structure improves summary statistics-based analyses. Bioinformatics *31*, i181–i189.
21. Xu, Z., Duan, Q., Yan, S., Chen, W., Li, M., Lange, E., and Li, Y. (2015). DISSCO: direct imputation of summary statistics allowing covariates. Bioinformatics *31*, 2434–2442.
22. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
23. Zuo, L., Zhang, C.K., Wang, F., Li, C.-S.R., Zhao, H., Lu, L., Zhang, X.-Y., Lu, L., Zhang, H., Zhang, F., et al. (2011). A novel, functional and replicable risk gene region for alcohol dependence identified by genome-wide association study. PLoS ONE *6*, e26726.
24. Fesinmeyer, M.D., North, K.E., Ritchie, M.D., Lim, U., Franceschini, N., Wilkens, L.R., Gross, M.D., Bůžková, P., Glenn, K.,

Quibrera, P.M., et al. (2013). Genetic risk factors for BMI and obesity in an ethnically diverse population: results from the population architecture using genomics and epidemiology (PAGE) study. Obesity (Silver Spring) *21*, 835–846. http://dx.doi.org/10.1002/oby.20268.

25. Chang, M.H., Ned, R.M., Hong, Y., Yesupriya, A., Yang, Q., Liu, T., Janssens, A.C.J.W., and Dowling, N.F. (2011). Racial/ethnic variation in the association of lipid-related genetic variants with blood lipids in the US adult population. Circ Cardiovasc Genet *4*, 523–533.
26. Waters, K.M., Stram, D.O., Hassanein, M.T., Le Marchand, L., Wilkens, L.R., Maskarinec, G., Monroe, K.R., Kolonel, L.N., Altshuler, D., Henderson, B.E., and Haiman, C.A. (2010). Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet. *6*, e1001078.
27. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540–2542.
28. Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to quantitative genetics (Essex, England: Longman).
29. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
30. 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. Nat. Genet. *42*, 565–569.
31. So, H.-C., Li, M., and Sham, P.C. (2011). Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. Genet. Epidemiol. *35*, 447–456.
32. Vattikuti, S., Guo, J., and Chow, C.C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet. *8*, e1002637.
33. 't Hoen, P.A.C., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F.J., Buermans, H.P.J., Karlberg, O., Brännvall, M., et al.; GEUVADIS Consortium (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. *31*, 1015–1022.
34. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.
35. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
36. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
37. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.
38. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al.; LifeLines Cohort Study (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.
39. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.
40. Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. *7*, e1001317.
41. Gaffney, D.J. (2013). Global properties and functional complexity of human gene regulatory variation. PLoS Genet. *9*, e1003501.
42. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.
43. Cho, Y.S., Chen, C.-H., Hu, C., Long, J., Ong, R.T., Sim, X., Takeuchi, F., Wu, Y., Go, M.J., Yamauchi, T., et al.; DIAGRAM Consortium; MuTHER Consortium (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat. Genet. *44*, 67–72.
44. Ma, R.C.W., and Chan, J.C.N. (2013). Type 2 diabetes in East Asians: similarities and differences with populations in Europe and the United States. Ann. N Y Acad. Sci. *1281*, 64–91.
45. Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A.S., et al.; Diabetes Genetics Replication and Meta-analysis Consortium; Myocardial Infarction Genetics Consortium (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. *44*, 483–489.
46. Marenberg, M.E., Risch, N., Berkman, L.F., Floderus, B., and de Faire, U. (1994). Genetic susceptibility to death from coronary heart disease in a study of twins. N. Engl. J. Med. *330*, 1041–1046.
47. Nora, J.J., Lortscher, R.H., Spangler, R.D., Nora, A.H., and Kimberling, W.J. (1980). Genetic–epidemiologic study of early-onset ischemic heart disease. Circulation *61*, 503–508.
48. 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.; DIAGRAM Consortium; MAGIC Investigators (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. PLoS Genet. *9*, e1003500.
49. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., Esko, T., Milani, L., et al.; Lifelines Cohort Study (2015). Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. Hum. Mol. Genet. *24*, 7445–7449.
50. Chen, X., Kuja-Halkola, R., Rahman, I., Arpegård, J., Viktorin, A., Karlsson, R., Hägg, S., Svensson, P., Pedersen, N.L., and

Magnusson, P.K.E. (2015). Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. Am. J. Hum. Genet. *97*, 708–714.

51. Zhu, Z., Bakshi, A., Vinkhuyzen, A.A.E., Hemani, G., Lee, S.H., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., Esko, T., Milani, L., et al.; LifeLines Cohort Study (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. Am. J. Hum. Genet. *96*, 377–385.
52. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. *11*, 499–511.