Title

Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits

Permalink
https://escholarship.org/uc/item/7dx0s4pr

Journal
American Journal of Human Genetics, 101(5)

ISSN
0002-9297

Authors
Shi, Huwenbo
Mancuso, Nicholas
Spendlove, Sarah
et al.

Publication Date
2017-11-01

DOI
10.1016/j.ajhg.2017.09.022

Peer reviewed

Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits

Huwenbo Shi,[1,*] Nicholas Mancuso,[2] Sarah Spendlove,[4] and Bogdan Pasaniuc[1,2,3]

Although genetic correlations between complex traits provide valuable insights into epidemiological and etiological studies, a precise quantification of which genomic regions disproportionately contribute to the genome-wide correlation is currently lacking. Here, we introduce ρ-HESS, a technique to quantify the correlation between pairs of traits due to genetic variation at a small region in the genome. Our approach requires GWAS summary data only and makes no distributional assumption on the causal variant effect sizes while accounting for linkage disequilibrium (LD) and overlapping GWAS samples. We analyzed large-scale GWAS summary data across 36 quantitative traits, and identified 25 genomic regions that contribute significantly to the genetic correlation among these traits. Notably, we find 6 genomic regions that contribute to the genetic correlation of 10 pairs of traits that show negligible genome-wide correlation, further showcasing the power of local genetic correlation analyses. Finally, we report the distribution of local genetic correlations across the genome for 55 pairs of traits that show putative causal relationships.

Introduction

Genomic regions that harbor variants contributing to multiple traits provide valuable insights into the underlying biological mechanisms with which genetic variation impacts complex traits.[1–7] Therefore, both *de novo* discovery of such regions as well as the quantification of the correlation in effect sizes at known shared regions are important to epidemiological and etiological studies. For example, genetic variants associated with multiple traits in genome-wide associations studies (GWASs) can be used as instrumental variables in Mendelian randomization analyses to suggest causal relationships among complex traits.[7–10] Unfortunately, many risk variants are left undetected by existing GWASs due to a combination of high polygenicity (i.e., many variants of small effects) and sample sizes which limits the power to detect genetic variants of small effect.[11] To improve accuracy at sub-GWAS significant regions, recent works[1,2] proposed to utilize the posterior probability of two traits sharing a causal variant at a given risk region to detect genetic overlap. Although powerful in detecting shared genetic risk variants, the posterior probability does not convey the direction or magnitude of the genetic effect at the overlapped genomic regions.[1,2] Alternative approaches have used genetic correlation (i.e., correlation of the genetic components of two traits), that summarizes both direction and magnitude of effects, to gain insights into genetic overlap of complex traits.[12–14] Traditional methods to estimate genetic correlation are hindered by the lack of availability of large-scale individual-level data due to privacy concerns as they require individual genotype and trait measurements on the same set of individuals.[12,14,15] More recent works have shown that GWAS summary data (i.e., effect sizes and standard errors at all variants typed in the study) are sufficient to estimate genome-wide genetic correlation under a polygenic trait architecture by aggregating information across all typed variants in the study.[16,17]

In this work, we investigate the correlation between traits due to typed genetic variants from a small region in the genome (i.e., local genetic correlation) as means to identify genomic regions that contribute disproportionately to the genetic sharing between traits. We introduce methods that estimate the local genetic correlation from GWAS summary data while allowing for overlapping GWAS samples and linkage disequilibrium (LD) among variants. We partition the genome-wide genetic sharing across approximately independent LD regions of 1.6 Mb in width on average.[18] To allow for a broad range of causal effect sizes, our approach makes no distributional assumptions on the causal effect sizes by treating them as fixed quantities. Our method can be viewed as a natural extension to pairs of traits of recently proposed methods that quantify local SNP heritability from GWAS summary data under a fixed-effect model.[19]

We illustrate the utility of local genetic correlation through an analysis of GWAS summary data of 36 quantitative complex traits. We identify 25 genomic regions that show significant local genetic correlation across 27 pairs of traits; e.g., region chr2: 21M–23M that harbors *APOB* (MIM: 107730) shows a significant genetic correlation for the pair of traits high-density lipoprotein (HDL) and triglycerides (TG). Notably, 6 (out of the 25) regions show significant local genetic correlation although the

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90024, USA; [2]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA; [3]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA; [4]Department of Biology, Brigham Young University, Provo, UT 84602, USA
*Correspondence: shihuwenbo@ucla.edu
https://doi.org/10.1016/j.ajhg.2017.09.022.

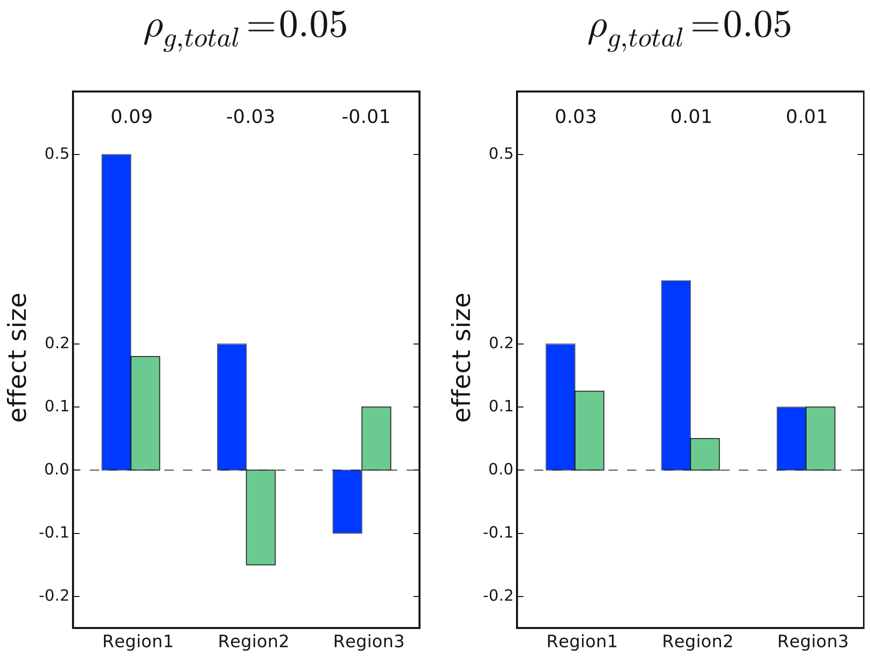

Figure 1. Examples of Two Different Distributions of Local Genetic Covariances that Result in the Same Total Genetic Covariance

Covariances shown at the top of each bar; total genetic covariance ($\rho_{g,total} = 0.05$). In the left example, the total genetic covariance is a summation of a large positive local genetic covariance at region 1 and two smaller negative local genetic covariances at region 2 and region 3 (e.g., regions 2 and 3 impact traits through a different pathway than region 1). In the right example, the total genetic covariance is a summation of small positive local genetic covariances (e.g., all three regions impact both traits through the same pathway). Positive local genetic covariance can be interpreted as a locus driving a pathway that regulates two traits in the same direction, and negative local genetic covariance the opposite direction.

genome-wide genetic correlation is not significantly different from 0; e.g., region chr6: 134M–136M shows a significant in local genetic correlation for mean cell volume (MCV) and platelet count (PLT) although the genome-wide genetic correlation MCV-PLT is negligible (0.02, 95% CI [−0.04, 0.07]). This shows that these traits are correlated at a local level (e.g., due to pleiotropy and/or shared pathways) that are not reflected in the genome-wide correlation (due to balancing effect of other loci; e.g., positive correlation partially canceling a negative correlation, see Figure 1). Regions with significant local genetic correlations can also be used to identify new risk loci. For example, although the region chr8: 9.2M–9.6M shows a significant local genetic correlation between HDL and LDL, it does not harbor GWAS variant for HDL and LDL. Finally, we explore putative causal relations between all the 36 studied traits using a recently proposed approach[2] and report 55 instances of pairs with putative causality. For most of these pairs, we show that the local genetic correlation ascertained for GWAS signals specific to each trait is consistent with the putative causal relation while providing a directly interpretable quantity of the magnitude of effect.

Material and Methods

Overview of Methods

Genetic covariance measures the similarity between a pair of traits driven by genetic variations and enjoys wide applications in understanding relations between complex traits.[13,20,21] Genetic covariance is traditionally estimated as a single measure across the entire genome to capture the genome-wide contribution of genetic variations to the correlation between phenotypes. Here, we introduce local genetic covariance, the similarity between pairs of traits driven by genetic variations localized at a specific region in the genome (e.g., one LD block), as a principled way to partition the shared genetic risk between traits. For example, a high genome-wide genetic covariance can be driven by one genomic region containing a shared risk variant or by a large number of regions, each with a small contribution reflecting putative causal relations (where all risk variants for one trait are risk variants for the other trait) and/or pleiotropy (risk variants contributing to both traits through shared pathways) (see Figure 1). Whereas genetic covariance quantifies the magnitude of co-variation of the genetic components of two traits in their original scale, genetic correlation quantifies co-variation in a standardized scale and is therefore comparable across pairs of traits and/or genomic regions for which magnitude of effect size may differ. As a motivating example, consider two traits modeled by $\phi = x_1\beta_1 + x_2\beta_2 + \epsilon$ and $\psi = x_1\gamma_1 + x_2\gamma_2 + \delta$, where x_1 and x_2 represent two independent SNPs. In the special case where $\boldsymbol{\gamma}$ is proportional to $\boldsymbol{\beta}$ by a factor of α, i.e., $\boldsymbol{\gamma} = \alpha\boldsymbol{\beta}$, the genetic covariance between the two traits is $\alpha(\beta_1^2 + \beta_2^2)$ and is governed by α. However, the genetic correlation between the two traits is always 1 for positive α (−1 for negative α) regardless of the magnitude of α.

We start by defining local genetic covariance under the fixed effect model, making a distinction between genetic covariance and covariance of the causal effects, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (see below). We then describe methods to estimate genetic covariance followed by an approach to standardize the local genetic covariance to estimate local genetic correlation.

Local Genetic Covariance under Fixed-Effect Model

Let $\phi = \boldsymbol{x}^\top\boldsymbol{\beta} + \epsilon$ and $\psi = \boldsymbol{x}^\top\boldsymbol{\gamma} + \delta$ be two traits measured at an individual, standardized so that $\mathrm{E}[\phi] = \mathrm{E}[\psi] = 0$ and $\mathrm{Var}[\phi] = \mathrm{Var}[\psi] = 1$, where $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$ are the fixed effect size vectors for the two traits; $\boldsymbol{x} \in \mathbb{R}^p$, the genotype vector of the individual at p SNPs, standardized so that $\mathrm{E}[\boldsymbol{x}] = 0$, and $\mathrm{Var}[\boldsymbol{x}] = \boldsymbol{V}$, the LD matrix; and ϵ, δ, random environmental effects independent of $\boldsymbol{x}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, with $\mathrm{E}[\epsilon] = \mathrm{E}[\delta] = 0$, $\mathrm{Var}[\epsilon] = \sigma_\epsilon^2$, $\mathrm{Var}[\delta] = \sigma_\delta^2$, and $\mathrm{Cov}[\epsilon, \delta] = \rho_e$. Under these assumptions, one can decompose the phenotypic covariance, ρ, between ϕ and ψ into a summation of genetic covariance and environmental covariance, as

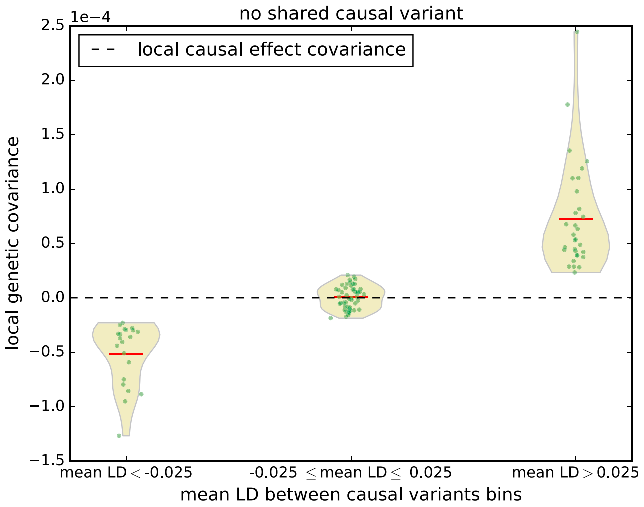

Figure 2. Distribution of Simulated Genetic Covariance and Causal Effect Covariance across 100 LD-Independent Regions on Chromosome 1 Binned by Average LD between Causal Variants

The red lines represent the average local genetic covariance in each bin. For each region, we simulated 2 traits, each with 3 causal variants with effect sizes set to 0.01, and with no shared causal variants (see Figure S1 for the case where the two traits share causal variants). Genetic covariance varies with respect to LD whereas causal effect covariance is always 0 (horizontal dotted line). Since genetic covariance can be thought as an upper bound of prediction accuracy using causal effects from one trait to another, a positive genetic covariance indicates that non-zero prediction accuracy could be attained by virtue of LD tagging.

$$
\begin{aligned}
\rho = \mathrm{Cov}[\phi,\psi] &= \mathrm{E}\,[\phi\psi] - \mathrm{E}\,[\phi]\mathrm{E}\,[\psi] = \mathrm{E}\left[(\boldsymbol{x}^\top\boldsymbol{\beta}+\epsilon)(\boldsymbol{x}^\top\boldsymbol{\gamma}+\delta)^\top\right] \\
&= \mathrm{E}\,[(\boldsymbol{x}^\top\boldsymbol{\beta})(\boldsymbol{x}^\top\boldsymbol{\gamma})] + \mathrm{E}\,[\epsilon\delta] = Cov[\boldsymbol{x}^\top\boldsymbol{\beta},\boldsymbol{x}^\top\boldsymbol{\gamma}] + \mathrm{Cov}[\epsilon,\delta] \\
&= \boldsymbol{\beta}^\top\mathrm{E}\,[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{\gamma} + \mathrm{Cov}[\epsilon,\delta] = \boldsymbol{\beta}^\top\boldsymbol{V}\boldsymbol{\gamma} + \rho_e,
\end{aligned}
\qquad \text{(Equation 1)}
$$

where $\rho_g = \mathrm{Cov}[\boldsymbol{x}^\top\boldsymbol{\beta},\boldsymbol{x}^\top\boldsymbol{\gamma}] = \boldsymbol{\beta}^\top\boldsymbol{V}\boldsymbol{\gamma}$ is the genetic covariance between the two traits (i.e., covariance between the genetic components of the two traits, $\boldsymbol{x}^\top\boldsymbol{\beta}$ and $\boldsymbol{x}^\top\boldsymbol{\gamma}$), and ρ_e the environmental covariance (i.e., covariance between the environmental effects of two traits, ϵ and δ). The magnitude and sign of local genetic covariance can be interpreted as the effect and direction of the local genetic component of one trait on that of the other. Thus, given the true effect size vectors, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and the LD matrix $\boldsymbol{V}$, one can obtain ρ_g by plugging in these quantities.

Genetic Covariance versus Covariance of the Causal Effects

An alternative approach to the covariance of the genetic components of the traits is to quantify the covariance of the causal effects (i.e., $\rho_{g,causal} = \boldsymbol{\beta}^\top\boldsymbol{\gamma}$). In the special case where there is no LD (i.e., $\boldsymbol{V} = \boldsymbol{I}$, the identity matrix), genetic covariance and covariance of the causal effects coincide, $\rho_g = \boldsymbol{\beta}^\top\boldsymbol{V}\boldsymbol{\gamma} = \boldsymbol{\beta}^\top\boldsymbol{I}\boldsymbol{\gamma} = \boldsymbol{\beta}^\top\boldsymbol{\gamma} = \rho_{g,causal}$. However, in general genetic covariance is different from covariance of the causal effects as function of the LD between the causal variants. More importantly, high local genetic covariance does not necessarily imply high covariance of the causal effects. In fact, high genetic covariance can be attained even when causal variants are different between the traits. To illustrate the difference, consider an example involving two SNPs. Let $\boldsymbol{\beta} = (1,0)$ and $\boldsymbol{\gamma} = (0,1)$ be the causal effect vectors of the two traits, i.e., the two traits have two distinct set of causal variants. And let

$$
\boldsymbol{V} = \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix}
$$

be the LD matrix between the SNPs. In this example, the covariance of the causal effects is $\rho_{g,causal} = \boldsymbol{\beta}^\top\boldsymbol{\gamma} = 0$, whereas the genetic covariance is $\rho_g = \boldsymbol{\beta}^\top\boldsymbol{V}\boldsymbol{\gamma} = 0.9$. Thus, at a region where the causal variants are distinct for the two traits, covariance of the causal effects is always zero, whereas genetic covariance may be non-zero depending on the LD (see Figure 2). The two definitions measure genetic sharing at different levels of resolution. Local genetic covariance measures sharing at regional level, giving a measure of how similar the regional genetic components are between the two traits, and has applications in predicting the regional genetic component of one trait from that of the other. In contrast, local causal effect covariance measures sharing at an individual SNP level, giving a measure of how similar the causal effects are between the two traits. Consider a scenario where two traits are each driven locally by a different SNP in the same gene. In this case, the local causal effect covariance is zero since the two traits share no causal SNP. However, the local genetic covariance is non-zero if the two SNPs are in LD, which induces similarity in the genetic component of the two traits and is an indication of the gene being shared across the two traits. Although in this work we focus on genetic covariance, for completeness we discuss an estimator for covariance of the causal effects ($\rho_{g,causal}$) in Appendix A.

Estimating Local Genetic Covariance from GWAS Summary Data

In two GWASs involving n_1 individuals for trait 1 (ϕ), n_2 individuals for trait 2 ($\boldsymbol{\psi}$), and n_s shared individuals, we assume

$$
\begin{bmatrix} \phi \\ \phi_s \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X}_s \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \epsilon \\ \epsilon_s \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\psi}_s \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{X'}_s \end{bmatrix}\boldsymbol{\gamma} + \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\delta}_s \end{bmatrix}, \qquad \text{(Equation 2)}
$$

where $(\phi,\phi_s)\in\mathbb{R}^{n_1}$ and $(\boldsymbol{\psi},\boldsymbol{\psi}_s)\in\mathbb{R}^{n_2}$ are the standardized trait values of all individuals in each GWAS; $(\boldsymbol{Y},\boldsymbol{X}_s)\in\mathbb{R}^{n_1\times p}$ and $(\boldsymbol{Z},\boldsymbol{X'}_s)\in\mathbb{R}^{n_2\times p}$, column standardized genotype matrices of all individuals in each GWAS, where $\boldsymbol{X}_s$ and $\boldsymbol{X'}_s$ represent the genotype matrices for the same set of individuals and SNPs but standardized differently in each GWAS; and $(\epsilon,\epsilon_s)\in\mathbb{R}^{n_1}$ and $(\boldsymbol{\delta},\boldsymbol{\delta}_s)\in\mathbb{R}^{n_2}$ are environmental effects of all individuals in each GWAS. We use the subscript s to represent individuals shared by both GWASs. We further assume that $\mathrm{E}\,[\epsilon] = \mathrm{E}\,[\boldsymbol{\delta}] = \mathrm{E}\,[\epsilon_s] = \mathrm{E}\,[\boldsymbol{\delta}_s] = 0$, $\mathrm{Var}[\epsilon] = \mathrm{Var}[\epsilon_s] = \sigma_\epsilon^2\boldsymbol{I}$, $\mathrm{Var}[\boldsymbol{\delta}] = \mathrm{Var}[\boldsymbol{\delta}_s] = \sigma_\delta^2\boldsymbol{I}$, $\mathrm{Cov}[\epsilon,\boldsymbol{\delta}] = 0$, and $\mathrm{Cov}[\epsilon_s,\boldsymbol{\delta}_s] = \rho_e\boldsymbol{I}$.

In a traditional GWAS, we obtain marginal effect size estimates, $\widehat{\boldsymbol{\beta}}_{gwas}$ and $\widehat{\boldsymbol{\gamma}}_{gwas}$, as

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{gwas} &= \frac{1}{n_1}\left[\boldsymbol{Y}^\top\ \boldsymbol{X}_s^\top\right]\begin{bmatrix} \phi \\ \phi_s \end{bmatrix} = \frac{1}{n_1}\left(\boldsymbol{Y}^\top\boldsymbol{Y} + \boldsymbol{X}_s^\top\boldsymbol{X}_s\right)\boldsymbol{\beta} + \frac{1}{n_1}\left(\boldsymbol{Y}^\top\epsilon + \boldsymbol{X}_s^\top\epsilon_s\right) \\
\widehat{\boldsymbol{\gamma}}_{gwas} &= \frac{1}{n_2}\left[\boldsymbol{Z}^\top\ \boldsymbol{X'}_s^\top\right]\begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\psi}_s \end{bmatrix} = \frac{1}{n_2}\left(\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{X'}_s^\top\boldsymbol{X'}_s\right)\boldsymbol{\gamma} + \frac{1}{n_2}\left(\boldsymbol{Z}^\top\boldsymbol{\delta} + \boldsymbol{X'}_s^\top\boldsymbol{\delta}_s\right).
\end{aligned}
\qquad \text{(Equation 3)}
$$

Assuming individuals in both GWASs are drawn from the same population with LD matrix $\boldsymbol{V}$, we have $\widehat{\boldsymbol{\beta}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\beta}, \frac{\sigma_\epsilon^2}{n_1}\boldsymbol{V}\right)$, $\widehat{\boldsymbol{\gamma}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\gamma}, \frac{\sigma_\delta^2}{n_2}\boldsymbol{V}\right)$. We also find

$$\begin{aligned}\mathrm{Cov}\left[\widehat{\boldsymbol{\beta}}_{gwas}, \widehat{\boldsymbol{\gamma}}_{gwas}\right] &= \mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{gwas}\widehat{\boldsymbol{\gamma}}_{gwas}^{\top}\right] - (\boldsymbol{V}\boldsymbol{\beta})(\boldsymbol{V}\boldsymbol{\gamma})^{\top} \\ &= \frac{\rho_e}{n_1 n_2}\mathrm{E}\left[\boldsymbol{X}_s^{\top}\boldsymbol{X}_s'\right] = \frac{\rho_e n_s}{n_1 n_2}\boldsymbol{V},\end{aligned} \qquad \text{(Equation 4)}$$

where the last equality follows from Isserlis' theorem.[22]

Under infinite sample sizes, $\mathrm{Var}[\widehat{\boldsymbol{\beta}}_{gwas}] = \mathrm{Var}[\widehat{\boldsymbol{\gamma}}_{gwas}] = \mathrm{Cov}[\widehat{\boldsymbol{\beta}}_{gwas}, \widehat{\boldsymbol{\gamma}}_{gwas}] = 0$, and we have $\boldsymbol{\beta} = \boldsymbol{V}^{-1}\widehat{\boldsymbol{\beta}}_{gwas}$, $\boldsymbol{\gamma} = \boldsymbol{V}^{-1}\widehat{\boldsymbol{\gamma}}_{gwas}$. Thus, local genetic covariance, $\rho_{g,local}$, can be computed as

$$\rho_{g,local} = \left(\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{-1}\right)\boldsymbol{V}\left(\boldsymbol{V}^{-1}\widehat{\boldsymbol{\gamma}}_{gwas}\right) = \widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{-1}\widehat{\boldsymbol{\gamma}}_{gwas}. \qquad \text{(Equation 5)}$$

However, when sample sizes are finite, from bilinear form theory,[23] the covariance between $\widehat{\boldsymbol{\beta}}_{gwas}$ and $\widehat{\boldsymbol{\gamma}}_{gwas}$ creates bias, resulting in

$$\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{-1}\widehat{\boldsymbol{\gamma}}_{gwas}\right] = \boldsymbol{\beta}^{\top}\boldsymbol{V}\boldsymbol{\gamma} + \frac{\rho_e}{n_1 n_2}\mathrm{tr}(\boldsymbol{V}) = \boldsymbol{\beta}^{\top}\boldsymbol{V}\boldsymbol{\gamma} + \frac{p(\rho - \rho_{g,local})n_s}{n_1 n_2}. \qquad \text{(Equation 6)}$$

Correcting for bias, we arrive at the unbiased estimator

$$\widehat{\rho}_{g,local} = \frac{n_1 n_2 \widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{-1}\widehat{\boldsymbol{\gamma}}_{gwas} - n_s p \rho}{n_1 n_2 - n_s p}. \qquad \text{(Equation 7)}$$

For rank-deficient LD matrix $\boldsymbol{V}$, one replaces $\boldsymbol{V}^{-1}$ with the pseudo-inverse ($\boldsymbol{V}^{\dagger}$) and p with $q = \mathrm{rank}(\boldsymbol{V})$, yielding the unbiased estimator

$$\widehat{\rho}_{g,local} = \frac{n_1 n_2 \widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{\dagger}\boldsymbol{\gamma}_{gwas} - n_s q \rho}{n_1 n_2 - n_s q}. \qquad \text{(Equation 8)}$$

Thus, in order to obtain an unbiased estimate of genetic covariance between a pair of traits, one needs to know their phenotypic covariance. When phenotypic covariance is not available, one can obtain an estimate from genome-wide summary association data using cross-trait LD Score regression,[16]

$$\mathrm{E}\left[z_{\phi,j}z_{\psi,j} \mid l_j\right] = \frac{\sqrt{n_1 n_2}\rho_g}{p}l_j + \frac{\rho n_s}{\sqrt{n_1 n_2}}, \qquad \text{(Equation 9)}$$

where $z_{\phi,j}$ and $z_{\psi,j}$ are the Z-scores of SNP j in the two traits, and l_j the LD score of SNP j. Cross-trait LD Score regression regresses the product of Z-scores at each SNP against its LD score, l_j, and accounts for bias generated by overlapping samples through the intercept term, $\frac{\rho n_s}{\sqrt{n_1 n_2}}$,[16] from which one can obtain an estimate of phenotypic covariance, ρ.

In the special case when $\widehat{\boldsymbol{\beta}}_{gwas}$ and $\widehat{\boldsymbol{\gamma}}_{gwas}$ are obtained for the same trait on the same set of individuals (i.e., $\widehat{\boldsymbol{\beta}}_{gwas} = \widehat{\boldsymbol{\gamma}}_{gwas}$, $n_1 = n_2 = n_s$, $\rho = 1$), Equation 7 reduces to the local SNP-heritability estimator.[19] When $n_s = 0$ (i.e., no shared individuals between the GWASs), the unbiased estimator is simply $\widehat{\rho}_{g,local} = \widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{-1}\widehat{\boldsymbol{\gamma}}_{gwas}$. An interpretation for this simple formula is that in the absence of sample overlap, the covariance in the noise, ϵ and δ, is 0 and does thus not introduce bias into the estimate of $\rho_{g,local}$.

Following bilinear form theory,[23] we can estimate the variance for $\widehat{\rho}_{g,local}$ as

$$\begin{aligned}\mathrm{Var}\left[\widehat{\rho}_{g,local}\right] = \left(\frac{n_1 n_2}{n_1 n_2 - n_s p}\right)^2 \Bigg[\left(\frac{p\rho_e n_s}{n_1 n_2}\right)^2 + \frac{\sigma_\epsilon^2\sigma_\delta^2 p}{n_1 n_2} + \frac{\sigma_\delta^2 h_{g\phi,local}^2}{n_2} \\ + \frac{\sigma_\epsilon^2 h_{g\psi,local}^2}{n_1} + 2\frac{n_s\rho_e\rho_{g,local}}{n_1 n_2}\Bigg].\end{aligned} \qquad \text{(Equation 10)}$$

For rank deficient LD matrix with $\mathrm{rank}(\boldsymbol{V}) = q$, one replaces p with q in Equation 10.

Accounting for Statistical Noise in LD Estimates

Limited sample size of external reference panels creates statistical noise in the estimated LD matrix that biases our estimates. Following our previous work,[19] we apply truncated-SVD regularization[24] to remove noise in external reference LD. We note that $\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{\dagger}\widehat{\boldsymbol{\gamma}}_{gwas} = \sum_{i=1}^{q}s_i = \sum_{i=1}^{q}\frac{1}{w_i}(\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{u}_i)(\widehat{\boldsymbol{\gamma}}_{gwas}^{\top}\boldsymbol{u}_i)$, where w_i and $\boldsymbol{u}_i$ are the eigenvalues and eigenvectors of the LD matrix $\boldsymbol{V}$ and $q = \mathrm{rank}(\boldsymbol{V})$. We use $\widehat{s}_i = \frac{1}{\widehat{w}_i}(\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\widehat{\boldsymbol{u}}_i)(\widehat{\boldsymbol{\gamma}}_{gwas}^{\top}\widehat{\boldsymbol{u}}_i)$ to denote the counterpart obtained from external reference LD matrix $\widehat{\boldsymbol{V}}$. We show through simulations that the bulk of $\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{\dagger}\widehat{\boldsymbol{\gamma}}_{gwas}$ comes from s_i where $i \ll q$ and that $s_i \approx \widehat{s}_i$ for $i \ll q$, thus justifying truncated-SVD as an appropriate regularization method when only external reference LD ($\widehat{\boldsymbol{V}}$) is available.

Let $g(\widehat{\boldsymbol{\beta}}_{gwas}, \widehat{\boldsymbol{\gamma}}_{gwas}, k) = \sum_{i=1}^{k}\widehat{s}_i = \sum_{i=1}^{k}\frac{1}{\widehat{w}_i}(\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\widehat{\boldsymbol{u}}_i)(\widehat{\boldsymbol{\gamma}}_{gwas}^{\top}\widehat{\boldsymbol{u}}_i)$ be the truncated-SVD regularized estimates for $\widehat{\boldsymbol{\beta}}_{gwas}^{\top}\boldsymbol{V}^{\dagger}\widehat{\boldsymbol{\gamma}}_{gwas}$, then it can be shown that

$$\mathrm{E}\left[g\left(\widehat{\boldsymbol{\beta}}_{gwas}, \widehat{\boldsymbol{\gamma}}_{gwas}, k\right)\right] = \frac{n_s k(\rho - \rho_g)}{n_1 n_2} + \sum_{i=1}^{k}\widehat{w}_i(\boldsymbol{\beta}^{\top}\widehat{\boldsymbol{u}}_i)(\boldsymbol{\gamma}^{\top}\widehat{\boldsymbol{u}}_i). \qquad \text{(Equation 11)}$$

Assuming $\widehat{w}_i = w_i$ and $\widehat{\boldsymbol{u}}_i = \boldsymbol{u}_i$ for $i \ll k$, Equation 11 is a biased approximation of $\rho_{g,local}$, with bias $\frac{n_s k(\rho - \rho_g)}{n_1 n_2}$. Correcting for the bias, we arrive at the estimator

$$\widehat{\rho}_{g,local} = \frac{n_1 n_2 g\left(\widehat{\boldsymbol{\beta}}_{gwas}, \widehat{\boldsymbol{\gamma}}_{gwas}, k\right) - n_s \rho k}{n_1 n_2 - n_s k}, \qquad \text{(Equation 12)}$$

which has variance

$$\begin{aligned}\mathrm{Var}\left[\widehat{\rho}_{g,local}\right] = \left(\frac{n_1 n_2}{n_1 n_2 - n_s k}\right)^2 \Bigg[\left(\frac{k\rho_e n_s}{n_1 n_2}\right)^2 + \frac{\sigma_\epsilon^2\sigma_\delta^2 k}{n_1 n_2} + \frac{\sigma_\delta^2 h_{g\phi,local}^2}{n_2} \\ + \frac{\sigma_\epsilon^2 h_{g\psi,local}^2}{n_1} + 2\frac{n_s\rho_e\rho_{g,local}}{n_1 n_2}\Bigg].\end{aligned} \qquad \text{(Equation 13)}$$

Extension to Multiple Independent Regions

For genome partitioned into m regions, let

$$\begin{aligned}\phi &= \boldsymbol{x}_1^{\top}\boldsymbol{\beta}_1 + \cdots + \boldsymbol{x}_m^{\top}\boldsymbol{\beta}_m + \epsilon \\ \psi &= \boldsymbol{x}_1^{\top}\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{x}_m^{\top}\boldsymbol{\gamma}_m + \delta\end{aligned} \qquad \text{(Equation 14)}$$

denote the phenotype measurements of two traits at an individuals, where we assume that SNPs in different pairs of regions are independent, i.e., $\mathrm{E}[\boldsymbol{x}_{ik}\boldsymbol{x}_{jl}] = 0$ for all $i \neq j$, $k \in \{1, \cdots, p_i\}$, and $l \in \{1, \cdots, p_j\}$, where p_i and p_j are the number of SNPs in region i and j. Under these assumptions, we decompose the phenotypic covariance, ρ, between ϕ and ψ, into a summation of per-region genetic covariance and environmental covariance

$$\begin{aligned}\rho = \mathrm{Cov}[\phi, \psi] &= \mathrm{E}\left[\left(\boldsymbol{x}_1^{\top}\boldsymbol{\beta}_1 + \cdots + \boldsymbol{x}_m^{\top}\boldsymbol{\beta}_m + \epsilon\right)\left(\boldsymbol{x}_1^{\top}\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{x}_m^{\top}\boldsymbol{\gamma}_m + \delta\right)^{\top}\right] \\ &= \mathrm{E}\left[\left(\boldsymbol{x}_1^{\top}\boldsymbol{\beta}_1\right)\left(\boldsymbol{x}_1^{\top}\boldsymbol{\gamma}_1\right)\right] + \cdots + \mathrm{E}\left[\left(\boldsymbol{x}_m^{\top}\boldsymbol{\beta}_m\right)\left(\boldsymbol{x}_m^{\top}\boldsymbol{\gamma}_m\right)\right] + \mathrm{E}[\epsilon\delta] \\ &= \sum_{i=1}^{m}\mathrm{Cov}\left[\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_i, \boldsymbol{x}_i^{\top}\boldsymbol{\gamma}_i\right] + \mathrm{Cov}[\epsilon, \delta] = \sum_{i=1}^{m}\boldsymbol{\beta}_i^{\top}\boldsymbol{V}_i\boldsymbol{\gamma}_i + \rho_e\end{aligned}, \qquad \text{(Equation 15)}$$

where $\rho_{g,local,i} = \mathrm{Cov}[\boldsymbol{x}_i^\top \boldsymbol{\beta}_i, \boldsymbol{x}_i^\top \boldsymbol{\gamma}_i] = \boldsymbol{\beta}_i^\top \boldsymbol{V}_i \boldsymbol{\gamma}_i$ is the local genetic covariance between the pair of traits attributed to genetic variants at region i. Following strategies outlined in previous sections, we arrive at the estimator for genetic covariance at the i^{th} region,

$$\widehat{\rho}_{g,local,i} = \frac{n_1 n_2 g\left(\widehat{\boldsymbol{\beta}}_{gwas,i}, \widehat{\boldsymbol{\gamma}}_{gwas,i}, k\right) - n_s\left(\rho - \sum_{j=1,j\neq i}^{m} \widehat{\rho}_{g,local,j}\right) k_i}{n_1 n_2 - n_s k_i}, \tag{Equation 16}$$

which defines a system of linear equation involving m unknown variables and m equations. Following bilinear form theory, we obtain variance estimate for $\widehat{\rho}_{g,local,i}$ as

$$\begin{aligned}\mathrm{Var}\left[\widehat{\rho}_{g,local,i}\right] = \left(\frac{n_1 n_2}{n_1 n_2 - n_s k_i}\right)^2 \Bigg[\left(\frac{k_i \rho_e n_s}{n_1 n_2}\right)^2 + \frac{\sigma_\epsilon^2 \sigma_\delta^2 k_i}{n_1 n_2} + \frac{\sigma_\delta^2 h_{g\phi,local,i}^2}{n_2} \\ + \frac{\sigma_\epsilon^2 h_{g\psi,local,i}^2}{n_1} + 2\frac{n_s \rho_e \rho_{g,local,i}}{n_1 n_2}\Bigg] \quad , \\ + \sum_{j=1,j\neq i}^{m} \left(\frac{n_s k_j}{n_1 n_2 - n_s k_i}\right)^2 \mathrm{Var}\left[\widehat{\rho}_{g,local,j}\right]\end{aligned} \tag{Equation 17}$$

which also defines a system of linear equations with m equations and m variables. In the special case where there is no sample overlap ($n_s = 0$), $\widehat{\rho}_{g,local,i}$ reduces to $g(\widehat{\boldsymbol{\beta}}_{gwas}, \widehat{\boldsymbol{\gamma}}_{gwas}, k)$ with $\mathrm{Var}[\widehat{\rho}_{g,local,i}] = \frac{\sigma_\epsilon^2 \sigma_\delta^2 k_i}{n_1 n_2} + \frac{\sigma_\delta^2 h_{g\phi,local,i}^2}{n_2} + \frac{\sigma_\epsilon^2 h_{g\psi,local,i}^2}{n_1} \approx \frac{\sigma_\epsilon^2 \sigma_\delta^2 k_i}{n_1 n_2}$, i.e., both the local genetic covariance and its variance can be estimated independent of all other windows.

When $k_1 = \cdots = k_m = k$, i.e., all regions use the same number of eigenvectors in the truncated-SVD regularization, summing over i on both sides of Equation 16 yields

$$\begin{aligned}\widehat{\rho}_g = \sum_{i=1} \widehat{\rho}_{g,local,i} &= \frac{n_1 n_2}{n_1 n_2 - n_s k} \sum_{i=1}^{m} g\left(\widehat{\boldsymbol{\beta}}_{gwas,i}, \widehat{\boldsymbol{\gamma}}_{gwas,i}, k\right) - \frac{k n_s}{n_1 n_2 - n_s k} \sum_{i=1}^{m}\left(\rho - \sum_{j=1,j\neq i}^{m} \widehat{\rho}_{g,local,j}\right) \\ &= \frac{n_1 n_2}{n_1 n_2 - n_s k} \sum_{i=1}^{m} g\left(\widehat{\boldsymbol{\beta}}_{gwas,i}, \widehat{\boldsymbol{\gamma}}_{gwas,i}, k\right) - \frac{k n_s}{n_1 n_2 - n_s k} \sum_{i=1}^{m}\left(\rho - \widehat{\rho}_g + \widehat{\rho}_{g,local,i}\right) \\ &= \frac{n_1 n_2}{n_1 n_2 - n_s k} \sum_{i=1}^{m} g\left(\widehat{\boldsymbol{\beta}}_{gwas,i}, \widehat{\boldsymbol{\gamma}}_{gwas,i}, k\right) + \frac{k n_s m - k n_s}{n_1 n_2 - n_s k} \widehat{\rho}_g - \frac{k n_s m \rho}{n_1 n_2 - n_s k}.\end{aligned} \tag{Equation 18}$$

Solving for $\widehat{\rho}_g$ yields

$$\widehat{\rho}_g = \frac{n_1 n_2 \sum_{i=1}^{m} g\left(\widehat{\boldsymbol{\beta}}_{gwas,i}, \widehat{\boldsymbol{\gamma}}_{gwas,i}, k\right) - k n_s m \rho}{n_1 n_2 - k n_s m}, \tag{Equation 19}$$

which has variance

$$\mathrm{Var}\left[\widehat{\rho}_g\right] = \left(\frac{n_1 n_2}{n_1 n_2 - k n_s m}\right)^2 \sum_{i=1}^{m} \mathrm{Var}\left[g\left(\widehat{\boldsymbol{\beta}}_{gwas,i}, \widehat{\boldsymbol{\gamma}}_{gwas,i}, k\right)\right]. \tag{Equation 20}$$

Thus, if k is chosen such that $(n_1 n_2 - k n_s m)$ is small (i.e., $\frac{n_1 n_2}{n_1 n_2 - k n_s m}$ large), the estimate of total genetic covariance will have large standard error. To reduce standard error in the estimates (at the cost of some bias), we recommend choosing k such that $\frac{n_1 n_2}{n_1 n_2 - k n_s m}$ is less than 2. When testing for statistical significance, we assume that the estimates of local and genome-wide genetic covariance and correlation follow a normal distribution.

Standardizing Local Genetic Covariance

We estimate the local genetic correlation for the i^{th} region as

$$\widehat{r}_{g,local,i} = \frac{\widehat{\rho}_{g,local,i}}{\sqrt{\widehat{h}_{g\phi,local,i}^2}\sqrt{\widehat{h}_{g\psi,local,i}^2}}, \tag{Equation 21}$$

where $\widehat{h}_{g\phi,local,i}^2$ and $\widehat{h}_{g\psi,local,i}^2$ denote the local SNP heritability of trait ϕ and ψ at the i^{th} region. In some cases, this estimator of local genetic correlation may yield an estimate with magnitude greater than 1, and we cap the estimate at -1 or 1. In simulations, we show that $\widehat{r}_{g,local,i}$ is approximately unbiased when both traits are heritable at the i^{th} region. In practice, however, the terms $\widehat{h}_{g\phi,local,i}^2$ and $\widehat{h}_{g\psi,local,i}^2$ can be close to zero, greatly inflating the standard error of $\widehat{r}_{g,local,i}$. Thus, we recommend estimating local genetic correlation only at regions with significant local SNP heritability. One can also estimate local genetic correlation at a set of regions. For example, to estimate genetic correlation at regions indexed by the index set $\boldsymbol{C}$, one applies the formula

$$\widehat{r}_{g,\boldsymbol{C}} = \frac{\sum_{i\in \boldsymbol{C}} \widehat{\rho}_{g,local,i}}{\sqrt{\sum_{i\in \boldsymbol{C}} \widehat{h}_{\phi,g,local,i}^2}\sqrt{\sum_{i\in \boldsymbol{C}} \widehat{h}_{\psi,g,local,i}^2}}. \tag{Equation 22}$$

We estimate standard error of local genetic correlation at a single region through a parametric bootstrap approach[25] and local genetic correlation at a set of regions through jackknife.

Simulation Framework

Starting from half (202 individuals) of the EUR reference panel from the 1000 Genomes Project,[26] we simulated genotype data for 50,000 individuals at HapMap3[27] SNPs with minor allele frequency (MAF) greater than 5% in 100 randomly selected LD-independent regions defined in Berisa and Pickrell[18] on chromosome 1 using HAPGEN2.[27] We used the other half of the EUR reference panel (203 individuals) to obtain external reference LD matrices.

We simulated phenotypes from the genotypes according to the linear model $\phi = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ and $\boldsymbol{\psi} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\delta}$, where $\boldsymbol{X}$ is the column-standardized genotype matrix. We drew the effects of causal SNPs ($\boldsymbol{\beta}_C$, $\boldsymbol{\gamma}_C$) from the distribution

$$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\frac{h_{g\phi}^2}{|C|}\boldsymbol{I} & \frac{\rho_g}{|C|}\boldsymbol{I} \\ \frac{\rho_g}{|C|}\boldsymbol{I} & \frac{h_{g\psi}^2}{|C|}\boldsymbol{I}\end{bmatrix}\right), \tag{Equation 23}$$

where C is the index set of causal SNPs, and set the effects of all other SNPs to be zero. We then drew (ϵ, δ) from the distribution

$$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\left(1-h^2_{g\phi}\right)\boldsymbol{I} & \rho_e\boldsymbol{I}\\ \rho_e\boldsymbol{I} & \left(1-h^2_{g\psi}\right)\boldsymbol{I}\end{bmatrix}\right). \qquad \text{(Equation 24)}$$

Finally, we simulated GWAS summary statistics using methods outlined in previous sections. For each $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ drawn from the normal distribution, we simulated 1,000 sets of summary statistics by varying ϵ and $\boldsymbol{\delta}$ and applied ρ-HESS to estimate genetic covariance and genetic correlation for each set of the simulated summary statistics.

Empirical Datasets

We obtained GWAS summary data for 36 quantitative complex traits and diseases from 15 GWAS consortia or institutions (see Table 1), all of which are based on individuals of European ancestry and have sample size greater than 20,000. We used approximately independent genomic regions previously defined[18] to partition the genome and restricted our analyses on HapMap3 SNPs with minor allele frequency (MAF) greater than 5% in the European population in the 1000 Genomes data.[26] We also removed stand-ambiguous SNPs prior to our analyses. We follow the method previously outlined[19] to estimate and re-inflate λ_{gc} and to choose the number of eigenvectors to include in estimating local genetic covariance and SNP heritability.

Local Genetic Correlation at Regions Ascertained for GWAS Signals

Recent works leverage the difference in correlations of Z-scores at genomic regions ascertained for GWAS signals specific to each trait to prioritize putative causal models between pairs of complex traits.[2,3] We evaluated the local genetic correlation at regions harboring GWAS signals specific to each trait across all 298 pairs of traits exhibiting significant genome-wide genetic correlation. We estimate local genetic correlations only for pairs of traits for which the number of loci harboring GWAS hits specific to each trait is greater than 10. The confidence intervals (1.96 times jackknife standard error on each side) of the ascertained local genetic correlations ($\widehat{r}_{g,local,trait1}$ and $\widehat{r}_{g,local,trait2}$) do not overlap; one of the confidence intervals overlap with 0 and the other does not.

Results

Local Genetic Correlation Estimation in Simulations

We evaluated the performance of our approach (ρ-HESS) through simulations across a wide range of disease architectures. We included cross-trait LDSC,[16] an approach that assumes a random-effect model, in the comparison for completeness purposes. When LD is estimated in-sample, ρ-HESS provides an unbiased estimate of local genetic covariance and nearly unbiased estimates of genetic correlation (i.e., genetic covariance divided by the square root of local SNP heritability, see Material and Methods) (Figure S2). Next, we quantified the performance in the more realistic case when in-sample LD is unavailable and needs to be estimated from external reference panels. Although both cross-trait LDSC and ρ-HESS provide accurate estimates of genetic correlation, we observe superior accuracy with higher precision for ρ-HESS (Figures 3, S4, S6, and S7). We attribute the lower standard error of ρ-HESS to the truncated-SVD regularization of the LD matrix which effectively reduces the degree of freedom of the bi-linear form in Equation 7 (Figure S10). Different genomic regions vary in their total amount of LD and we observed that the accuracy of genetic correlation estimation decreases with the total amount of regional LD (Figure S11). This is expected as high LD regions lead to high rank deficiencies in the LD matrix and small eigenvalues, thus increasing the level of statistical noise in the estimation. We also evaluated the performance of local genetic correlation estimation in simulations where we varied the number of causal variants in each region. Overall, we observe that our estimator of genetic covariance and correlation is not sensitive to the underlying polygenicity (i.e., number of causal SNPs) (Figures 3, S5, S8, and S9). Finally, we also evaluated the performance of the estimator when causal variants are all drawn from DHS regions[48] and observed that the performance is not sensitive to the uneven distribution of causal variants (Figure S3).

Local Genetic Correlation across 36 Quantitative Traits

We analyzed GWAS summary data from 36 complex traits to obtain local genetic correlations at 1,703 approximately LD-independent regions in the genome (~1.6 Mb in width on average).[18] First, as a quality control step, we aggregated the local estimates into genome-wide estimates of genetic correlation (see Material and Methods) and compared to the cross-trait LDSC estimates. Reassuringly, we find a high degree of consistency with genetic correlations estimated by cross-trait LDSC regression ($R = 0.77$; Figures 4 and S13). Our estimator provides lower standard errors as compared to cross-trait LDSC (likely due to the truncated-SVD regularization procedure) and yields consistently lower estimates for pairs of traits from the same consortium where we conservatively assume full sample overlap (see Discussion). Overall, we identify 298 pairs of traits with significant genome-wide genetic correlation ($p < 0.05/630$). These include previously reported correlations, e.g., body mass index (BMI) and triglyceride (TG), as well as complex traits that have not been studied before using genetic correlation, e.g., red blood cell count (RBC) and fasting insulin (FI) (Figure 4).

Next, we searched for genomic regions that disproportionately contribute to the genetic correlation of the 36 analyzed traits; we excluded the HLA region due to complex LD patterns. We identify 25 genomic regions that show both significant local genetic correlation (two-tailed $p < 0.05/1{,}703$) as well as significant local SNP heritability (one-tailed $p < 0.05/1703$) (see Table 2, Figures S14–S16). For example, the estimate of local genetic correlation between HDL and TG at chr11: 116–117 Mb is −0.82 (95% CI [−0.95, −0.69]), suggesting highly shared genetic architecture at this region for HDL and TG.

Table 1. A Summary of the 36 GWAS Summary Datasets Analyzed

Trait Name	Abbreviation	Consortium	# Gen Corr All Consortium	# Gen Corr outside Consortium	Approx. Sample Size
Age at menarche[28]	AM	REPROGEN	21 (4)	21 (4)	133K
Body mass index[29]	BMI	GIANT	27 (17)	23 (14)	231K
Height[30]	HEIGHT	GIANT	17 (2)	13 (1)	241K
Hip circumference[31]	HIP	GIANT	23 (14)	19 (10)	144K
Waist circumference[31]	WC	GIANT	26 (18)	22 (15)	153K
Waist-to-hip ratio[31]	WHR	GIANT	27 (19)	23 (16)	143K
Haemoglobin[32]	HB	HAEMGEN	21 (10)	18 (8)	51K
Mean cell haemoglobin[32]	MCH	HAEMGEN	9 (1)	8 (1)	44K
MCH concentration[32]	MCHC	HAEMGEN	6 (4)	2 (1)	47K
Mean cell volume[32]	MCV	HAEMGEN	12 (3)	10 (1)	49K
Packed cell volume[32]	PCV	HAEMGEN	18 (11)	14 (8)	45K
Red blood cell count[32]	RBC	HAEMGEN	20 (10)	17 (8)	46K
Number of platelets[33]	PLT	HAEMGEN	9 (1)	6 (1)	67K
Fasting glucose[34]	FG	MAGIC	19 (9)	16 (8)	46K
Fasting insulin[34]	FI	MAGIC	20 (12)	18 (12)	46K
HBA1C[35]	HBA1C	MAGIC	19 (14)	18 (13)	46K
HOMA-B[34]	HOMA-B	MAGIC	17 (11)	15 (11)	46K
HOMA-IR[34]	HOMA-IR	MAGIC	21 (12)	21 (12)	46K
High-density lipoprotein[36]	HDL	GLGC	23 (12)	21 (11)	96K
Low-density lipoprotein[36]	LDL	GLGC	19 (6)	17 (4)	91K
Total cholesterol[36]	TC	GLGC	18 (3)	15 (1)	96K
Triglycerides[36]	TG	GLGC	26 (14)	23 (11)	92K
Forearm BMD[37]	FA	GEFOS	4 (1)	2 (0)	53K
Femoral neck BMD[37]	FN	GEFOS	4 (2)	2 (0)	53K
Lumbar spine BMD[37]	LS	GEFOS	7 (1)	5 (0)	53K
Education years[38]	EY	SSGAC	26 (5)	24 (4)	294K
Neuroticism[39]	NEURO	SSGAC	5 (2)	3 (0)	171K
Subjective well-being[39]	SWB	SSGAC	4 (1)	2 (0)	298K
Age first birth[40]	AFB	BIOS	23 (5)	23 (5)	251K
Birth weight[41]	BW	EGG	13 (1)	13 (1)	68K
Urinary albumin-to-creatinine ratio[42]	UACR	DCCT-EDIC	11 (1)	11 (1)	53K
Rest heart rate[43]	HR	EPPINGA	14 (0)	14 (0)	265K
Serum urate concentrations[44]	URATE	GUGC	25 (14)	25 (14)	107K
Body fat[45]	BF	Lu et al.	26 (17)	26 (17)	58K
Extra-glomerular filtration rate of creatinin[46]	CRN	CKDGEN	10 (1)	10 (1)	133K
Age at menopause[47]	MP	BCAC	6 (0)	6 (0)	70K

We list the total number of traits with significant non-zero genome-wide genetic correlation (two-tailed $p < 0.05/630$) and the total number of traits outside the consortium with significant non-zero genome-wide genetic correlation in the fourth and fifth column, respectively. Number of traits for which the magnitude of genetic correlation is both significantly non-zero and greater than 0.2 is shown in parentheses.

Indeed, the region chr11: 116M–117M harbors *APOA1* (MIM: 107680), which is known to be associated with multiple lipid traits.[36] Interestingly, 4 out of the 25 regions do not contain GWAS-significant SNPs ($p < 5 \times 10^{-8}$) for either one or both traits and can be viewed as new risk regions for these traits.

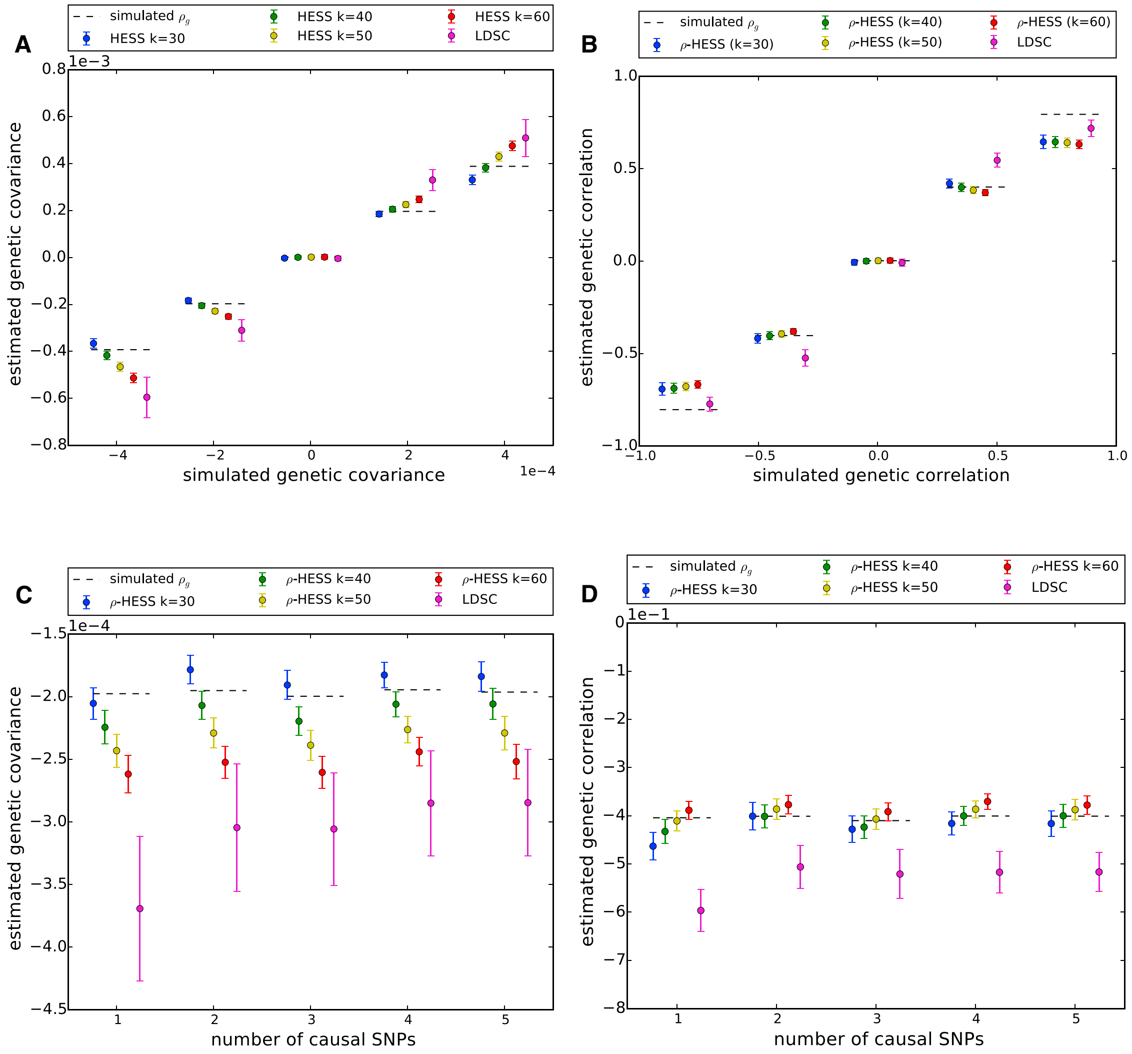

Figure 3. Performance of ρ-HESS and Cross-trait LDSC using External Reference LD across 100 LD-Independent Regions, with Each Region Having 1,000 Simulations

Here, each dot represents the mean (more than 100 regions) of the average performance (more than 1,000 simulations per region), with error bars representing 1.96 times the standard error on both sides. Overall, ρ-HESS provides approximately unbiased estimates of local genetic covariance (A) and correlation (B) and is not sensitive to the underlying genetic architectures (covariance in C and correlation in D). We also observe that ρ-HESS is less biased, is more consistent, and has smaller standard error than cross-trait LDSC.

Since genetic correlation is an aggregation of local genetic covariance, for pairs of traits with highly positive or negative genetic correlation, we expect the distribution of local genetic covariances to be shifted toward the positive or negative side (see Figure S17), whereas for pairs of traits with low genetic correlation, we expect the distribution of local genetic covariances to be centered around zero (see Figures 5 and6). Indeed, pairs of traits with higher genome-wide genetic correlation tend to harbor more loci with significant local genetic covariance (see Figure S14). For instance, only one region exhibits significant local genetic covariance for the pair of traits age at menarche (AM) and height ($r_g = 0.13$, 95% CI [0.10, 0.13]), whereas four loci show significant local genetic covariance for the pair of traits LDL and TG ($r_g = 0.45$, 95% CI [0.42, 0.49]).

Local Correlations for Pairs of Traits with Negligible Genome-wide Correlation

Several pairs of traits show negligible genome-wide genetic correlation although they share GWAS risk regions. For example HDL and LDL share several GWAS risk loci[36] but the genome-wide genetic correlation is negligible (−0.05, 95% CI [−0.09, −0.01]).[16] The absence of significant genome-wide genetic correlation between these pairs of traits can be attributed to either symmetric distribution of local genetic covariance (positive local genetic covariance cancels out negative local genetic covariance, see Figure 1) and/or lack of power to declare significance for genome-wide genetic correlation. Thus, we hypothesize that at the region-specific level, many loci may manifest significant local genetic covariance even if the genome-wide genetic correlation between a pair of traits is not significant. Indeed, 11 genomic regions show significant local

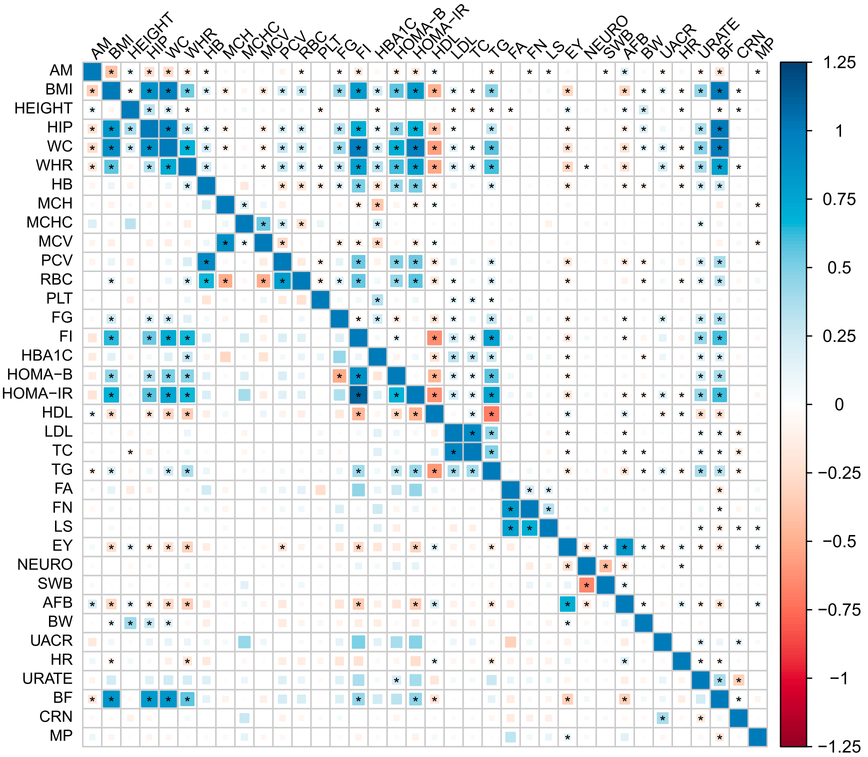

Figure 4. Genetic Correlation across the 36 Complex Traits Obtained by ρ-HESS and Cross-trait LDSC[17]

The magnitude of the correlation is represented by the color and the size of the square. Among the 630 pairs of traits, ρ-HESS (top half) (cross-trait LDSC [bottom half]) identified 298 (115) pairs showing significant genetic correlation (marked with dots).

genetic correlation (two-tailed $p < 0.05/1{,}703$) for HDL and LDL (see Figure 5). Some of these loci, e.g., chr2: 21M–23M, chr11: 116M–117M, and chr19: 44M–46M, harbor *APOB*, *APOA1*, and *APOE* (MIM: 107741), respectively, which are known to be involved in lipid genetics.[36,49,50] Across all pairs of traits with non-significant genome-wide correlation, we identify 6 regions across 10 pairs of traits with significant local genetic correlation (two-tailed $p < 0.05/1{,}703$) and local SNP heritability (one-tailed $p < 0.05/1{,}703$) (see Table 2, Figure S16). For example, the region chr6: 134M–136M harbors the *HBS1L* (MIM: 612450)[32,51] and contributes to local genetic covariance across many blood traits (MCH, MCV, RBC, and PLT).

Genetic Correlation Ascertained for GWAS Risk Loci

Assessing the correlation in the effects at genomic regions ascertained for trait-specific GWAS regions can be used to prioritize putative causal models between complex traits. We utilized a recently proposed approach[2] to assign putative causal relation to 55 pairs of traits. Restricting to 40 of the 55 pairs of traits that contain at least 10 regions with trait-specific GWAS signals (see Material and Methods), we quantified the local genetic correlation at genomic regions containing GWAS loci specific to each trait (see Table S2, Figure 7). Overall, the local genetic correlation is highly consistent with the putative causal relationships inferred by correlating the top signals at these loci.[2] For example, when considering body mass index (BMI) and triglyceride levels (TG), the correlation at BMI-specific regions is significantly greater than TG-specific loci ($\widehat{r}_{g,local,BM} = 0.47$ 95% CI [0.37, 0.57] versus $\widehat{r}_{g,local,TG} = -0.02$ 95% [−0.14, 0.10]), indicating that loci that increase BMI tend to consistently increase TG, whereas loci that increase TG do not consistently affect BMI, consistent with the putative model that BMI causally increases TG (see Figure 6).[2,3] We also observe correlations consistent with a model in which years of education (EY) consistently decreases hemoglobin level (HB), LDL, and TG (see Table S2), in line with previous conclusions on the effect of education on health.[52,53] However, we note that education attainment (or other studied traits) may be confounded by other factors such as social status and that one should exercise caution when inferring causality from genetic data. Finally, we also report pairs of traits in which the genetic correlation approach attains different results from bi-directional regression on the top signals.[2] For example, when considering body mass index (BMI) and age at menarche (AM), the local correlation approach do not yield different estimates ($r_{g,local,BMI} = -0.49$ 95% CI [−0.63, −0.35] versus $r_{g,local,AM} = -0.47$ 95% CI [−0.59, −0.35]), whereas the approach of Joseph et al.[2] suggests a putative causal relation. This discrepancy can be due to different model assumptions, e.g., single causal variant versus allelic heterogeneity, with further investigations needed to assign causality from these data.

Discussion

We have described ρ-HESS, a method to estimate local genetic correlation from GWAS summary association data. Through extensive simulations, we demonstrated that our method is approximately unbiased and provides consistent results irrespective of causal architecture. We analyzed large-scale GWAS summary association data of 36 quantitative traits. Compared with cross-trait LDSC, our methods identified considerably more pairs of traits displaying significant genome-wide genetic correlation likely because of the truncated-SVD regularization of the

Table 2. Loci that Show Significant Local Genetic Covariance (Two-Tailed p $<$ 0.05/1,703) and Local SNP Heritability (One-Tailed p $<$ 0.05/1,703) for Both Traits

Trait1	Trait2	Locus	$h^2_{g,local,trait1}$	$h^2_{g,local,trait2}$	$r_{g,local}$
AM	HEIGHT	chr9: 107M–109M	0.15 (0.02)	0.05 (0.01)	0.61 ([0.34,0.87])
BMI	HIP	chr16: 53M–55M	0.22 (0.02)	0.19 (0.03)	0.99 ([0.76,1.00])
BMI	HIP	chr18: 57M–59M	0.14 (0.02)	0.13 (0.02)	0.99 ([0.71,1.00])
BMI	WC	chr16: 53M–55M	0.22 (0.02)	0.21 (0.03)	1.00 ([0.78,1.00])
BMI	WC	chr18: 57M–59M	0.14 (0.02)	0.13 (0.02)	1.00 ([0.72,1.00])
BW	HEIGHT	chr12: 65M–67M	0.14 (0.02)	0.23 (0.02)	0.93 ([0.70,1.00])
HDL	TG	chr2: 21M–23M	0.16 (0.03)	0.22 (0.03)	−0.94 ([−1.00, −0.65])
HDL	TG	chr8: 19M–20M	0.65 (0.04)	0.82 (0.04)	−1.00 ([−1.00, −0.91])
HDL	TG	chr11: 116M–117M	0.40 (0.04)	1.27 (0.06)	−0.82 ([−0.95,-0.69])
HDL	TG	chr15: 58M–59M	1.18 (0.06)	0.18 (0.03)	0.89 ([0.68,1.00])
HEIGHT	HIP	chr16: 4M–5M	0.06 (0.01)	0.10 (0.02)	0.73 ([0.41,1.00])
HIP	WC	chr16: 53M–55M	0.19 (0.03)	0.21 (0.03)	0.99 ([0.73,1.00])
HIP	WC	chr18: 57M–59M	0.13 (0.02)	0.13 (0.02)	1.00 ([0.69,1.00])
LDL	TG	chr1: 61M–63M	0.14 (0.03)	0.28 (0.03)	0.98 ([0.67,1.00])
LDL	TG	chr2: 21M–23M	0.84 (0.05)	0.22 (0.03)	0.62 ([0.46,0.78])
LDL	TG	chr8: 126M–128M	0.16 (0.03)	0.32 (0.04)	0.94 ([0.63,1.00])
LDL	TG	chr19: 18M–19M	0.18 (0.03)	0.21 (0.03)	0.99 ([0.72,1.00])
PLT	RBC	chr6: 134M–136M	0.26 (0.05)	0.66 (0.09)	−0.99 ([−1.00, −0.69])
HDL	HEIGHT	chr11: 47M–49M	0.17 (0.02)	0.07 (0.01)	0.61 ([0.42,0.80])
HDL	LDL	chr2: 21M–23M	0.16 (0.03)	0.84 (0.05)	−0.56 ([−0.74, −0.39])
HDL	LDL	chr8: 9M–9M	0.14 (0.02)	0.12 (0.02)	0.99 ([0.70,1.00])
MCH	MCV	chr6: 24M–25M	0.49 (0.07)	0.37 (0.06)	0.97 ([0.67,1.00])
MCH	MCV	chr6: 134M–136M	0.86 (0.09)	0.70 (0.08)	0.98 ([0.76,1.00])
MCH	PLT	chr6: 134M–136M	0.86 (0.09)	0.26 (0.05)	1.00 ([0.72,1.00])
MCH	RBC	chr6: 134M–136M	0.86 (0.09)	0.66 (0.09)	−0.98 ([−1.00, −0.75])
MCV	PLT	chr6: 134M–136M	0.70 (0.08)	0.26 (0.05)	1.00 ([0.72,1.00])
MCV	RBC	chr6: 134M–136M	0.70 (0.08)	0.66 (0.09)	−0.98 ([−1.00, −0.74])
MP	HEIGHT	chr5: 175M–177M	0.31 (0.04)	0.10 (0.01)	−0.63 ([−0.82, −0.45])
URATE	MCH	chr6: 24M–25M	0.13 (0.02)	0.53 (0.07)	0.56 ([0.33,0.79])
URATE	MCV	chr6: 24M–25M	0.13 (0.02)	0.41 (0.06)	0.66 ([0.39,0.92])

We list pairs of traits for which the genome-wide genetic correlation is significant (two-tailed p $<$ 0.05/630) and negligible in top and bottom half of this table, respectively. Here, we focus only on the pairs of traits excluding TC (see Table S1 for pairs of traits involving TC). Numbers in parentheses represent standard errors for local SNP heritability estimates and 95% confidence intervals for local genetic correlation estimates.

LD matrix, which decreases the standard error of the estimates. We identify genomic regions that are significantly correlated across pairs of traits regardless of the significance of genome-wide correlation. Finally, we performed bi-directional analyses over the local genetic correlations to identify putative causal relationships, and report local genetic correlations at loci harboring GWAS signal specific to each trait.

We conclude with several limitations highlighting areas for future work. First, our estimator requires phenotype correlation between two traits, as well as the number of shared individuals between the two GWASs. We estimate the phenotype correlation through cross-trait LDSC assuming full sample overlap between GWAS within the same consortium and no sample overlap between GWAS across two consortia. Second, we note that our bi-directional analyses over local genetic correlation can be further extrapolated to infer putative causal models between complex traits. We refrain from making conclusive causal inferences from the bi-directional

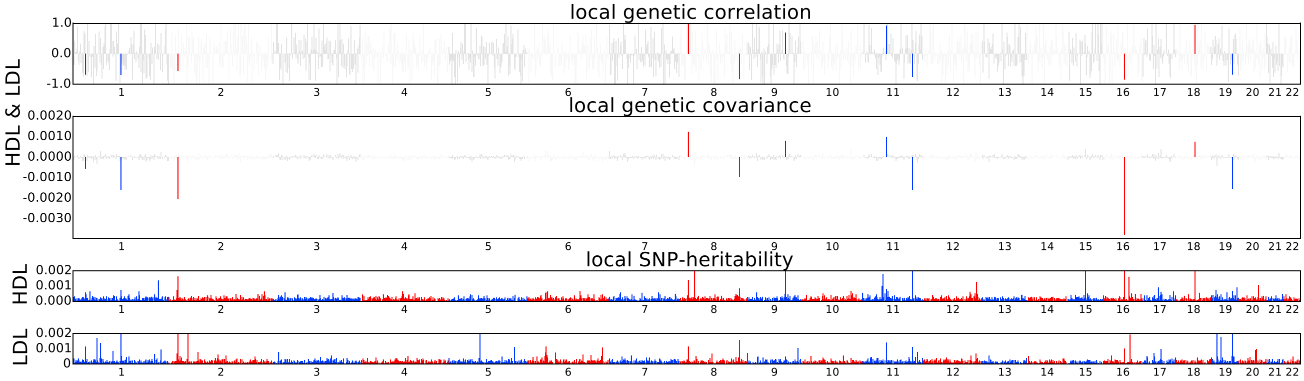

Figure 5. Manhattan-Style Plots Showing the Estimates of Local Genetic Covariance for the Pairs of Traits HDL and LDL
Although the genome-wide genetic correlation between HDL and LDL does not reach the significance level ($p < 0.05/630$), 11 loci exhibit significant local genetic covariance.

analyses because exact inference of causal relations is largely complicated by unobserved confounders such as socioeconomic status, population stratification, and/or biological pathways. Furthermore, most of the GWAS summary association data are adjusted for covariates such as age and gender to increase statistical power,[54] and previous works have shown that adjusting for covariates can potentially lead to false positives.[55] Third, in our real data analyses, we made the assumption that the loci are independent of each other. In reality, however, correlations may exist across adjacent loci due to long-range LD and can lead to biased estimates. Nevertheless, we note that previous works have indicated the effect of LD leakage to be minimal,[19,56] and we conjecture that this statement still holds in estimating local genetic correlation. Lastly, we use truncated-SVD to regularize LD matrix and to reduce standard error in the estimates of local genetic correlation, at the cost of introducing bias. Currently, we use a fixed number of eigenvectors in the truncated-SVD regularization, across all the loci. However, this approach may not be optimal for genomic regions with different LD structure and leave a principled approach of estimating the number of eigenvectors as future work.

Appendix A

Quantifying Shared Genetics via Covariance of the Causal Effects

An alternative measure of shared genetics is the covariance of the causal effects ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) of the two traits. Under the fixed-effect model, we define covariance of the causal effects, $\rho_{g,causal}$, as the dot product between the causal effect size vectors of the two traits,

$$\rho_{g,causal} = \boldsymbol{\beta}^{\top}\boldsymbol{\gamma}. \qquad \text{(Equation A1)}$$

Here, we make the assumption that the average effect size of each SNP is 0.

The definition of covariance of the causal effects in Equation A1 coincides with genetic covariance under the random-effect model. As shown in the supplementary data of Bulik-Sullivan et al.,[16] if one assumes that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have zero mean and

$$\mathrm{Var}[(\boldsymbol{\beta},\boldsymbol{\gamma})] = \frac{1}{p}\begin{bmatrix} h^2_{g\phi} & \rho_g \\ \rho_g & h^2_{g\psi} \end{bmatrix}, \qquad \text{(Equation A2)}$$

then it can be shown that the genetic covariance between two traits is

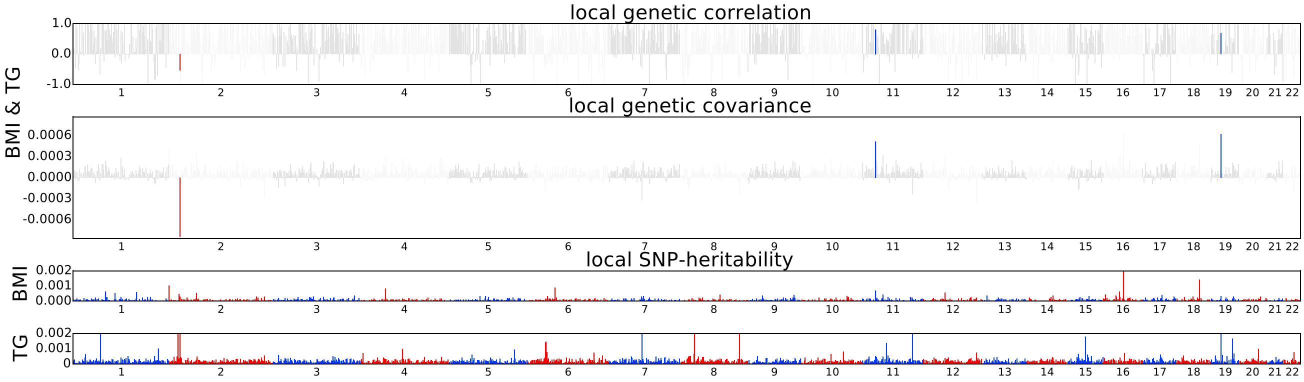

Figure 6. Manhattan-Style Plots Showing the Estimates of Local Genetic Covariance for the Pairs of Traits BMI and TG
That the local genetic covariance between BMI and TG is mostly one-sided implies plausible causal relationship between the two traits.

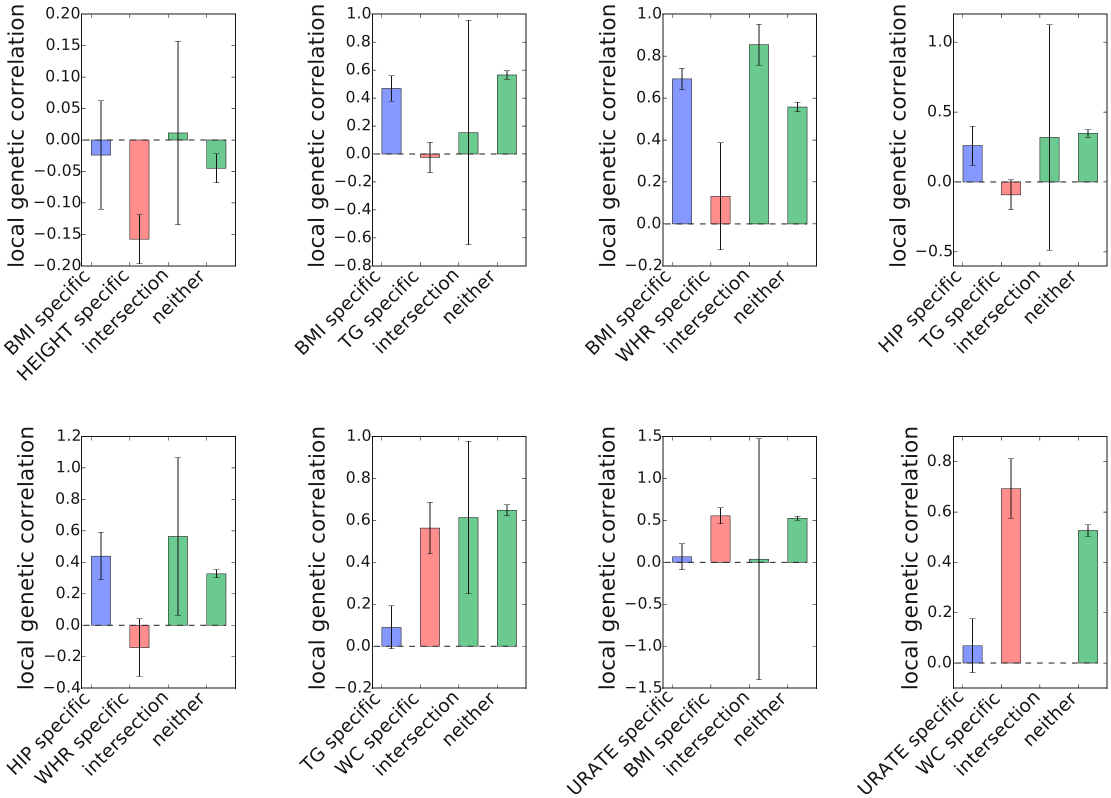

Figure 7. Estimates of Local Genetic Correlation at Loci Ascertained for GWAS Risk Variants for Eight Example Pairs of Traits that Show Plausible Causal Relationship

We obtained standard error using a jackknife approach. Error bars represent 1.96 times the standard error on each side.

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{x}^\top\boldsymbol{\beta},\boldsymbol{x}^\top\boldsymbol{\gamma}] &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{\beta}_i\boldsymbol{\gamma}_j\right] = \sum_{i=1}^{p}\mathrm{E}\left[\boldsymbol{x}_i^2\boldsymbol{\beta}_i\boldsymbol{\gamma}_i\right]\\ &= \sum_{i=1}^{p}\mathrm{E}\left[\boldsymbol{x}_i^2\right]\mathrm{E}\left[\boldsymbol{\beta}_i\boldsymbol{\gamma}_i\right] = \rho_g.\end{aligned} \qquad \text{(Equation A3)}$$

The random-effect model makes the implicit assumption that many SNPs are causal, which is appropriate for genome-wide analysis but not for local analysis, where few SNPs are likely to be causal.

Estimating Covariance of the Causal Effects from GWAS Summary Data

For completeness, we derive an estimator for $\rho_{g,causal}$. We assume a linear model for the two traits (see Material and Methods). The effect size estimates from GWAS, $\widehat{\boldsymbol{\beta}}_{gwas}$ and $\widehat{\boldsymbol{\gamma}}_{gwas}$, follow $\widehat{\boldsymbol{\beta}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\beta}, \frac{1-h_\phi^2}{n_1}\boldsymbol{V}\right)$ and $\widehat{\boldsymbol{\gamma}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\gamma}, \frac{1-h_\psi^2}{n_2}\boldsymbol{V}\right)$, with $\mathrm{Cov}[\widehat{\boldsymbol{\beta}}_{gwas}, \widehat{\boldsymbol{\gamma}}_{gwas}] = \frac{\rho_e n_s}{n_1 n_2}\boldsymbol{V}$, where n_1 and n_2 are the sample size for the two GWASs and n_s is the number of shared samples (see Material and Methods).

As the sample size, n_1 and n_2, of the two GWASs go to infinity, we have $\boldsymbol{\beta}_{gwas} = \lim_{n\to\infty}\widehat{\boldsymbol{\beta}}_{gwas} = \boldsymbol{V}\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_{gwas} = \lim_{n\to\infty}\widehat{\boldsymbol{\gamma}}_{gwas} = \boldsymbol{V}\boldsymbol{\gamma}$, which implies $\boldsymbol{\beta} = \boldsymbol{V}^{-1}\boldsymbol{\beta}_{gwas}$ and $\boldsymbol{\gamma} = \boldsymbol{V}^{-1}\boldsymbol{\gamma}_{gwas}$, suggesting the following estimator for covariance of the causal effects,

$$\rho_{g,causal} = \boldsymbol{\beta}^\top\boldsymbol{\gamma} = \boldsymbol{\beta}_{gwas}^\top\boldsymbol{V}^{-2}\boldsymbol{\gamma}_{gwas}. \qquad \text{(Equation A4)}$$

In reality, however, finite sample sizes of GWAS results in noise in the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, creating bias in the estimate of $\rho_{g,causal}$. From bilinear form theory, it can be shown that

$$\begin{aligned}\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{gwas}^\top\boldsymbol{V}^{-2}\widehat{\boldsymbol{\gamma}}_{gwas}\right] &= \boldsymbol{\beta}^\top\boldsymbol{\gamma} + \frac{\rho_e}{n}\mathrm{tr}\left(\boldsymbol{V}^{-2}\boldsymbol{V}\right)\\ &= \boldsymbol{\beta}^\top\boldsymbol{\gamma} + \frac{\rho_e}{n}\mathrm{tr}\left(\boldsymbol{V}^{-1}\right),\end{aligned} \qquad \text{(Equation A5)}$$

suggesting the unbiased estimator of $\rho_{g,causal}$,

$$\widehat{\rho}_{g,causal} = \widehat{\boldsymbol{\beta}}_{gwas}^\top\boldsymbol{V}^{-2}\widehat{\boldsymbol{\gamma}}_{gwas} - \frac{n_s\rho_e}{n_1 n_2}\mathrm{tr}\left(\boldsymbol{V}^{-1}\right), \qquad \text{(Equation A6)}$$

where the environmental covariance can be estimated through cross-trait LD Score regression.[16]

Supplemental Data

Supplemental Data include 17 figures and 2 tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2017.09.022.

Acknowledgments

This research was supported by NIH (United States Public Health Service) grants R01-HG009120, R01-GM053275, and U01-CA194393. We are grateful to Gleb Kichaev, Malika Kumar, Suraj Alva, and James Boocock for their helpful discussions that greatly improved the quality of this manuscript. We also thank Dr. Nicole Soranzo for kindly sharing summary data for the platelet traits.

Received: April 12, 2017
Accepted: September 22, 2017
Published: November 2, 2017

Web Resources

European Genome-phenome Archive (EGA) (accession number EGAS00000000132), https://www.ebi.ac.uk/ega

GEFOS consortium, http://www.gefos.org/?q=content/data-release-2015

GIANT consortium, https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

GLGC consortium, http://csg.sph.umich.edu//abecasis/public/lipids2013/

HESS and ρ-HESS, http://bogdan.bioinformatics.ucla.edu/software/hess/

LDSC, https://github.com/bulik/ldsc

MAGIC, https://www.magicinvestigators.org/downloads/

OMIM, http://www.omim.org/

Psychiatric Genomics Consortium, https://www.med.unc.edu/pgc/acl_users/credentials_cookie_auth/require_login?came_from=http%3A//www.med.unc.edu/pgc/old-pages/downloads

ReproGen, http://www.reprogen.org/

SSGAC, https://www.thessgac.org/data

References

1. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.
2. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.
3. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am. J. Hum. Genet. *100*, 473–487.
4. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
5. Price, A.L., Spencer, C.C., and Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. Proc. Biol. Sci. *282*, 20151684.
6. Sheehan, N.A., Didelez, V., Burton, P.R., and Tobin, M.D. (2008). Mendelian randomisation and causal inference in observational epidemiology. PLoS Med. *5*, e177.
7. Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., Johnson, T., et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet *380*, 572–580.
8. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat. Med. *27*, 1133–1163.
9. Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. *23* (R1), R89–R98.
10. Smith, G.D., and Ebrahim, S. (2003). ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. *32*, 1–22.
11. 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.
12. 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.
13. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and 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.
14. Neale, M., and Cardon, L. (1992). In Methodology for Genetic Studies of Twins and Families, *Volume 67* (Springer Science & Business Media).
15. Haseman, J.K., and Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. *2*, 3–19.
16. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen Consortium, Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., Perry, J.R., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
17. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.
18. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.
19. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. Am. J. Hum. Genet. *99*, 139–153.
20. Hegmann, J.P., and Possidente, B. (1981). Estimating genetic correlations from inbred strains. Behav. Genet. *11*, 103–114.
21. Carey, G. (1988). Inference about genetic correlations. Behav. Genet. *18*, 329–338.
22. Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika *12*, 134–139.
23. Shayle, R. (1971). Searle. *Linear models* (John Wiley & Sons, Inc.), p. 65.
24. Hansen, P.C. (1987). The truncatedsvd as a method for regularization. BIT *27*, 534–553.

25. Efron, B. (2012). Bayesian inference and the parametric bootstrap. Ann. Appl. Stat. *6*, 1971–1997.

26. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

27. Richard, A.; and International HapMap Consortium (2003). The international hapmap project. Nature *426*, 789–796.

28. Perry, J.R., Day, F., Elks, C.E., Sulem, P., Thompson, D.J., Ferreira, T., He, C., Chasman, D.I., Esko, T., Thorleifsson, G., et al.; Australian Ovarian Cancer Study; GENICA Network; kConFab; LifeLines Cohort Study; InterAct Consortium; and Early Growth Genetics (EGG) Consortium (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature *514*, 92–97.

29. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; and International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.

30. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

31. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Mägi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al.; ADIPOGen Consortium; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GEFOS Consortium; GENIE Consortium; GLGC; ICBP; International Endogene Consortium; LifeLines Cohort Study; MAGIC Investigators; MuTHER Consortium; PAGE Consortium; and ReproGen Consortium (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature *518*, 187–196.

32. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.

33. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature *480*, 201–208.

34. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al.; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium; Anders Hamsten on behalf of Procardis Consortium; and MAGIC investigators (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. *42*, 105–116.

35. Soranzo, N., Sanna, S., Wheeler, E., Gieger, C., Radke, D., Dupuis, J., Bouatia-Naji, N., Langenberg, C., Prokopenko, I., Stolerman, E., et al.; WTCCC (2010). Common variants at 10 genomic loci influence hemoglobin $A_1(C)$ levels via glycemic and nonglycemic pathways. Diabetes *59*, 3229–3239.

36. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.

37. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al.; AOGC Consortium; and UK10K Consortium (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature *526*, 112–117.

38. Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.B., Emilsson, V., Meddens, S.F., et al.; LifeLines Cohort Study (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature *533*, 539–542.

39. Okbay, A., Baselmans, B.M., De Neve, J.E., Turley, P., Nivard, M.G., Fontana, M.A., Meddens, S.F., Linnér, R.K., Rietveld, C.A., Derringer, J., et al. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat. Genet. *48*, 624–633.

40. Barban, N., Jansen, R., de Vlaming, R., Vaez, A., Mandemakers, J.J., Tropf, F.C., Shen, X., Wilson, J.F., Chasman, D.I., Nolte, I.M., et al.; BIOS Consortium; and LifeLines Cohort Study (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. Nat. Genet. *48*, 1462–1472.

41. Horikoshi, M., Beaumont, R.N., Day, F.R., Warrington, N.M., Kooijman, M.N., Fernandez-Tajes, J., Feenstra, B., van Zuydam, N.R., Gaulton, K.J., Grarup, N., et al.; CHARGE Consortium Hematology Working Group (2016). Genome-wide associations for birth weight and correlations with adult disease. Nature *538*, 248–252.

42. Teumer, A., Tin, A., Sorice, R., Gorski, M., Yeo, N.C., Chu, A.Y., Li, M., Li, Y., Mijatovic, V., Ko, Y.A., et al.; DCCT/EDIC (2016). Genome-wide association studies identify genetic loci associated with albuminuria in diabetes. Diabetes *65*, 803–817.

43. Eppinga, R.N., Hagemeijer, Y., Burgess, S., Hinds, D.A., Stefansson, K., Gudbjartsson, D.F., van Veldhuisen, D.J., Munroe, P.B., Verweij, N., and van der Harst, P. (2016). Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nat. Genet. *48*, 1557–1563.

44. Köttgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., Pistis, G., Ruggiero, D., O'Seaghdha, C.M., Haller, T., et al.; LifeLines Cohort Study; CARDIoGRAM Consortium; DIAGRAM Consortium; ICBP Consortium; and MAGIC Consortium (2013). Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nat. Genet. *45*, 145–154.

45. Lu, Y., Day, F.R., Gustafsson, S., Buchkovich, M.L., Na, J., Bataille, V., Cousminer, D.L., Dastani, Z., Drong, A.W., Esko, T., et al. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. Nat. Commun. *7*, 10495.

46. Pattaro, C., Teumer, A., Gorski, M., Chu, A.Y., Li, M., Mijatovic, V., Garnaas, M., Tin, A., Sorice, R., Li, Y., et al.; ICBP Consortium; AGEN Consortium; CARDIOGRAM; CHARGe-Heart Failure Group; and ECHOGen Consortium (2016). Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. Nat. Commun. *7*, 10023.

47. Day, F.R., Ruth, K.S., Thompson, D.J., Lunetta, K.L., Pervjakova, N., Chasman, D.I., Stolk, L., Finucane, H.K., Sulem, P., Bulik-Sullivan, B., et al.; PRACTICAL consortium; kConFab Investigators; AOCS Investigators; Generation Scotland; EPIC-InterAct Consortium; and LifeLines Cohort Study (2015). Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. Nat. Genet. *47*, 1294–1303.

48. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.

49. Getz, G.S., and Reardon, C.A. (2009). Apoprotein E as a lipid transport and signaling protein in the blood, liver, and artery wall. J. Lipid Res. *50* (*Suppl*), S156–S161.

50. Pallaud, C., Gueguen, R., Sass, C., Grow, M., Cheng, S., Siest, G., and Visvikis, S. (2001). Genetic influences on lipid metabolism trait variability within the Stanislas Cohort. J. Lipid Res. *42*, 1879–1890.

51. Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. *41*, 1182–1190.

52. Mary, A. (2009). Silles. The causal effect of education on health: Evidence from the united kingdom. Econ. Educ. Rev. *28*, 122–128.

53. Baker, D.P., Leon, J., Smith Greenaway, E.G., Collins, J., and Movit, M. (2011). The education effect on population health: a reassessment. Popul. Dev. Rev. *37*, 307–332.

54. Mefford, J., and Witte, J.S. (2012). The covariate's dilemma. PLoS Genet. *8*, e1003096.

55. Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am. J. Hum. Genet. *96*, 329–339.

56. Loh, P.R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., Schizophrenia Working Group of Psychiatric Genomics Consortium, de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1292.