Transethnic genetic correlation estimates from summary statistics

Brielin C. Brown, Asian Genetic Epidemiology Network-Type 2 Diabetes (AGEN-T2D) Consortium, Chun Jimmie Ye, Alkes L. Price, Noah Zaitlen

Abstract

The increasing number of genetic association studies conducted in multiple populations provides 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 develop a method for estimating the *transethnic genetic correlation*: the correlation of causal variant effect sizes at SNPs common in populations. We take advantage of the entire spectrum of SNP associations and use only summary-level GWAS data. 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 apply our method to gene expression, rheumatoid arthritis, and type-two diabetes data and overwhelmingly find that the genetic correlation is 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 due to a combination of genetic and environmental differences. For example, northern Europeans are on average taller than southern Europeans[1] and African Americans have an increased rate of hypertension relative to European Americans[2]. The genetic contribution to population phenotypic differentiation is driven by differences in causal allele frequencies, effect sizes, and genetic architectures. 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?

While the vast majority of GWAS have been conducted in European populations[5], the growing number of non-European and multi-ethnic studies[4,6,7] provide an opportunity to study genetic effect distributions 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[8]. While 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 to estimating genetic correlation across populations that uses only summary level GWAS data. Similar to other recent summary statistics based methods[9–20], 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[21–25] we utilize the entire spectrum of GWAS associations while accounting for LD in order 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[18,26]. In multiple populations, differences in allele frequency motivate multiple possible definitions of genetic correlation. Because a variant may have a higher effect size but lower frequency in one population, we consider both the correlation of allele effect sizes as well as the correlation of allelic impact. We define the *transethnic genetic effect correlation* (ρ_{ge}, previously defined by Lee et al[26] and implemented in GCTA) as the correlation coefficient of the per-allele SNP effect sizes, and the *transethnic genetic impact correlation* (ρ_{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, while the genetic impact correlation gives more weight to common alleles than 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 correlation of effect sizes (the genetic effect correlation ρ_{ge}) is 1, but this provides an incomplete picture of the relationship between the two populations, as the allele has a much bigger impact on the distribution of the phenotype in population 2. Therefore, we define the genetic impact correlation ρ_{gi} as the correlation of effect sizes after normalizing genotypes to have mean 0 and variance 1. In our hypothetical case $\rho_{gi} < \rho_{ge}$, however 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 ρ_{ge} may be much less than 1, but the impact of the allele in the

two populations will be similar and therefore ρ_{gi} will be close to 1. While 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[27]). While unlikely to represent reality, this model has been used successfully in practice[8,16,17,28,29]. The infinitesimal assumption yields a multivariate normal distribution on the observed test statistics (Z-scores), which is a function of the heritability and genetic correlation. Rather than pruning SNPs in LD[10,30,31], 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*. Though 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 standard error 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 apply *popcorn* to European and Yoruban gene expression data[32] as well as GWAS summary statistics from European and East Asian rheumatoid arthritis and type-two diabetes cohorts[,33,34]. Our analysis of gEUVADIS shows that our summary statistic based estimator is concordant with the mixed model based estimator. We find that the mean transethnic genetic correlation across all genes is low (ρ_{ge}= 0.320 (0.009)), but increases substantially when the gene is highly heritable in both populations (ρ_{ge}= 0.772 (0.017)). in RA and T2D, we find the genetic effect correlation to be 0.463 (0.058) and 0.621 (0.088), respectively.

Across all phenotypes considered, we overwhelmingly find that the transethnic genetic correlation is significantly less than one. This observation highlights the need to study phenotypes in multiple populations as it implies that, up to the effects of un-observed variants, effect sizes at common SNPs tend to differ between populations. This indicates that GWAS results may not transfer between populations, and therefore disease risk prediction in non-Europeans based on current GWAS results may be problematic,

necessitating a multi-population approach to gain insight into inter-population differences in the genetic architecture of complex traits.

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 Σ_1^2, Σ_2^2 and $\Sigma_1\Sigma_2$, then, using these estimates, we find the maximum likelihood values and estimate standard errors of the parameters of interest: h_1^2, h_2^2 and ρ_{ge} or ρ_{gi}. The details follow.

Consider GWAS of a phenotype conducted in two different populations. Assume we have N_1 individuals genotyped on M SNPs in study one and N_2 individuals genotyped on the same SNPs study two. Let X_1, X_2 be the matrices of mean-centered genotypes in study one and study two, respectively, and let Y_1, Y_2 be their normalized phenotypes. Let f_1, f_2 be vectors of the allele frequencies of the M SNPs common to both populations. Assuming Hardy-Weinberg equilibrium within each population separately, the allele variances are $\sigma_1^2 = 2f_1(1 - f_1)$, $\sigma_2^2 = 2f_2(1 - f_2)$. Let β_1, β_2 be the (unobserved) per-allele effect sizes for each SNP in studies one and two, respectively. The heritability in study one is then $h_1^2 = \Sigma_i \sigma_{1i}^2 \beta_{1i}^2$ (and likewise for study two). The objective of this work is to estimate transethnic genetic correlation from summary statistics of common variants $Z_1 = \frac{(X_1/\sigma_1)^\top Y_1}{\sqrt{N_1}}$ (and likewise for study two) and estimates of population LD matrices (Σ_1 and Σ_2) from external reference panels. Define the *genetic effect correlation* $\rho_{ge} = Cor(\beta_1, \beta_2)$ and the *genetic impact correlation* $\rho_{gi} = Cor(\sigma_1\beta_1, \sigma_2\beta_2)$.

We assume 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 two). When effect sizes β are assumed inversely proportional to allele frequency, as is commonly done[16,29], we show (Appendix) 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:

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

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

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

Given these equations for variance, the quantities ρ_{gi} or ρ_{ge} and h_1^2, h_2^2 can be estimated by maximizing the multivariate normal likelihood, $l(\rho_{g\{i,e\}}, h_1^2, h_2^2 | Z, \Sigma, \sigma) \propto -\ln(|C|) - Z^\top C^{-1} Z$, where C is either of the above covariance matrices (1) and (2). Because Σ_1 and Σ_2 are estimated from finite external reference panels, maximum likelihood estimation of the above multivariate normal 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 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 down weight Z-scores of SNPs in proportion their LD. Second, rather than compute the full products Σ_1^2, Σ_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[16]. The full details of the derivation and optimization are provided in the appendix.

Results

Simulated Genotypes and Simulated Phenotypes

We simulated 50,000 European-like (EUR) and 50,000 East Asian-like (EAS) individuals at 248,953 SNPs from chromosomes 1-3 with allele frequency above 1% in both European and East Asian HapMap3 populations with HapGen2[35]. HapGen2 implements a haplotype recombination with mutation model that results in excess local relatedness among the simulated individuals. To account for this local structure, we used Plink2[36] to filter individuals with genetic relatedness above 0.05, resulting in 4499 EUR-like indivudals and 4837 EAS-like individuals. From these simulated individuals, 500 per population were chosen uniformly at random to serve as an external reference panel for estimating Σ_1 and Σ_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*. ρ_{gi} was analytically computed from the simulated effect sizes and allele frequencies in the simulated reference genotypes. Quantitative phenotypes were generated under a linear model with i.i.d. noise and normalized to have mean 0 and variance 1, while binary phenotypes were generated under a liability threshold model where individuals are labeled cases when their liability exceeds a threshold $\tau = \Phi^{-1}(1 - K)$, with K the population disease prevalence[37].

We varied h_1^2, h_2^2, ρ_{ge}, and ρ_{gi}, as well as the number of individuals in each study (N_1, N_2), the number of SNPs (M), the population prevalence K, and proportion of causal variants (p) in the simulated GWAS 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 standard error of the estimator decreases as the number of SNPs and individuals in the study increases. 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[38], therefore we sought to quantify the extent that this bias may effect our estimates. We simulated phenotypes under six different disease models. *Independent*: effect size independent of allele frequency. *Inverse*: effect size inversely proportional to allele frequency. *Rare*: only SNPs with allele frequency under 10% affect the trait. *Common*: only SNPs with allele frequency between 40% and 50% affect the trait. *Difference*: effect size proportional to difference in allele frequency. *Adversarial*: difference model with sign of beta set to increase the phenotype in the population where the allele is most common. Additional genetic architectures are possible, including ones where effect sizes are not a direct function of MAF[39].

We simulated phenotypes using genotypes with allele frequency above 1% or 5% and compared the true and estimated genetic impact and effect correlation among all models (Table 1). We find that when only SNPs with frequency above 5% in both populations are used, the difference in ρ_{ge} and ρ_{gi} is minimal except in the most adversarial cases. Even in the adversarial model, the true difference is only 7%. Though unlikely to represent reality, the four nonstandard disease models result in substantial bias in our estimators. When SNPs with 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 proceed with a 5% MAF cutoff in our analysis of real data, and use the notation h_c^2 to refer to the heritability of SNPs with 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.

Validation of Popcorn using gene expression in GEUVADIS

We compared the common-SNP heritability (h_c^2) and genetic correlation estimates of *popcorn* to GCTA in the gEUVADIS dataset for which raw genotypes are publicly available. gEUVADIS consists of 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, FIN, GBR, TSI) and 89 are of African ancestry (YRI). Raw RNA-sequencing reads obtained from the European Nucleotide Archive were aligned to the transcriptome using UCSC annotations matching hg19 coordinates. RSEM was used to estimate the abundances of each annotated isoform and total gene abundance is calculated as the sum of all isoform abundances normalized to one million total counts or transcripts per million (TPM). For eQTL mapping, Caucasian and Yoruban samples were analyzed separately. For each population, TPMs were median-normalized to account for differences in sequencing depth in each sample and standardized to mean 0 and variance 1. Of the 29763 total genes, 9350 with TPM > 2 in both populations were chosen for this analysis.

For each gene, we conducted a *cis*-eQTL association study at all SNPs within 1 megabase of the gene body with allele frequency above 5% in both populations using 30 principal components as covariates. We found that GCTA and *popcorn* agree on the global distribution of heritability (Figure S3), and that GCTA's estimates of genetic correlation

have a similar distribution to *popcorn's* genetic effect and genetic impact correlation estimates (Figure 2). While the number of SNPs and individuals included in each gene analysis are too small to obtain 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 find 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 arise for several reasons. First, we remove 68% of the transcripts that are lowly expressed in either the YRI or EUR data. 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), while the average genetic impact correlation was 0.313 (0.010). Notably, the genetic correlation increases as the *cis*-h_c^2 of expression in both populations increases (Figure 3). In particular, when the *cis*-h_c^2 of the gene is at least 0.2 in both populations the genetic effect correlation was 0.772 (0.017) while the genetic impact correlation was 0.753 (0.018).

In order to verify that there were no small-sample size or conditioning biases in our analysis, we analyzed the genetic correlation of simulated phenotypes over the gEUVADIS genotypes. We sampled pairs of heritabilities from the estimated expression heritability distribution and simulated pairs of phenotypes to have the given heritability and a genetic effect correlation of 0.0 over randomly chosen 4000 base 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, the average estimated genetic effect correlation was not significantly different from 0.0 conditional on the estimates of heritability being above a certain threshold (Figure S4).

We find that while the average genetic correlation is low, the genetic correlation increases with the *cis*-h_c^2 of the gene, indicating that as *cis*-genetic regulation of gene expression increases it does so similarly in both YRI and EUR populations. This may help interpret the recent observation that while the global genetic correlation of gene expression across tissues is low[40], *cis*-eQTL's tend to replicate across tissues[41]. As the presence of a *cis*-

eQTL indicates substantial *cis*-genetic regulation, an analysis of eQTL replication across tissues is implicitly conditioning on the heritability of gene expression being high and therefore may indicate much higher genetic correlation than average.

Summary statistics of RA and T2D

Finally, we sought to examine the transethnic ρ_{gi} and ρ_{ge} in RA and T2D cohorts for which raw genotypes are not available. We obtained summary statistics of GWAS for rheumatoid arthritis and type-2 diabetes conducted in European and East Asian populations. 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 MHC region (chromosome 6, 25–35 Mb) from the RA summary statistics. We estimated the common-SNP heritability and genetic correlation using 2,539,629 SNPs genotyped or imputed in both RA studies and 1,054,079 SNPs genotyped or imputed in both T2D studies with allele frequency above 5% in 1000 genomes EUR and EAS populations. The h_c^2 and genetic correlation estimates are presented in Table 2. Our RA h_c^2 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 estimates of 0.32 (0.037) in Europeans[42]. Similarly, our T2D h_c^2 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[42]. We stress that this discrepancy is likely due to the difference between common-SNP heritability h_c^2 and total narrow-sense heritability h_g^2. Furthermore, estimates of the heritability of T2D from family studies can vary significantly[43,44].

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

Discussion

We have developed the transethnic genetic effect and genetic impact correlation and provided an estimator for these quantities based only on summary-level GWAS information

and suitable reference panels. We have applied our estimator to several phenotypes: rheumatoid arthritis, type-2 diabetes and gene expression. While the gEUVADIS dataset lacks the power required to make inferences about 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 find 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 is 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 may be identical ($\rho_{ge} = 1$) while polygenic selection acts to change only the allele frequencies. Similarly, the effect sizes may be different ($\rho_{ge} < 1$) without selection. Differences in effect sizes at common SNPs can result from many phenomena. We expect un-typed and un-imputed variants differentially linked to observed SNPs to contribute significantly, along with rare or population-specific variants differentially linked to observed SNPs. If a gene-gene or gene-environment interaction exists, but only marginal effects are tested, the observed marginal effects may be different in each population due to allele frequency differences even if the interaction effect is the same in both populations, and this will result in decreased genetic correlation. While within-locus (dominance) interactions may also play a role[45], the magnitude of this effect has been debated[46]. We emphasize that we cannot differentiate between these effects on the basis of this analysis alone, and further research is required to establish the magnitude of the contribution of each of these effects to inter-population effect size differences.

Estimates of the transethnic genetic correlation are important for several reasons. They may help inform best practices for transethnic meta-analysis, potentially offering 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). Our observation that for RA, T2D, and gene expression the genetic correlation is low shows that out of population phenotypic predictive power is quite

low. Similarly, it implies that disease risk assessment in non-Europeans based on current GWAS results may be problematic, necessitating increased study of disease in many populations to gain insight into differences in genetic architecture and improve risk assessment.

While 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 variance explained $\rho_{ge} = \mathrm{Cor}(\sigma_1^2\beta_1^2, \sigma_2^2\beta_2^2)$ and proportion of shared causal variants between the two populations. Interestingly, while our goal was to construct an estimator that determined the extent of genetic sharing independent of allele frequency, we observe that the correlation of genetic effect and genetic impact are similar. Furthermore, our simulations show that under a random effects model utilizing only SNPs with allele frequency above 5% in both populations the true genetic effect and genetic impact correlation 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 avoiding the computational bottleneck required to estimate 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 are exactly the LD-scores ($\Sigma_{1ii}^2 = l_i$). One could ignore our likelihood-based approach and define *cross-population scores* $c_i = \sum_m r_{1im} r_{2im}$ in order to exploit the linear relationship $\mathbb{E}[Z_{1i} Z_{2i}] = \frac{\sqrt{N_1 N_2}}{M} \rho_{gi} \sqrt{h_1^2 h_2^2} c_i$ (a similar approach can be taken for the genetic effect correlation). Since LD-score regression has been successfully used to compute 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 they perform similarly as heritability estimators (Figure S5).

Of course, our method is not without drawbacks. First, it requires a large sample size and large number of SNPs to achieve standard errors low enough to generate accurate estimates. Until recently large sample GWAS have been rare in non-European populations, though they are becoming more common. Similarly, reference panel quality may suffer in non-European populations and this may impact downstream analysis[47]. Second, it is limited to analyzing relatively common SNPs, both because having an accurate disease model is important for the analysis of rare variants and because effect size and correlation coefficient estimates have a high standard error at rare SNPs[16]. 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 ρ is bounded between -1 and 1. This may induce bias when the true value is close to the boundary and the sample size is small. Finally, admixed populations induce very long-range LD that is not accounted for in our approach and we are therefore limited to un-admixed populations[16].

Our analysis leaves open several avenues for future work. We approximately maximize the likelihood of an $M \times M$ multivariate normal distribution via a method that uses only the diagonal elements of each block, discarding covariance information between Z-scores. A better approximation may lower the standard error 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 as well as variants at frequencies between 1-5% or lower than 1%. Our simulations indicate that having an accurate disease model is important for determining the difference between the genetic effect and genetic impact correlation 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 α and ϒ 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

Consider two GWAS of a phenotype conducted in different populations populations. Assume we have N_1 individuals genotyped or imputed to M SNPs in study one and N_2 individuals genotyped or imputed to M SNPs in study two. Let X_1, X_2 and Y_1, Y_2 be the matrices of mean-centered genotypes and phenotypes of the individuals in study one and two, respectively, with f_1, f_2 the allele frequencies of the M SNPs common to both populations. Assuming Hardy-Weinberg equilibrium, the allele variances are $\sigma_1^2 = 2f_1(1 - f_1), \sigma_2^2 = 2f_2(1 - f_2)$. Let β_1, β_2 be the (unobserved) per-allele effect sizes for each SNP in studies one and two, respectively. Define the *genetic impact correlation* $\rho_{gi} = \mathrm{Cor}(\sqrt{\sigma_1^2}\beta_1, \sqrt{\sigma_2^2}\beta_2)$ and the *genetic effect correlation* $\rho_{ge} = \mathrm{Cor}(\beta_1, \beta_2)$. We present a maximum likelihood framework for estimating the heritability of the phenotype in study 1 and it's standard error, the heritability of the phenotype in study 2 and it's standard error, and the genetic effect and impact correlation of the phenotype between the studies and it's standard error given only the summary statistics Z_1, Z_2 and reference genotypes G_1, G_2 representing the populations in the studies. We assume that genotypes are drawn randomly from populations with expected correlation matrices Σ_1 (and similarly for study two), and that every SNP is causal with a normally distributed effects size (though this assumption is not necessary in practice, see Figure S1).

Genetic impact correlation

Let $X_1' = \frac{X_1}{\sqrt{\sigma_1^2}}$ (and similarly for study 2) be normalized genotype matrices. We consider the standard linear model for generation of the phenotypes, where

$$Y_1 = X_1'\beta_1 + \epsilon_1$$
$$Y_2 = X_2'\beta_2 + \epsilon_2$$

For convenience of notation let $h_{ix}^2 = \rho_{gi}\sqrt{h_1^2 h_2^2}$. We assume 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:

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

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} (1 - h_1^2)\mathbb{I}_M & 0 \\ 0 & (1 - h_2^2)\mathbb{I}_M \end{bmatrix} \right) \tag{2}$$

where h_1^2, h_2^2 are the heritability of the disease in study one and two, respectively, and ρ_{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 $(h_1^2, h_2^2, \rho_{gi})$. Given a distribution for Z and an observation of Z we can then choose parameters which 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 β when the sample size is large enough. Since $\mathbb{P}(Z) \propto \mathbb{P}(Z|\beta)\mathbb{P}(\beta)$ and the product of normal distributions is normal, we only need to compute the unconditional mean and variance of Z to know its distribution. Specifically, let $Z = [Z_1^\top, Z_2^\top]^\top$, then it's mean is

$$\mathbb{E}[Z] = \mathbb{E}\begin{bmatrix} \frac{X_1'^\top Y_1}{\sqrt{N_1}} \\ \frac{X_2'^\top Y_2}{\sqrt{N_2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{N_1}}(\mathbb{E}[X_1'^\top X_1']\mathbb{E}[\beta_1] + \mathbb{E}[X_1'^\top]\mathbb{E}[\epsilon_1]) \\ \frac{1}{\sqrt{N_2}}(\mathbb{E}[X_2'^\top X_2']\mathbb{E}[\beta_2] + \mathbb{E}[X_2'^\top]\mathbb{E}[\epsilon_2]) \end{bmatrix} = 0$$

The within-population variance is:

$$\begin{aligned}
\mathrm{Cov}[Z_{1i}, Z_{1j}] &= \mathbb{E}[Z_{1i}Z_{1j}] = \mathbb{E}_{X,\beta,\epsilon}[\mathbb{E}[Z_{1i}Z_{1j}|X,\beta,\epsilon]] \\
&= \frac{1}{N_1}\mathbb{E}_{X,\beta,\epsilon}[X_{1i}'^\top(X_1'\beta_1 + \epsilon_1)(X_1'\beta_1 + \epsilon_1)^\top X_{1j}'] \\
&= \frac{1}{N_1}\mathbb{E}_{X,\beta}[X_{1i}'^\top X_1'\beta_1\beta_1^\top X_1'^\top X_{2j}] + \frac{1}{N_1}\mathbb{E}_{X,\epsilon}[X_{1i}'^\top \epsilon_1\epsilon_1^\top X_{1j}'] \\
&= \frac{h_1^2}{MN_1}\mathbb{E}_X[X_{1i}'^\top X_1' X_1'^\top X_{1j}'] + \frac{1-h_1^2}{N_1}\mathbb{E}_X[X_{1i}'^\top X_{1j}'] \\
&= \frac{h_1^2}{MN_1}(N_1 M r_{1ij} + N_1\sum_{m=1}^{M} r_{1im}r_{1jm} + N_1^2\sum_{m=1}^{M} r_{1im}r_{1jm}) + \frac{1-h_1^2}{N_1}r_{1ij} \\
&= r_{1ij} + \frac{N_1+1}{M}h_1^2\Sigma_{1(i)}\Sigma_1^{(j)}
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{Cov}[Z_{1i}, Z_{2j}] &= \frac{1}{\sqrt{N_1N_2}}\mathbb{E}_{X,\beta}[X_{1i}'^\top X_1'\beta_1\beta_2^\top X_2'^\top X_{2j}'] + \frac{1}{\sqrt{N_1N_2}}\mathbb{E}_{X,\epsilon}[X_{1i}'^\top \epsilon_1\epsilon_2^\top X_{2j}'] \\
&= \frac{h_{ix}^2}{M\sqrt{N_1N_2}}\mathbb{E}_X[X_{1i}'^\top X_1' X_2'^\top X_{2j}'] \\
&= \frac{h_{ix}^2}{M\sqrt{N_1N_2}}(N_1N_2\sum_{m=1}^{M} r_{1im}r_{2jm}) \\
&= \frac{\sqrt{N_1N_2}}{M}h_{ix}^2\Sigma_{1(i)}\Sigma_2^{(j)}
\end{aligned}$$

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

thus

$$C = \mathrm{Var}(Z) = \begin{bmatrix} \Sigma_1 + \frac{N_1+1}{M} h_1^2 \Sigma_1^2 & h_{gx}^2 \frac{\sqrt{N_1 N_2}}{M} \Sigma_1 \Sigma_2 \\ h_{gx}^2 \frac{\sqrt{N_1 N_2}}{M} \Sigma_2 \Sigma_1 & \Sigma_2 + \frac{N_2+1}{M} h_2^2 \Sigma_2^2 \end{bmatrix} \tag{3}$$

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

Genetic effect correlation

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

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{h_1^2}{\|\sigma_1^2\|_1} \mathbb{I}_M & \frac{h_{ex}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}} \mathbb{I}_M \\ \frac{h_{ex}}{\sqrt{\|\sigma_1^2\|_1 \|\sigma_2^2\|_1}} \mathbb{I}_M & \frac{h_2^2}{\|\sigma_2^2\|_1} \mathbb{I}_M \end{bmatrix} \right) \tag{4}$$

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

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

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}, then we find ρ_{gi} or ρ_{ge} by maximizing the likelihood corresponding to C_{12}:

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

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

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

Because we are discarding between-SNP covariance information ($\mathrm{Cov}(Z_{1i}, Z_{1j})$), highly correlated SNPs will be overcounted 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'th entry in it's corresponding

LD-matrix product. Therefore we let $w^{gi}_{jki} = 1/\left(\Sigma_j\Sigma_k\right)_{ii}$ and $w^{ge}_{jki} = 1/\left(\Sigma_j\sqrt{\sigma^2_j\sigma^2_k}\Sigma_k\right)_{ii}$ to reduce the variance in our estimates of the parameters h^2_1, h^2_2, ρ_{gi} and ρ_{ge}.

Furthermore, rather than compute the full products Σ^2_1, Σ^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 $(\Sigma_a\Sigma_b)_{ii} = \sum_{w=i-W}^{w=i+W} r_{aiw}r_{biw}$. Though maximum likelihood estimation admits a straightforward estimate of the standard error via the fisher information, we found these estimates to be inaccurate in practice. Instead, we use block jackknife with block size equal to $\min(100, \frac{M}{200})$ SNPs to ensure that blocks are large enough to remove 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 effects sizes in population 2 be β_2. Let the true phenotypes in population 1 be $Y = X_1\beta_1 + \epsilon_1$ while the predicted phenotypes are $P = X_1\beta_2$. We are interested in the correlaiton of the predicted and true phenotypes $\rho^{MAX}_{YP} = \mathrm{Cor}(Y, P)$. Notice that, given X, the true and predicted phenotype of each individual is an affine transformation of a multivariate normal random variable (β)

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

and therefore (Y_i, P_i) for individual i is multivariate normal with expected covariance matrix

$$\begin{aligned}
\mathbb{E}_X\left[\mathrm{Cov}(Y_i, P_i)\right] &= \mathbb{E}_X \begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix} \begin{bmatrix} \frac{1}{\|\sigma^2_1\|_1}\mathbb{I}_M & \frac{h_{ex}}{\sqrt{\|\sigma^2_1\|_1\|\sigma^2_2\|_1}}\mathbb{I}_M \\ \frac{h_{ex}}{\sqrt{\|\sigma^2_1\|_1\|\sigma^2_2\|_1}}\mathbb{I}_M & \frac{h^2_2}{\|\sigma^2_2\|_1}\mathbb{I}_M \end{bmatrix} \begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix}^\top \\
&= \mathbb{E}_X \begin{bmatrix} \frac{\sum_m X^2_{im}}{\|\sigma^2_1\|_1} & \frac{h_{ex}\sum_m X^2_{im}}{\sqrt{\|\sigma^2_1\|_1\|\sigma^2_2\|_1}} \\ \frac{h_{ex}\sum_m X^2_{im}}{\sqrt{\|\sigma^2_1\|_1\|\sigma^2_2\|_1}} & \frac{h^2_2\sum_m X^2_{im}}{\|\sigma^2_2\|_1} \end{bmatrix} \\
&= \begin{bmatrix} 1 & h_{ex}\sqrt{\frac{\|\sigma^2_1\|_1}{\|\sigma^2_2\|_1}} \\ h_{ex}\sqrt{\frac{\|\sigma^2_1\|_1}{\|\sigma^2_2\|_1}} & h^2_2\frac{\|\sigma^2_1\|_1}{\|\sigma^2_2\|_1} \end{bmatrix}
\end{aligned}$$

and therefore the expected correlation $\mathbb{E}[\mathrm{Cor}(Y_i, P_i)]$ is $\frac{h_{ex}}{\sqrt{h^2_2}}\sqrt{\frac{\|\sigma^2_1\|_1}{\|\sigma^2_2\|_1}\frac{\|\sigma^2_2\|_1}{\|\sigma^2_1\|_1}} = \rho_{ge}\sqrt{h^2_1}$. The expected population correlation tends to the sample correlation as the number of samples increases, therefore

$$\rho^{MAX}_{YP} = \mathrm{Cor}(Y, P) \to \rho_{ge}\sqrt{h^2_1} \tag{5}$$

as $N \to \infty$

Description of Supplemental data

Supplemental data include six figures and one table.

Acknowledgements

The authors would like to acknowledge Lior Pachter and Hilary Finucane for insightful discussion about the problem. BCB is supported by the NSF GRFP. ALP is supported by NIH grant R01 HG006399. NZ is supported by NIH grant K25HL121295.

Web Resources

Popcorn is available at https://github.com/brielin/popcorn

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., De La Vega, F.M., and Burchard, E.G. (2011). Genomics for the world. Nature *475*, 163–165.
6. (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet *43*, 339–344.
7. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
8. 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. (2013). Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent. Am. J. Hum. Genet. *93*, 463–470.
9. 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.
10. 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.
11. 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. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375.
12. 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.

13. 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.
14. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. Bioinformatics *31*, i206–i213.
15. 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.
16. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet *47*, 291–295.
17. 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. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet *47*, 1228–1235.
18. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat Genet *47*, 1236–1241.
19. Park, D.S., Brown, B., Eng, C., Huntsman, S., Hu, D., Torgerson, D.G., Burchard, E.G., and Zaitlen, N. (2015). Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses. Bioinformatics *31*, i181–i189.
20. 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.
21. International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. Nature *460*, 748–752.
22. 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.
23. Fesinmeyer, M., North, K., Ritchie, M., Lim, U., Franceschini, N., Wilkens, L., Gross, M., Bůžková, P., Glenn, K., Quibrera, P., et al. (2013). Genetic risk factors for body mass index and obesity in an ethnically diverse population: results from the Population Architecture using Genomics and Epidemiology (PAGE) Study. Obes. Silver Spring Md *21*, 10.1002/oby.20268.
24. Chang, M., 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.
25. 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., et al. (2010). Consistent Association of Type 2 Diabetes Risk Variants Found in Europeans in Diverse Racial and Ethnic Groups. PLoS Genet. *6*, e1001078.
26. 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.
27. Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to quantitative genetics (Essex, England: Longman).
28. 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.
29. 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.
30. 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. n/a – n/a.
31. 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.

32. '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. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. *31*, 1015–1022.
33. 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. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.
34. Cho, Y.S., Chen, C.-H., Hu, C., Long, J., Hee Ong, R.T., Sim, X., Takeuchi, F., Wu, Y., Go, M.J., Yamauchi, T., et al. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet *44*, 67–72.
35. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.
36. 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*,.
37. 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.
38. 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.
39. 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. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.
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. 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. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. *44*, 483–489.
43. 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.
44. 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.
45. 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.
46. 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. (2015). Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. Am. J. Hum. Genet. *96*, 377–385.
47. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet *11*, 499–511.
48. 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: Diabetes in East Asians. Ann. N. Y. Acad. Sci. *1281*, 64–91.

Figure Titles and Legends

Figure 1: True and estimated genetic impact and effect correlation. All simulations conducted with simulated EUR and EAS heritability of 0.5 using 4499 simulated EUR and 4836 simulated EAS individuals at 248,953 SNPs.

Figure 2: Distribution of genetic correlation comparison between popcorn and GCTA. Distribution was computed using a gaussian kde on the set of genetic correlation estimates.

Figure 3: Genetic correlation as a function of heritability for gene expression. The mean and standard error 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) is plotted against h (x-axis).

Tables

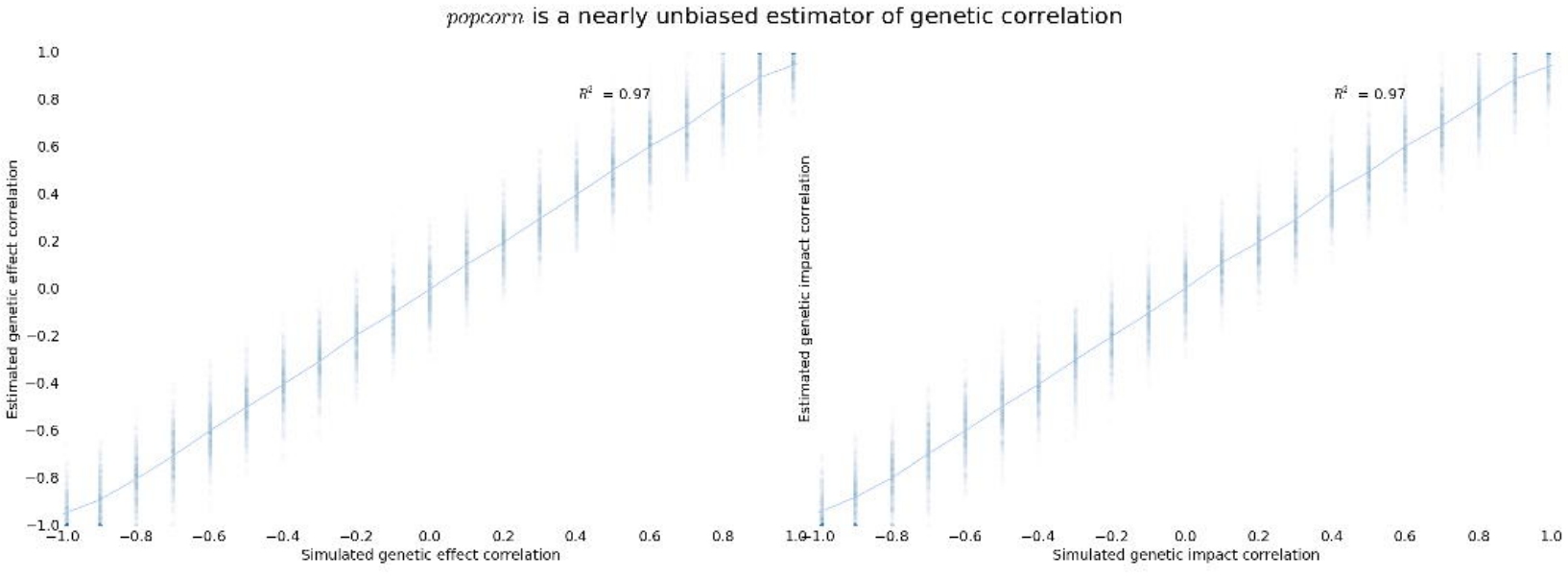
	MAF > 0.01				MAF > 0.05			
Model	ρ_{ge}	ρ_{gi}	$\hat{\rho}_{ge}$	$\hat{\rho}_{gi}$	ρ_{ge}	ρ_{gi}	$\hat{\rho}_{ge}$	$\hat{\rho}_{gi}$
Independent	0.500	0.478	0.500	0.460	0.500	0.488	0.509	0.469
Inverse	0.431	0.500	0.567	0.496	0.479	0.500	0.555	0.482
Rare	0.500	0.467	0.382	0.863	0.500	0.496	0.998	0.756
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

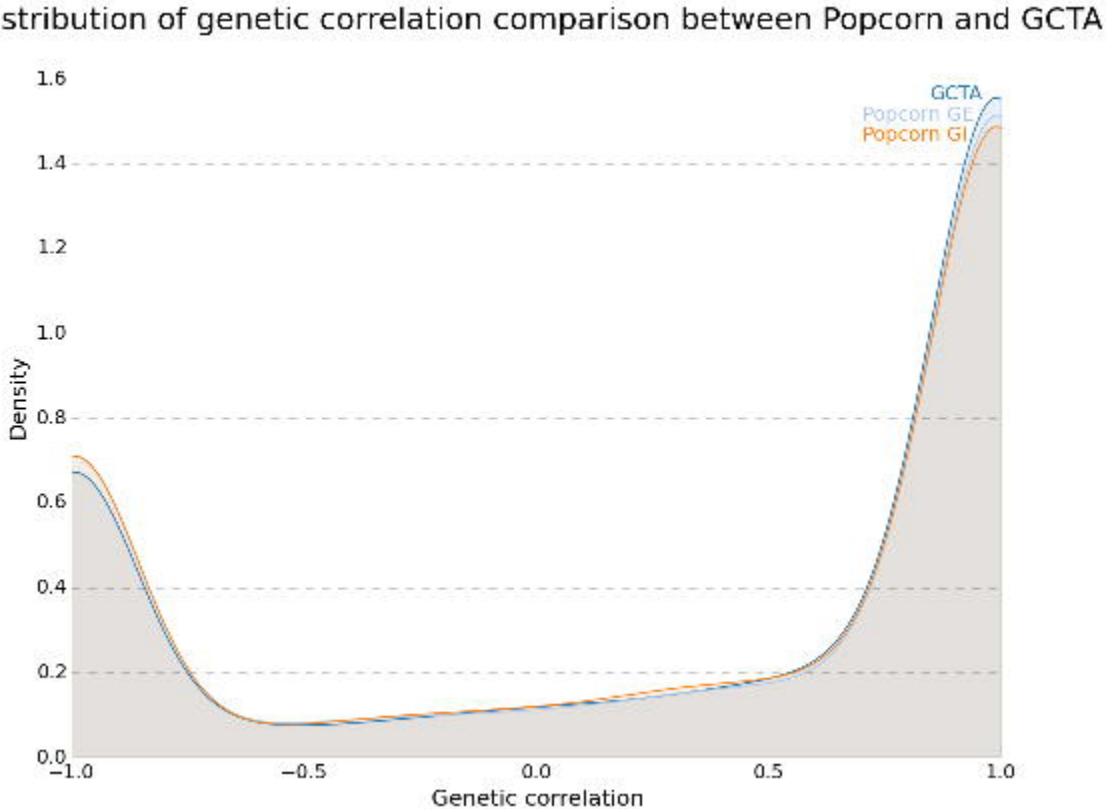
Table 1: True and estimated values of the genetic impact and effect correlation in simulated EUR-like and EAS-like genotypes. Results are the average of 100 simulations with phenotype heritability of 0.5 in each population.

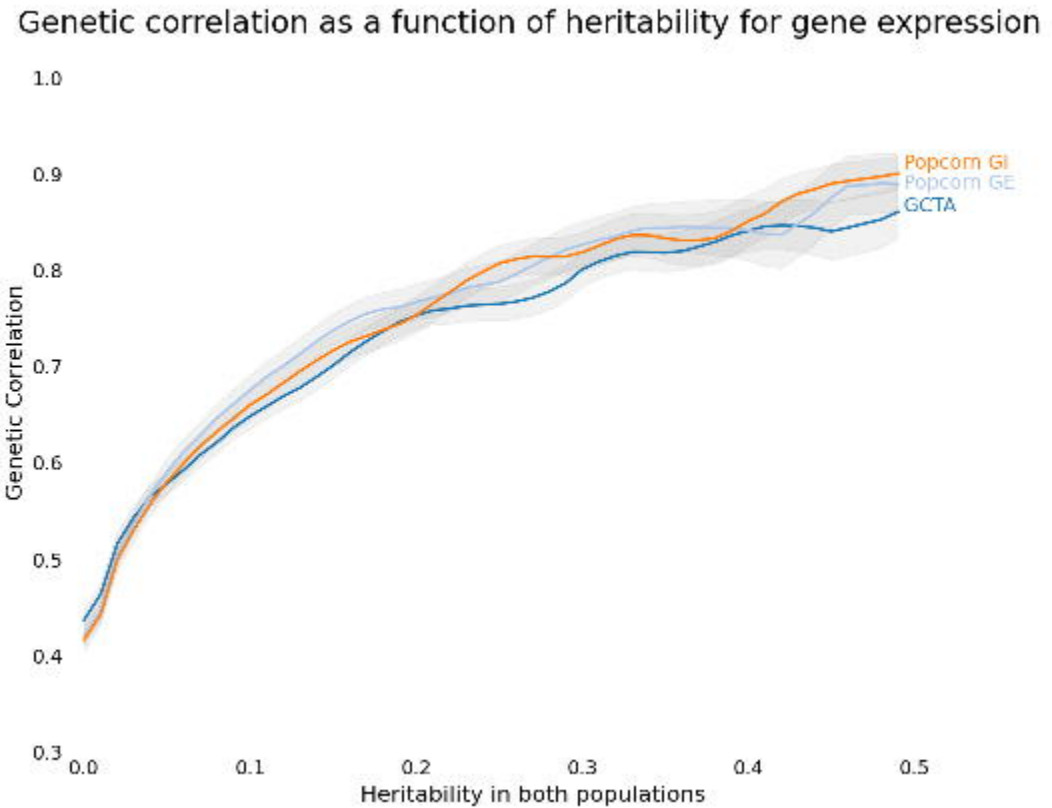
		h_{EUR}^2 liability	h_{EAS}^2 liability	ρ_{ge}	ρ_{gi}
RA	Est. (SE)	0.18 (0.02)	0.22 (0.03)	0.46 (0.06)	0.46 (0.06)
	95% CI	[0.15, 0.21]	[0.16, 0.28]	[0.34, 0.58]	[0.34, 0.58]
	$p_{>0}$	3.90e-32	1.89e-17	1.37e-15	8.16e-16
	$p_{<1}$	0.0	3.1e-197	2.53e-20	4.87e-22
T2D	Est. (SE)	0.24 (0.01)	0.11 (0.02)	0.62 (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-05	2.06e-06

Table 2: Heritability and genetic correlation of RA and T2D between EUR and EAS populations. EUR RA data contained 8,875 cases and 29,367 controls for a study prevalence of 0.23. EAS RA data contained 4,873 cases and 17,642 controls for a study prevalence of 0.22. RA disease prevalence was assumed to be 0.5% in both populations[7]. T2D EUR data contained 12171 cases and 56862 controls for a study prevalence of 0.18. T2D EAS data

contained 6952 cases and 11865 controls for a study prevalence of 0.37. T2D EUR prevalence was assumed to be 8%[33] while T2D EAS prevalence was assumed to be 9%[48].


popcorn is a nearly unbiased estimator of genetic correlation
R^2 = 0.97
Estimated genetic effect correlation
Simulated genetic effect correlation
R^2 = 0.97
Estimated genetic impact correlation
Simulated genetic impact correlation


Distribution of genetic correlation comparison between Popcorn and GCTA
GCTA
Popcorn GE
Popcorn GI
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Genetic correlation


Genetic correlation as a function of heritability for gene expression
Popcorn GI
Popcorn GE
GCTA
Genetic Correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.1
0.2
0.3
0.4
0.5
Heritability in both populations