Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data

Huwenbo Shi,[1] Gleb Kichaev,[1] and Bogdan Pasaniuc[1,2,3,*]

Variance-component methods that estimate the aggregate contribution of large sets of variants to the heritability of complex traits have yielded important insights into the genetic architecture of common diseases. Here, we introduce methods that estimate the total trait variance explained by the typed variants at a single locus in the genome (local SNP heritability) from genome-wide association study (GWAS) summary data while accounting for linkage disequilibrium among variants. We applied our estimator to ultra-large-scale GWAS summary data of 30 common traits and diseases to gain insights into their local genetic architecture. First, we found that common SNPs have a high contribution to the heritability of all studied traits. Second, we identified traits for which the majority of the SNP heritability can be confined to a small percentage of the genome. Third, we identified GWAS risk loci where the entire locus explains significantly more variance in the trait than the GWAS reported variants. Finally, we identified loci that explain a significant amount of heritability across multiple traits.

Introduction

Large-scale genome-wide association studies (GWASs) have identified thousands of SNPs associated with hundreds of traits and diseases.[1–4] However, only a fraction of the trait variance can be explained by the risk SNPs reported by GWASs. The so-called "missing-heritability problem" is partly due to the fact that GWASs impose a stringent significance threshold, which neglects small-effect variants that fail to reach genome-wide significance at current sample sizes. As an alternative, analysis of variance components aggregates the effect of all SNPs regardless of their significance[5] and has yielded important insights into the genetic architecture of complex traits.[6–11]

Heritability has been traditionally estimated with twins or pedigree[12] information, and more recent works have shown that SNP-based heritability (i.e., the proportion of trait variance explained by a given set of SNPs) can be estimated from unrelated individuals.[8] Standard approaches for estimating SNP heritability rely on estimating the genetic relationships between pairs of individuals (estimated genome-wide or across a subset of the genome).[8,13,14] Therefore, these analyses require individual-level genotype data, which prohibits their applicability to ultra-large GWASs that, as a result of privacy concerns, are typically available only at the summary level. To solve this bottleneck, recent methods have shown that SNP heritability, both across the genome and for different functional categories in the genome, can be accurately estimated with only summary GWAS data.[6,7] Although these methods have enabled powerful analyses making insights into the genetic basis of complex traits, they rely on the infinitesimal-model assumption (i.e., that all SNPs contribute to the trait), which is invalid at most risk loci.[6,7] To overcome this drawback, alternative approaches have proposed imposing a prior on the sparsity of effect sizes to further increase accuracy of estimating SNP heritability.[15] A potentially more robust approach is to not assume any distribution for the effect sizes at causal variants and treat them as fixed effects in the estimation procedure. Indeed, recent works have shown that SNP heritability can be estimated under maximum likelihood from polygenic scores under a fixed-effects model assuming no linkage disequilibrium (LD) among SNPs.[11]

Here, we introduce Heritability Estimator from Summary Statistics (HESS), an approach for estimating the trait variance explained by all typed SNPs at a single locus in the genome while accounting for LD among SNPs. We build upon recent works[11,16] that have treated causal effect sizes as fixed effects and model the genotypes at the locus as random correlated variables. Our estimator can be viewed as a weighted summation of the squares of the projection of GWAS effect sizes onto the eigenvectors of the LD matrix at the considered locus, where the weights are inversely proportional to the corresponding eigenvalues. Through extensive simulations, we show that HESS is unbiased when in-sample LD is available, regardless of disease architecture (i.e., the number of causals and distribution of effect sizes). We extend our method to use LD estimated from reference panels[17] and show that a principal-component-based regularization of the LD matrix[18] yields approximately unbiased and more consistent estimates of local SNP heritability than existing methods.[6]

We applied HESS to partition common SNP heritability at each locus in the genome by using GWAS summary data for 30 traits spanning over 10 million SNPs and 2.4 million phenotype measurements. First, we show that common SNPs explain a large fraction (ranging anywhere

[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
*Correspondence: pasaniuc@ucla.edu
http://dx.doi.org/10.1016/j.ajhg.2016.05.013.

from 20% to 90% across the studied quantitative traits) of the total familial heritability estimated from twin studies. Second, we showcase the utility of estimates of local SNP heritability in finding loci that explain more trait variance than the top associated SNP at the locus—an effect likely due to multiple signals of association. Third, we contrast the polygenicity of all 30 traits by comparing the fraction of total SNP heritability attributable to loci with the highest local SNP heritability. We have found that most of the 30 selected traits are highly polygenic and that a small number of traits are driven by a small number of loci. Finally, we report 36 "heritability hotspots"—genomic regions that attain a significant contribution to the SNP heritability of multiple traits. Taken together, our results provide insights into traits where further GWASs and/or fine-mapping studies are likely to recover a significant amount of the missing heritability.

Material and Methods

Overview of Methods

We introduce estimators for the trait variance explained by typed variants at a single locus (local SNP heritability, $h^2_{g,\text{local}}$) from summary GWAS data (i.e., Z scores, effect sizes, and their SEs). We derive our estimator under the assumption that effect sizes at typed variants are fixed and genotypes are drawn from a distribution with a pre-specified covariance structure. The covariance (i.e., the pairwise correlation between any variants at a locus, LD) can be estimated in sample from the GWAS genotype data or from external reference panels (e.g., 1000 Genomes Project[17]). Our estimator can be viewed as a weighted summation of the squared projections of GWAS effect sizes onto the eigenvectors of the LD matrix at the considered locus. The finite sample size of a GWAS, as well as the reference panels used for estimating LD, induces statistical noise that needs to be accounted for if estimation is to be accurate. Given that the top projections make up the bulk of the summation, truncated singular value decomposition (SVD) lends itself as an appropriate regularization method to account for noise in the estimated LD matrices. Finally, we extend our approach to consider multiple independent loci each contributing to the trait and show how our local estimator can be employed when the total genome-wide SNP heritability is known (or estimated from other methods).

Estimating SNP Heritability at a Single Locus from GWAS Summary Data

Let $y_i = \mathbf{x_i}^T\boldsymbol{\beta} + \epsilon_i$, where y_i is the trait value for individual i, $\mathbf{x_i}$ are the standardized (i.e., mean 0 and unit variance) genotypes of individual i at p typed SNPs in the locus, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is the vector of fixed effect sizes for the p SNPs, and $\epsilon_i \sim N(0, \sigma_e^2)$ is the environmental effect. If we assume that $\boldsymbol{\beta}$ is fixed and $\mathbf{X}$ is random, the phenotypic variance is

$$\text{Var}[\mathbf{y}] = \text{Var}[\mathbf{X}\boldsymbol{\beta}] + \sigma_e^2 = \boldsymbol{\beta}^T\text{Cov}[\mathbf{X}]\boldsymbol{\beta} + \sigma_e^2 = \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta} + \sigma_e^2, \quad \text{(Equation 1)}$$

where $\mathbf{V}$ is a $p \times p$ variance-covariance matrix of the genotype vector (i.e., the LD matrix). If we make a standard assumption that the phenotypes are standardized (i.e., $\text{Var}[\mathbf{y}] = 1$), it follows that the amount of variance contributed by the p SNPs to the trait (i.e., local SNP heritability) is $h^2_{g,\text{local}} = \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}$. If the true effect-size vector $\boldsymbol{\beta}$ and the LD matrix $\mathbf{V}$ are given, then computing $h^2_{g,\text{local}}$ is trivial. In reality, however, the vector $\boldsymbol{\beta}$ is unknown and is estimated in a GWAS involving n samples and p SNPs, where $\widehat{\beta}_{\text{GWAS},i}$ is estimated as the marginal standardized regression coefficient for SNP i:

$$\widehat{\beta}_{\text{GWAS},i} = \frac{1}{n}\mathbf{X}_i^T\mathbf{y} = \frac{1}{n}\mathbf{X}_i^T\left(\begin{bmatrix}\mathbf{X}_1 & \cdots & \mathbf{X}_p\end{bmatrix}\boldsymbol{\beta} + \epsilon\right)$$
$$= \begin{bmatrix}\frac{1}{n}\mathbf{X}_i^T\mathbf{X}_1 & \cdots & \frac{1}{n}\mathbf{X}_i^T\mathbf{X}_p\end{bmatrix}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}_i^T\epsilon = \sum_{j=1}^{p} r_{ij}\beta_j + \frac{1}{n}\mathbf{X}_i^T\epsilon, \quad \text{(Equation 2)}$$

where $\mathbf{X}_i$ denotes standardized genotypes for SNP i across the n individuals, and r_{ij} denotes the LD between SNPs i and j. If we extend to p SNPs at the locus, it follows that $\widehat{\boldsymbol{\beta}}_{\text{GWAS}} = \mathbf{V}\boldsymbol{\beta} + (1/n)\mathbf{X}^T\epsilon$, where $\mathbf{V}$ is the LD matrix. When $\boldsymbol{\beta}$ is fixed and $\boldsymbol{\epsilon}$ is random, $\widehat{\boldsymbol{\beta}}_{\text{GWAS}}$ is a random variable with $\text{E}[\widehat{\boldsymbol{\beta}}_{\text{GWAS}}] = \text{E}[\mathbf{V}\boldsymbol{\beta} + (1/n)\mathbf{X}^T\epsilon] = \mathbf{V}\boldsymbol{\beta}$ and $\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{GWAS}}] = \text{Var}[\mathbf{V}\boldsymbol{\beta} + (1/n)\mathbf{X}^T\epsilon] = (1/n^2)\mathbf{X}^T\text{Cov}[\epsilon]\mathbf{X} = (\sigma_e^2/n)\mathbf{V} = \mathbf{V}(1 - h^2_{g,\text{local}})/n$. By central limit theorem, $\widehat{\boldsymbol{\beta}}_{\text{GWAS}} \sim N(\mathbf{V}\boldsymbol{\beta}, \mathbf{V}(1 - h^2_{g,\text{local}})/n)$.

As the GWAS sample size (n) increases, $\widehat{\boldsymbol{\beta}}_{\text{GWAS}}$ converges to $\boldsymbol{\beta}_{\text{GWAS}} = \mathbf{V}\boldsymbol{\beta}$. By simple substitution in Equation 1, it follows that an estimator for $h^2_{g,\text{local}}$ is

$$(\boldsymbol{\beta}^T_{\text{GWAS}}\mathbf{V}^{-1})\mathbf{V}(\mathbf{V}^{-1}\boldsymbol{\beta}_{\text{GWAS}}) = \boldsymbol{\beta}^T_{\text{GWAS}}\mathbf{V}^{-1}\boldsymbol{\beta}_{\text{GWAS}}. \quad \text{(Equation 3)}$$

Unfortunately, the finite sample size of the GWAS induces statistical noise in the estimation of $\boldsymbol{\beta}_{\text{GWAS}}$, which leads to biased estimation if we simply replace $\boldsymbol{\beta}_{\text{GWAS}}$ with $\widehat{\boldsymbol{\beta}}_{\text{GWAS}}$ above as $\text{E}[\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{-1}\widehat{\boldsymbol{\beta}}_{\text{GWAS}}] = \text{tr}(\mathbf{V}^{-1}\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{GWAS}}]) + \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}$. However, we can correct for the bias $\text{tr}(\mathbf{V}^{-1}\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{GWAS}}])$ as follows.

Let $\widehat{h}^2_{g,\text{local}}$ be an unbiased estimator of $h^2_{g,\text{local}}$; then by definition, $E[\widehat{h}^2_{g,\text{local}}]$ must satisfy $E[\widehat{h}^2_{g,\text{local}}] = h^2_{g,\text{local}}$. Then it follows that

$$\text{E}\left[\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{-1}\widehat{\boldsymbol{\beta}}_{\text{GWAS}}\right] = \text{tr}\left(\frac{1 - h^2_{g,\text{local}}}{n}\mathbf{V}^{-1}\mathbf{V}\right) + h^2_{g,\text{local}}$$
$$= \frac{1 - \text{E}\left[\widehat{h}^2_{g,\text{local}}\right]}{n}p + \text{E}\left[\widehat{h}^2_{g,\text{local}}\right]. \quad \text{(Equation 4)}$$

A sufficient condition for Equation 4 to hold is $(1 - \widehat{h}^2_{g,\text{local}})p/n + \widehat{h}^2_{g,\text{local}} = \widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{-1}\widehat{\boldsymbol{\beta}}_{\text{GWAS}}$. Solving for $\widehat{h}^2_{g,\text{local}}$ gives an unbiased estimator for $h^2_{g,\text{local}}$:

$$\widehat{h}^2_{g,\text{local}} = \frac{n\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{-1}\widehat{\boldsymbol{\beta}}_{\text{GWAS}} - p}{n - p}. \quad \text{(Equation 5)}$$

Following quadratic form theory,[19] the variance of $\widehat{h}^2_{g,\text{local}}$ is

$$\text{Var}\left[\widehat{h}^2_{g,\text{local}}\right] = \left(\frac{n}{n-p}\right)^2\left(2p\left(\frac{1 - h^2_{g,\text{local}}}{n}\right) + 4h^2_{g,\text{local}}\right)\left(\frac{1 - h^2_{g,\text{local}}}{n}\right). \quad \text{(Equation 6)}$$

Because $h^2_{g,\text{local}}$, the true local SNP heritability, is unknown, we use $\widehat{h}^2_{g,\text{local}}$ instead. For $h^2_{g,\text{local}}$ near 0, $\text{Var}[\widehat{h}^2_{g,\text{local}}] \approx (4/(n-p)^2)h^2_{g,\text{local}} + (2p/(n-p)^2)$ through Taylor expansion around 0. Thus, the plug-in principle yields an estimation of $\text{Var}[\widehat{h}^2_{g,\text{local}}]$ approximately equal to the truth in the expectation. For small $\widehat{h}^2_{g,\text{local}}$ (as expected for most loci and traits), $\text{Var}[\widehat{h}^2_{g,\text{local}}]$ is dominated by $2p/(n-p)^2$.

Accounting for Rank Deficiencies in LD

In the above derivation, we made the assumption that the inverse of the LD matrix $\mathbf{V}$ exists. In practice, however, as a result of pairs of SNPs in perfect LD, $\mathbf{V}$ is usually rank deficient, and thus $\mathbf{V}^{-1}$ does not exist. In such cases, we use the Moore-Penrose pseudoinverse[20] $\mathbf{V}^{\dagger}$. Let $q = \text{rank}(\mathbf{V})$; by rank decomposition, $\mathbf{V} = \mathbf{V}_A\mathbf{V}_B$, where $\mathbf{V}_A \in \mathbf{R}^{p\times q}$ and $\mathbf{V}_B \in \mathbf{R}^{q\times p}$ are matrices with full column rank and full row rank, respectively. Then, $\text{tr}(\mathbf{V}^{\dagger}\mathbf{V}) = \text{tr}(\mathbf{V}_B^{\dagger}\mathbf{V}_A^{\dagger}\mathbf{V}_A\mathbf{V}_B) = \text{tr}(\mathbf{V}_B\mathbf{V}_B^{\dagger}\mathbf{V}_A^{\dagger}\mathbf{V}_A) = \text{tr}(\mathbf{I}_q) = q$. Accounting for the rank-deficient LD matrix, we obtain an unbiased estimator, $\widehat{h}^2_{g,\text{local}} = (n\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{\dagger}\widehat{\boldsymbol{\beta}}_{\text{GWAS}} - q)/(n-q)$. We make the same adjustment (replacing p with q) in the variance estimator for $\widehat{h}^2_{g,\text{local}}$.

Adjusting for Noise in External Reference LD

When genotype data of GWAS samples are not available, we substitute the in-sample LD matrix $\mathbf{V}$ with external reference LD matrix $\widehat{\mathbf{V}}$ estimated from the 1000 Genomes Project[17] with a population that matches the GWAS samples. However, because of limited sample size, external reference LD matrices contain statistical noise that biases our estimate. We apply truncated-SVD regularization to remove noise from the external reference LD matrix as follows.

First, note that $\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{\dagger}\widehat{\boldsymbol{\beta}}_{\text{GWAS}} = \sum_{i=1}^{q} s_i = \sum_{i=1}^{q}(1/w_i)(\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{u}_i)^2$, where w_i and $\mathbf{u}_i$ are the eigenvalues and eigenvectors, respectively, of the LD matrix $\mathbf{V}$, and $q = \text{rank}(\mathbf{V})$. For external reference LD matrix $\widehat{\mathbf{V}}$ with eigenvalues and eigenvectors $\widehat{w}_i$ and $\widehat{\mathbf{u}}_i$, respectively, the same decomposition holds except that s_i is replaced with $\widehat{s}_i = (1/\widehat{w}_i)(\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\widehat{\mathbf{u}}_i)^2$. In our previous works,[21,22] we proposed regularizing $\widehat{\mathbf{V}}$ by using ridge-regression penalty. This regularization method is equivalent to replacing $\widehat{w}_i$ with $\widehat{w}_i + \lambda$, where λ is the ridge-regression penalty. The ridge-regression penalty shrinks the quadratic term $\widehat{\boldsymbol{\beta}}_{\text{GWAS}}\widehat{\mathbf{V}}^{\dagger}\widehat{\boldsymbol{\beta}}_{\text{GWAS}}$ toward 0, which can lead to downward bias. We also notice that a large λ is needed to drive down the noise ($\widehat{s}_i$ for large i), which diminishes the true signal at the same time. Here, we show through simulations that most of the signal in $\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{\dagger}\widehat{\boldsymbol{\beta}}_{\text{GWAS}}$ comes from s_i, where $i \ll q$, and that $\widehat{s}_i \approx s_i$ for $i \ll q$ (see Figure S1). These results motivate us to apply truncated SVD to remove noise in $\widehat{\mathbf{V}}$, i.e., we estimate $\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\mathbf{V}^{\dagger}\widehat{\boldsymbol{\beta}}_{\text{GWAS}}$ by $\sum_{i=1}^{k} 1/\widehat{w}_i(\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\widehat{\mathbf{u}}_i)^2$, where $k \ll q$. Let $g(\widehat{\boldsymbol{\beta}}_{\text{GWAS}}, k) = \sum_{i=1}^{k}(1/\widehat{w}_i)(\widehat{\boldsymbol{\beta}}^T_{\text{GWAS}}\widehat{\mathbf{u}}_i)^2$; through eigen decomposition of $\widehat{\mathbf{V}}$, it can be shown that

$$\text{E}\left[g\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS}}, k\right)\right] = \frac{k\left(1 - h^2_{g,\text{local}}\right)}{n} + \sum_{i=1}^{k}\widehat{w}_i\left(\widehat{\mathbf{u}}_i^T\boldsymbol{\beta}\right)^2. \quad \text{(Equation 7)}$$

Because the true local SNP heritability is $h^2_{g,\text{local}} = \sum_{i=1}^{q} w_i(\mathbf{u}_i^T\boldsymbol{\beta})^2$, if we assume $\widehat{\mathbf{u}}_i = \mathbf{u}_i$ for $i \ll q$, Equation 7 is an approximation of $h^2_{g,\text{local}}$ with bias $k(1 - h^2_{g,\text{local}})/n$. Correcting for this bias yields the estimator for the single-locus case:

$$\tilde{h}^2_{g,\text{local}} = \frac{ng\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS}}, k\right) - k}{n-k}. \quad \text{(Equation 8)}$$

In theory, the variance of $\tilde{h}^2_{g,\text{local}}$ is $\text{Var}[\tilde{h}^2_{g,\text{local}}] \approx (4/(n-k)^2)\widehat{h}^2_{g,\text{local}} + (2k/(n-k)^2)$. In practice, however, this gives an underestimation of the truth. Thus, we replace k with $q = \text{rank}(\mathbf{V})$.

Extension to Multiple Independent Loci

For genomes partitioned into m independent loci, the linear model for individual i's trait value becomes $y_i = \mathbf{x}^T_{i,1}\boldsymbol{\beta}_1 + \ldots + \mathbf{x}^T_{i,m}\boldsymbol{\beta}_m + \epsilon_i$, where $\mathbf{x}_{i,j}$ denotes the genotypes at the p_i SNPs in the i^{th} locus for individual i, and $\boldsymbol{\beta}_i$ denotes the effect sizes of SNPs in this locus. On the basis of the revised model, we decompose $\text{Var}[\mathbf{y}]$ into

$$\begin{aligned}\text{Var}[\mathbf{y}] &= \text{Var}[\mathbf{X}_1\boldsymbol{\beta}_1] + \cdots + \text{Var}[\mathbf{X}_m\boldsymbol{\beta}_m] + \sigma^2_e \\ &= h^2_{g,\text{local},1} + \cdots + h^2_{g,\text{local},m} + \sigma^2_e, \end{aligned} \quad \text{(Equation 9)}$$

where $h^2_{g,\text{local},i}$ denotes the local SNP heritability contributed by the i^{th} locus. In the case of multiple independent loci, the noise term σ^2_e is equal to $1 - \sum_{j=1}^{m} h^2_{g,\text{local},j}$. Thus, in order to correct for the bias generated by σ^2_e, we need to know $h^2_{g,\text{local},j}$ for all j. After accounting for bias and adjusting for noise in external reference LD ($\widehat{\mathbf{V}}_i$) according to the strategies outlined in previous sections, we arrive at the estimator

$$\widehat{h}^2_{g,\text{local},i} = \frac{ng\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS},i}, k_i\right) - \left(1 - \sum_{j=1, j\neq i}^{m}\widehat{h}^2_{g,\text{local},j}\right)k_i}{n - k_i}, \quad \text{(Equation 10)}$$

which defines a system of linear equations involving m variables ($\widehat{h}^2_{g,\text{local},i}$) and m equations. We can solve a similar system of linear equations to obtain the variance estimate

$$\begin{aligned}\text{Var}\left[\widehat{h}^2_{g,\text{local},i}\right] = &\left(\frac{n}{n-k_i}\right)^2\left(2k_i\frac{\widehat{\sigma}^2_e}{n} + 4\widehat{h}^2_{g,\text{local},i}\right)\frac{\widehat{\sigma}^2_e}{n} \\ &+ \left(\frac{k_i}{n-k_i}\right)^2\sum_{j=1, j\neq i}^{m}\text{Var}\left[\widehat{h}^2_{g,\text{local},j}\right], \end{aligned} \quad \text{(Equation 11)}$$

where $\widehat{\sigma}^2_e = 1 - \sum_{j=1}^{m}\widehat{h}^2_{g,\text{local},j}$.

In the special case when $k_1 = \cdots = k_m = k$ (i.e., all loci use the same number of eigenvectors in the truncated-SVD regularization of LD matrices), Equation 10 simplifies as follows: $\widehat{h}^2_g = \sum_{i=1}^{m}\widehat{h}^2_{g,\text{local},i} = \sum_{i=1}^{m}(ng(\widehat{\boldsymbol{\beta}}_{\text{GWAS},i}, k) - (1 - \widehat{h}^2_g + \widehat{h}^2_{g,\text{local},i})k)/(n-k) = (n/(n-k))\sum_{i=1}^{m} g(\widehat{\boldsymbol{\beta}}_{\text{GWAS},i}, k) - (k/(n-k))(m - m\widehat{h}^2_g + \widehat{h}^2_g)$. This yields the following estimate for the total genome-wide SNP heritability:

$$\widehat{h}^2_g = \frac{n}{n-mk}\sum_{i=1}^{m} g\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS},i}, k\right) - \frac{mk}{n-mk}, \quad \text{(Equation 12)}$$

which has variance

$$\text{Var}\left[\widehat{h}^2_g\right] = \left(\frac{n}{n-mk}\right)^2\sum_{i=1}^{m}\text{Var}\left[g\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS},i}, k\right)\right] \approx \left(\frac{n}{n-mk}\right)^2\frac{2mk}{(n-k)^2}. \quad \text{(Equation 13)}$$

Thus, if k is chosen such that $n - mk$ is small (i.e., $n/(n-mk)$ is large), the estimates of genome-wide SNP heritability become unstable with large variance. To ensure stable estimates and reduce variance (at the cost of some bias), we recommend choosing k such that $n/(n-mk)$ is less than 2 when using our estimator for genome-wide estimation.

Known Genome-wide SNP Heritability

In many cases, the estimate of total genome-wide SNP heritability (h^2_g) and its variance ($\text{Var}[h^2_g]$) are known (e.g., estimated from individual-level data). In those cases, one can simply plug h^2_g into Equation 10 to obtain local estimates of heritability $h^2_{g,\text{local},i}$:

$$\widehat{h}^2_{g,\text{local},i} = g\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS},i}, k\right) - \frac{k}{n}\left(1 - h^2_g\right), \quad \text{(Equation 14)}$$

from which we conclude

$$\text{Var}\left[\widehat{h}^2_{g,\text{local},i}\right] = \text{Var}\left[g\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS},i}, k\right)\right] + \left(\frac{k}{n}\right)^2\text{Var}\left[h^2_g\right]. \quad \text{(Equation 15)}$$

Table 1. Estimates of Total SNP Heritability and the Amount of h_g^2 Attributable to Loci Containing GWAS Index SNPs, $h^2_{g,\text{local,GWAS}}$, and Index SNPs Only, h^2_{GWAS}

Trait	h_g^2 (% SE)	h^2_{pub} (%)	h_g^2/h^2_{pub}	h^2_{GWAS} (% SE)	$h^2_{g,\text{local,GWAS}}$ (% SE)	$h^{2*}_{g,\text{local,GWAS}}$ (% SE)	Enrichment[a] (SE)
BMI[1]	16.5 (0.5)	42[23]	0.39	1.6 (0.001)	3.1 (0.1)	3.1 (0.1)	3.7 (0.4)
Height[2]	59.4 (0.3)	69[23]	0.86	13.9 (0.002)	32.0 (0.2)	24.0 (0.2)	1.5 (0.1)
Hemoglobin[24]	17.9 (2.1)	37[25]	0.48	2.2 (0.003)	1.9 (0.3)	1.8 (0.3)	7.6 (1.4)
Mean cell hemoglobin[24]	29.3 (2.2)	52[26]	0.56	7.2 (0.003)	6.2 (0.4)	6.1 (0.4)	9.9 (1.9)
Concentration of mean cell hemoglobin[24]	10.9 (2.5)	48[27]	0.23	0.4 (0.003)	0.5 (0.2)	0.5 (0.2)	6.7 (1.8)
Mean cell volume[24]	26.3 (2.0)	52[26]	0.51	6.5 (0.004)	5.7 (0.4)	5.6 (0.4)	8.1 (1.3)
Packed cell volume[24]	16.7 (2.5)	30[25]	0.56	1.4 (0.003)	0.9 (0.2)	0.8 (0.2)	6.0 (1.4)
Red blood cell count[24]	22.0 (2.3)	56[26]	0.39	3.6 (0.004)	2.6 (0.3)	2.6 (0.3)	6.4 (1.6)
Number of platelets[28]	27.5 (1.5)	57[25]	0.48	3.5 (0.003)	3.9 (0.3)	3.9 (0.3)	5.7 (0.9)
Fasting glucose[29]	22.3 (2.3)	66[30]	0.34	2.6 (0.002)	1.7 (0.2)	1.6 (0.2)	8.0 (2.5)
Fasting insulin[29]	19.9 (2.4)	36[31]	0.55	–	–	–	–
HBA1C[32]	20.8 (2.3)	75[30]	0.28	1.8 (0.003)	0.9 (0.2)	0.9 (0.2)	6.6 (1.9)
HOMA-B[29]	20.3 (2.4)	72[33]	0.28	0.6 (0.001)	0.4 (0.1)	0.4 (0.1)	7.5 (1.9)
HOMA-IR[29]	19.9 (2.4)	38[31]	0.52	–	–	–	–
HDL[3]	39.4 (0.9)	42[34]	0.94	5.8 (0.002)	10.7 (0.2)	10.5 (0.2)	4.6 (1.3)
LDL[3]	33.0 (1.0)	40[34]	0.82	7.8 (0.002)	8.4 (0.2)	8.3 (0.2)	5.1 (0.9)
TC[3]	35.5 (0.9)	50[35]	0.71	8.0 (0.002)	9.3 (0.2)	9.3 (0.2)	4.3 (0.6)
TG[3]	34.8 (0.9)	40[36]	0.87	5.2 (0.002)	8.0 (0.2)	8.0 (0.2)	5.8 (1.4)
Education years[37]	19.9 (0.8)	40[37]	0.50	0.1 (0.002)	0.2 (0.0)	0.2 (0.0)	3.2 (1.4)
Forearm BMD[38]	17.4 (2.2)	84[39]	0.21	0.3 (0.001)	0.5 (0.1)	0.5 (0.1)	22.4 (7.7)
Femoral-neck BMD[38]	24.1 (2.1)	84[39]	0.29	2.0 (0.003)	2.0 (0.2)	2.0 (0.2)	7.1 (1.0)
Lumbar spine[38]	25.1 (2.0)	84[39]	0.30	2.2 (0.003)	2.2 (0.3)	2.2 (0.3)	6.1 (0.8)
Age at menarche[40]	27.8 (0.7)	49[41]	0.57	2.6 (0.002)	3.8 (0.2)	3.7 (0.2)	2.9 (0.2)
College[37]	19.4 (0.8)	40[37]	0.48	0.1 (0.001)	0.1 (0.0)	0.1 (0.0)	3.5 (0.9)
RA[42]	66.3 (0.9)	55[43]	1.21	11.2 (0.003)	22.0 (0.3)	22.1 (0.3)	9.8 (4.3)
SCZ[44]	64.5 (0.7)	81[45]	0.80	6.2 (0.004)	9.2 (0.2)	9.2 (0.2)	2.3 (0.1)
Crohn disease[46]	35.9 (1.8)	53[47]	0.68	3.8 (0.002)	5.9 (0.4)	5.9 (0.4)	4.8 (0.7)
Inflammatory bowel disease[46,b]	35.3 (1.4)	–	–	4.9 (0.002)	6.7 (0.3)	6.6 (0.3)	4.6 (0.5)
Ulcerative colitis[46]	31.9 (2.1)	58[47]	0.55	2.7 (0.002)	4.1 (0.3)	4.1 (0.3)	5.4 (1.0)
Type 2 diabetes[48]	25.4 (1.6)	26[49]	0.98	1.3 (0.002)	1.1 (0.2)	1.1 (0.2)	3.9 (0.7)

$h^{2*}_{g,\text{local,GWAS}}$ is the same as $h^2_{g,\text{local,GWAS}}$ except that GWAS index SNPs were excluded from the computation. In Table S2, we report $h^{2\dagger}_{g,\text{local,GWAS}}$, which we obtained by excluding all GWAS hits. We also report estimates of familial heritability (h^2_{pub}) obtained from twin or family studies. At the bottom of the table, we list case-control traits where our estimate of h_g^2 is biased as a result of ascertainment.

[a]Similar to Finucane et al.,[7] we define enrichment as the ratio between the fraction of h_g^2 attributable to $h^{2*}_{g,\text{local,GWAS}}$ and the genomic fraction covered by these loci. We obtained SEs by a jackknife over the loci.

[b]Inflammatory bowel disease refers to the union of Crohn disease and ulcerative colitis.

In general, the sum of local SNP heritability $\widehat{h}_g^2 = \sum_{i=1}^m \widehat{h}^2_{g,\text{local},i}$ is not necessarily equal to h_g^2 as a result of variance in $\widehat{h}^2_{g,\text{local},i}$. Given that $\text{Var}[\widehat{h}_g^2] = \text{Var}[\sum_{i=1}^m \widehat{h}^2_{g,\text{local},i}] \approx (2mk/(n-k)^2) + (mk/n)^2 \text{Var}[h_g^2]$, we recommend choosing k such that mk/n is less than 0.5 to ensure stable estimate and reduce variance. We assessed estimation of the local SNP heritability with or without known genome-wide SNP heritability by using the height GWAS data (see Table 1) with a previously reported $h_g^2 = 0.50$.[2] The estimates of local SNP heritability were virtually indistinguishable between the two approaches ($R = 1.0$; see Figure S8).

Simulation Framework

We used HAPGEN2[50] to simulate genotypes for 50,000 individuals by starting with half of the 505 European (EUR) individuals in the 1000 Genomes Project[17] for SNPs with minor allele frequency

(MAF) greater than 5% in randomly selected regions spanning 0.75–1.5 Mb on chromosome 1. We reserved the other half of the EUR individuals as an external reference panel. From the simulated genotypes of the 50,000 individuals, we then simulated phenotypes according to the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\mathbf{X}$ is the standardized genotype matrix with mean 0 and variance 1 at each column.

We investigated the performance of our method under a wide range of simulations. We first selected a subset C of $|C|$ causal SNPs at random and then simulated the effect sizes at these SNPs as $\boldsymbol{\beta}_C \sim N(0, (h^2/|C|)\mathbf{I}_{|C|})$, where h^2 is the heritability to be simulated. We drew ϵ from $N(0, (1-h^2)\mathbf{I}_n)$ such that $\mathrm{E}[\mathbf{y}] = 0$, $\mathrm{Var}[\mathbf{y}] = 1$, and the SNP heritability for this locus is h^2. For fixed $\boldsymbol{\beta}$, we then generated replications of trait values $\mathbf{y}$ by re-drawing ϵ. Finally, we computed summary statistics, $\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}}$, according to the procedures outlined in previous sections. We simulated 500 sets of summary statistics for each simulation scenario. Although C and $\boldsymbol{\beta}$ were fixed within each of the 500 sets of simulated summary statistics, they varied across different set of simulations.

We also investigated simulations where $\boldsymbol{\beta}$ varied across simulated individuals. In each of the 500 sets of simulated GWAS summary statistics, we first selected a subset C of $|C|$ causal SNPs at random. Then, for each individual, we drew $\boldsymbol{\beta}_{C,i}$ from $N(0, \boldsymbol{\alpha}_i h^2)$ for $i = 1, \cdots, |C|$, where $\boldsymbol{\alpha}$ governs the proportion of heritability contributed by each SNP and satisfies $\sum_{i=1}^{|C|} \boldsymbol{\alpha}_i = 1$. In the special case when $\boldsymbol{\alpha}_i = 1/|C|$ for all i, each causal SNP contributes the same proportion of heritability. Here, C and $\boldsymbol{\alpha}$ were fixed in each simulation set but varied across the 500 sets of simulations.

Given that in simulations we assume that all SNPs are typed and that environmental effect (ϵ) is drawn independently for each individual, cryptic relatedness among individuals in the 1000 Genomes Project[17] will have minimal effect on our estimates.

Empirical Datasets

We obtained publicly available GWAS summary data for 30 traits in individuals of European ancestry from 11 GWAS consortia (see Table 1). For quality control, we restricted our analysis to GWASs involving at least 20,000 samples and excluded sex chromosomes. We used the same definition of independent loci as in Berisa and Pickrell[51] (1.6 Mb on average). To reduce statistical noise in the LD matrix, we focused on estimating heritability attributable to common SNPs (i.e., SNPs with a MAF greater than 5% in the EUR 1000 Genomes data[17]). Prior to estimating heritability, we also removed SNPs with ambiguous alleles in comparison to the reference panel (Table S1) and applied our estimator as defined in Equation 10. For each trait, we chose k, the number of eigenvectors used for estimating local heritability across all loci, on the basis of the GWAS sample size (see Material and Methods)—a large k for a GWAS with a large sample size and a small k for a GWAS with a small sample size. To avoid inflation due to noise in LD, we capped k at a maximum of 50 (see Table S2). To ensure stable estimates, we also recommend filtering out eigenvectors with corresponding eigenvalues less than 1.

Most GWASs apply a genomic control (GC) factor (λ_{GC}) to χ^2 statistics to correct for inflation due to population structure[52] and publish GC-corrected effect-size estimation ($\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS,GC}}$). We note that all of the summary GWAS data we analyze in this work are adjusted for population structure to various degrees and have at least one round of genomic correction. However, recent works[6,53] have shown that λ_{GC} cannot distinguish between inflation and true polygenicity and overestimates the correction factor needed for population stratification. Although dividing the χ^2 statistics by λ_{GC} has little effect on computing the ratios between local and genome-wide heritability,[7] it can result in underestimation of both local and genome-wide SNP heritability—when applied on GC-corrected summary data directly, our method can produce negative and uninformative estimates of local and total SNP heritability. To account for this, we first estimate λ_{GC} from summary GWAS data and re-inflate the effect sizes ($\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS,GC}}$) with estimated $\sqrt{\lambda_{\mathrm{GC}}}$ before obtaining estimates of local SNP heritability. We estimate λ_{GC} on the basis of the observation that at a locus with no heritability (i.e., $h^2_{g,\mathrm{local},i} = 0$), $\mathrm{E}[\widehat{\boldsymbol{\beta}}^T_{\mathrm{GWAS,GC},i}\mathbf{V}_i^{\dagger}\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS,GC},i}] = (1/\lambda_{\mathrm{GC}})(q_i/n)$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS,GC},i} = \widehat{\boldsymbol{\beta}}_{\mathrm{GWAS},i}/\sqrt{\lambda_{\mathrm{GC}}}$ denotes the GC-corrected effect-size vector, and $\mathrm{E}[\widehat{\boldsymbol{\beta}}^T_{\mathrm{GWAS},i}\mathbf{V}_i^{\dagger}\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS},i}] = q_i/n$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS},i}$ is the vector of effect-size estimation without GC correction. To estimate λ_{GC}, we treat the bottom 50% of all loci with the smallest estimated local SNP heritability as loci for which $h^2_{g,\mathrm{local},i} = 0$ and regress the vector (q_i/n) against the vector $(\widehat{\boldsymbol{\beta}}^T_{\mathrm{GWAS,GC},i}\mathbf{V}_i^{\dagger}\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS,GC},i})$. We note that using the bottom 50% of all loci is a conservative measure to account for ascertainment in choosing loci and can result in estimated λ_{GC} less than 1. In practice, we only re-inflate $\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS,GC}}$ if the estimated λ_{GC} is greater than 1. We report estimated λ_{GC} for all 30 traits in Table S1. Overall, our estimated λ_{GC} is consistent with the reported λ_{GC}. For example, our estimated λ_{GC} for BMI (1.33), high-density lipoprotein (HDL; 1.13), low-density lipoprotein (LDL; 1.16), total cholesterol (TC; 1.16), and triglycerides (TG; 1.18) are consistent with the reported λ_{GC} for BMI (1.38)[1] and lipid traits (1.10–1.15).[3]

We define GWAS hits as SNPs with p values less than 5×10^{-8}. To avoid overestimation due to LD tagging, for each locus, we only select the most significant (i.e., smallest p value) GWAS hit as the index SNP. Heritability attributable to index SNPs, h^2_{GWAS}, is then estimated as $\sum_{i=1}^{I} \widehat{\beta}_i^2$, where $\widehat{\beta}_i$ is the effect size of the i^{th} index SNP, and I is the number of index SNPs. We estimate the variance of $\widehat{h}^2_{\mathrm{GWAS}}$ as $\mathrm{Var}[\widehat{h}^2_{\mathrm{GWAS}}] = \sum_{i=1}^{I}\mathrm{Var}[\widehat{\beta}_i^2] = \sum_{i=1}^{I}\mathrm{Var}[(Z_i/\sqrt{n})^2] = \sum_{i=1}^{I}\mathrm{Var}[(1/n)\chi_i^2] = 2I/n^2$.

For case-control traits, an adjustment factor is needed to correct for ascertainment.[54] We note that this adjustment factor is derived on the basis of the infinitesimal model and does not apply to our method, which assumes a fixed-effects model. Therefore, we report only unadjusted heritability estimates for case-control traits. However, we note that the ratio between local and genome-wide SNP heritability is not affected by this scaling factor.

Results

Performance of HESS in Simulations

We used simulations to assess the performance of our proposed approach under a variety of disease architectures. First, we confirmed that by accounting for rank deficiency in the LD matrix, we obtained unbiased estimation, whereas the approach that uses the number of SNPs to correct for bias generated by the quadratic form[16] leads to a severe underestimation of heritability (Figure S2). Second, we found that using the top 10–50 eigenvectors of the LD matrix (see Material and Methods) provides a good

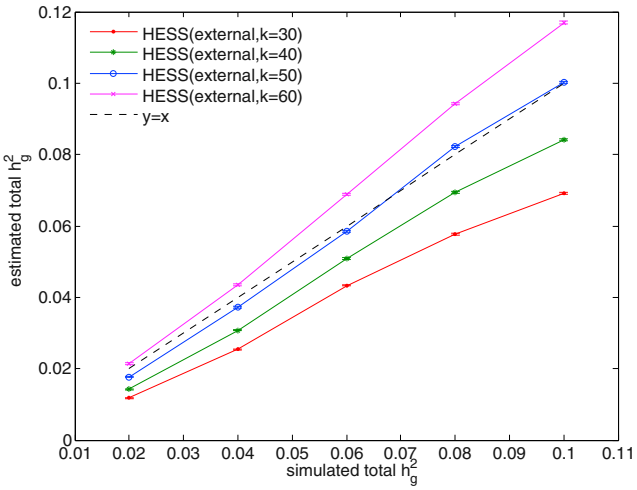

Figure 1. Estimates of Total SNP Heritability in the Whole-Chromosome Simulation for Different Numbers of Eigenvectors Included

We saw a slight downward bias when the number of eigenvectors, k, was small (e.g., $k = 30$) and an upward bias when k was large (e.g., $k = 60$). When $k = 50$, we attained an approximately unbiased estimate of total SNP heritability. Error bars represent twice the SE.

approximation of the estimated heritability when LD is estimated from external reference panels (Figure S1).

Because we used approximately independent loci,[51] we also assessed potential bias due to cross-tagging of heritability resulting from LD across adjacent loci. We simulated summary statistics based on 10,000 randomly selected SNPs spread across the entire chromosome 22; 20% of these SNPs were causal, and total SNP heritability varied from 2% to 10%. For each simulation scenario, we simulated 500 sets of summary statistics and used Equation 10 to estimate local SNP heritability. We estimated total SNP heritability by summing all local estimates of SNP heritability. We found that when HESS used the top $k = 30$ eigenvectors in the truncated-SVD regularization of LD matrices, it yielded a downwardly biased estimate of total SNP heritability, whereas at $k = 50$, HESS was approximately unbiased (Figure 1). Therefore, we used $k = 50$ as the default unless otherwise noted.

Next, we compared HESS to the recently proposed LD Score Regression (LDSC),[6,7] which provides estimates of heritability from GWAS summary data. Although LDSC is not designed for local analyses as a result of model assumptions on polygenicity, it is able to estimate the trait variance attributable to any sets of SNPs. As expected, in our simulations, where all individuals shared the same effect-size vector ($\boldsymbol{\beta}$), we found that LDSC was sensitive to the underlying polygenicity and, in general, yielded biased estimation of heritability. In contrast, HESS provided an unbiased estimation of heritability across all simulated disease architectures when in-sample LD was available. For example, in simulations where 20% of the variants at the locus were causal and explained 0.05% of the heritability, HESS yielded an estimate of 0.054% (SE = 0.004%), whereas LDSC yielded 0.025% (SE = 0.0009%) (Figure 2). We attribute this to the fact that HESS does not make an assumption on the distribution of effect sizes at causal variants by treating them as fixed effects in the model. When LD from the sample was unavailable and had to be estimated from reference panels, both methods were biased, but HESS (with $k = 30$ or 50 eigenvectors in the truncated-SVD regularization of the LD matrix) yielded results closer to the simulated heritability at randomly selected loci with different widths (Figure 2, Figure S3, and Figure S4). Similar results were obtained in simulations where $\boldsymbol{\beta}$ was drawn independently for each individual (see Figure S5). This is expected because when conditioned on a fixed $\boldsymbol{\beta}$, HESS is unbiased (i.e., $\mathrm{E}[\widehat{h_g^2} \mid \boldsymbol{\beta}] = h_g^2$), and then the expectation of the HESS estimate across all possible $\boldsymbol{\beta}$ is still unbiased (i.e., $\mathrm{E}[\widehat{h_g^2}] = \mathrm{E}[\mathrm{E}[\widehat{h_g^2} \mid \boldsymbol{\beta}]] = \mathrm{E}[h_g^2] = h_g^2$).

Finally, unlike LDSC, which employs a jackknife approach to estimate variance in the estimated heritability (thus requiring multiple loci), HESS provides a variance estimator according to quadratic form theory (see Material and Methods). Because external reference LD is typically computed on the basis of much smaller samples than in-sample LD, subtle patterns in in-sample LD cannot be captured by external reference LD. Thus, external reference LD matrices usually have lower rank than their corresponding in-sample LD matrices, resulting in underestimation of $\mathrm{Var}[\widehat{h}^2_{g,\mathrm{local},i}]$ (see Equation 11). We verified this in simulations and found that the variance estimator yielded unbiased estimates when in-sample LD was available and underestimated theoretical variance when external reference LD was used (Figure S6). We also note that cryptic relatedness in GWAS samples can drive down the effective sample size (n); thus, our estimates of SEs could be deflated for GWASs in which the effective sample size is significantly smaller than the actual sample size.

Common Variants Explain a Large Fraction of Heritability

Having demonstrated the utility of HESS in simulations, we next applied our method to empirical GWAS summary data across 30 complex traits and diseases spanning more than two million phenotypic measurements (see Material and Methods, Table 1, and Table S1). We estimated the local SNP heritability at 1,703 approximately independent loci[51] by using EUR individuals from 1000 Genomes to estimate LD.[17] We first investigated the total contribution of common variants (MAF > 5%) to the heritability of complex traits. We summed up the local estimates provided by our method to obtain an estimate for the total genome-wide heritability for all genotyped SNPs. For traits where the SNP heritability was previously reported, we found a broad consistency between our estimate and the existing estimates from the literature (see Table 1). For example, HESS estimated a genome-wide SNP heritability (h_g^2) of 16.5% (SE = 0.5%) for BMI and 59.4% (SE = 0.3%) for height, whereas the previously reported estimates were 21.6% (2.2%) for BMI[1] and 62.5% for height.[2]

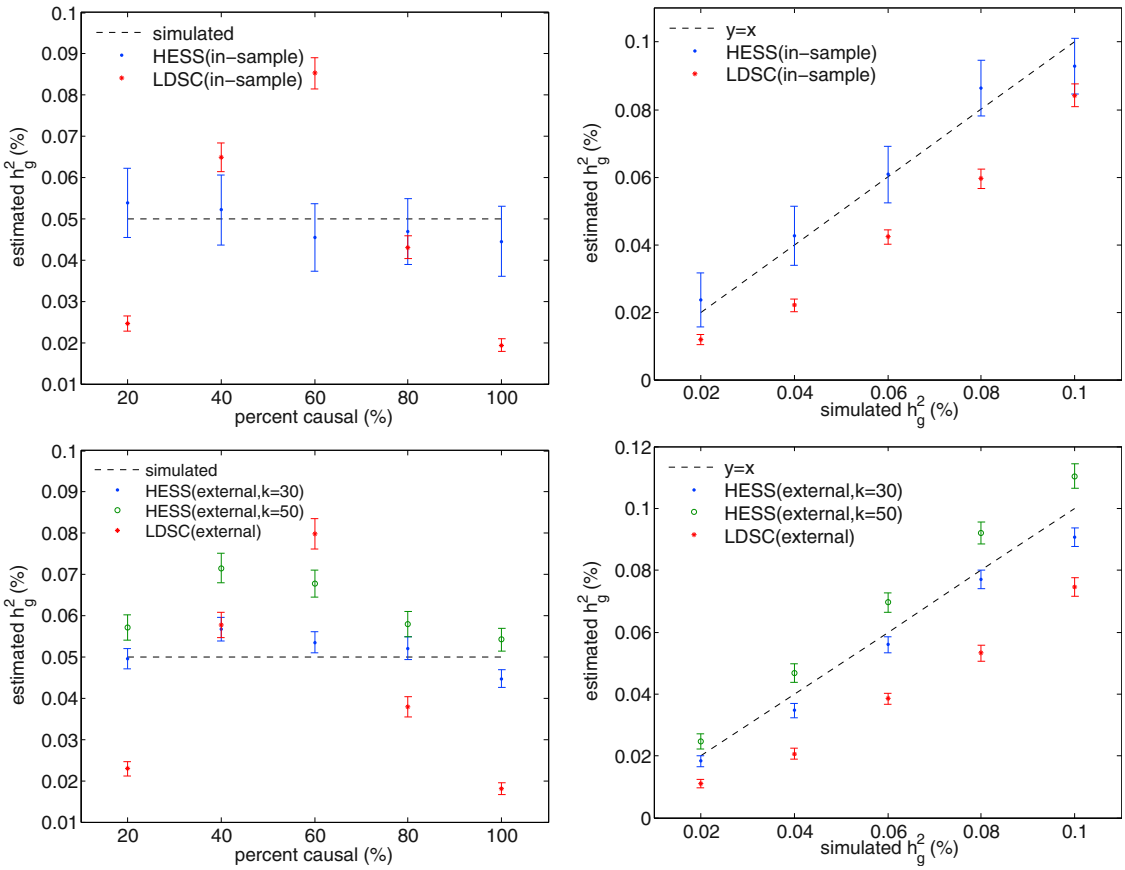

Figure 2. HESS Provides Superior Accuracy over LDSC in Estimating Local Heritability
HESS attains unbiased estimates when in-sample LD is used (top) and approximately unbiased estimates when reference LD is used (bottom). The mean and SE in these figures were computed on the basis of 500 simulations, each involving 50,000 simulated GWAS datasets. Error bars represent twice the SE.

We also found that our estimates of total SNP heritability broadly correlate with those obtained by LDSC (see Figure S7). Most importantly, we found that common SNPs explain a large fraction (ranging from 21% for forearm bone mineral density [BMD] to 94% for HDL) of the previously reported familial heritability for all quantitative traits we interrogated (Table 1). Although we observed a very high contribution of common SNPs to case-control traits as well, we note that our estimator can be biased as a result of ascertainment in this case (see Material and Methods).

Hidden Heritability at Known Risk Loci

Recent works[10,55] have shown that the total heritability explained by all variants at GWAS risk loci ($h^2_{g,\mathrm{local,GWAS}}$) is higher than heritability explained by GWAS index SNPs (h^2_{GWAS}). This suggests that a fraction of the missing heritability is due to multiple causal variants or poor tagging of hidden causal variants at known risk loci. We used HESS to quantify the gap between these two estimates of heritability at known risk loci. We found several traits for which $h^2_{g,\mathrm{local,GWAS}}$ was significantly larger than h^2_{GWAS}. For example, $h^2_{g,\mathrm{local,GWAS}}$ was over two times higher (32.0% [SE = 0.2%]) than h^2_{GWAS} (13.9% [SE = 0.002%]) for height (Table 1). The difference can be accounted for by incomplete tagging of hidden causal variant(s) or allelic heterogeneity (i.e., multiple causal variants). Indeed, conditional analysis identified 36 GWAS loci containing multiple signals of association (for a total of 87 GWAS risk SNPs at these loci) for height.[56] Restricting to the 28 loci containing at least 2 of the 87 GWAS risk SNPs, we estimated $h^2_{g,\mathrm{local,GWAS}}$ at 4.6% (SE = 0.06%), a 2.4-fold increase over h^2_{GWAS} = 1.9% (SE = 0.003%). These loci, 5.8% of all GWAS loci for height, contributed to 14.2% of the difference between $h^2_{g,\mathrm{local,GWAS}}$ and h^2_{GWAS} across all loci, thus suggesting that the difference is most likely due to multiple signals of association. To confirm this hypothesis, we applied a conditional analysis from summary GWAS data by using GCTA-COJO[56] for the traits HDL, TG, rheumatoid arthritis (RA), and schizophrenia (SCZ). We observed that a moderate fraction (2%–16%) of GWAS loci showed multiple signals of association (see Table 2), thus confirming that contrasting $h^2_{g,\mathrm{local,GWAS}}$ with h^2_{GWAS} is indicative of multiple signals of association.

In contrast, the majority of traits showed similar $h^2_{g,\mathrm{local,GWAS}}$ and h^2_{GWAS} (see Table 1), suggesting that these

Table 2. GCTA-COJO Analysis of Summary Statistics for the Traits HDL, TG, RA, and SCZ

Trait	No. of Loci with GWAS Hits	No. of GWAS Loci with Multiple Signals	$h^2_{g,\text{local,GWAS}}$ (% SE)	h^2_{GWAS} (% SE)	Fraction (%)[a]
HDL[3]	92	15	6.1 (0.14)	2.8 (0.003)	67.3
TG[3]	66	9	4.6 (0.12)	3.0 (0.002)	57.1
RA[42]	51	4	14.8 (0.19)	4.3 (0.005)	97.3
SCZ[44]	103	2	0.3 (0.003)	0.2 (0.003)	3.6

We define loci with multiple association signals as loci containing at least two of the risk SNPs reported by GCTA-COJO.[56] Here, we computed $h^2_{g,\text{local,GWAS}}$ and h^2_{GWAS} by restricting to the loci with multiple association signals.
[a]The fraction of difference between $h^2_{g,\text{local,GWAS}}$ and h^2_{GWAS} across all loci that is accounted for by loci with multiple signals of association.

loci have a single causal variant that is very well tagged by the index GWAS variant. For example, it is known that LDL is strongly regulated by a single non-coding functional variant at the SORT1 locus[3,57] and that BMD (femoral neck) is strongly regulated by WNT16.[58,59] We also observed traits (e.g., mean cell hemoglobin, mean cell volume, and red blood cell count) for which $h^2_{g,\text{local,GWAS}}$ was estimated to be less than h^2_{GWAS}. This seemingly contradictory result is due to the fact that fewer eigenvectors in the truncated-SVD regularization of LD matrices were used for estimating $h^2_{g,\text{local,GWAS}}$ for GWASs with small sample sizes (see Table S2), resulting in downward bias (see Material and Methods).

Contrasting Polygenicity across Multiple Complex Traits

Most studied common traits exhibit a strong polygenic architecture (i.e., an abundance of small-effect loci contributing to a trait).[1–3,9] We recapitulated this observation by using the HESS analysis (Figures 3, S16, and S17) and found a strong correlation between chromosome length and the fraction of heritability it explains for most traits that we have analyzed here (Figures 4 and 5). Consistent with previous findings,[60] we also observed regions (such as fat mass and obesity associated [*FTO*] on chromosome 16 and human leukocyte antigen [HLA] on chromosome 6) contributing disproportionately to the fraction of heritability for HDL, BMI, and RA.

Next, we sought to quantify the variability in polygenicity across traits. We rank ordered loci on the basis of their estimated local SNP heritability, summed their contribution, and plotted it against the percentage of genome they occupy (Figure 6). For highly polygenic traits, we expected the cumulative fraction of total SNP heritability to be proportional to the fraction of genome covered, whereas for less polygenic traits, we expected to see a small genomic fraction accounting for a large fraction of the total SNP heritability. For example, in SCZ and height, the top 1% of loci with the highest local SNP heritability contributed to 4.2% (SE = 1.0%) and 6.5% (SE = 1.5%), respectively, of the total SNP heritability of these traits. This is consistent with previous reports on the degree of polygenicity of these traits.[2,3,9] At the other extremes, RA and lipid traits (HDL, LDL, TC, and TG) had a lower degree of polygenicity, such that the top 1% of loci accounted for 14%–30% of the total SNP heritability. However, the low polygenicity of RA was mostly driven by the HLA region on chromosome 6. After removing estimates of local SNP heritability at loci overlapping the HLA region for all traits, we observed that RA showed a moderate degree of polygenicity for the rest of the genome (see Figure S9). We also note that the different degrees of polygenic signals across traits reflect both a difference in disease architecture (i.e., distribution of effect sizes) and a difference in the sample sizes for the GWAS summary data.

A different perspective of polygenicity is to restrict to GWAS risk loci (because they clearly contain risk variants) and contrast the proportion of explained variance with the proportion of the genome they occupy. We observed a wide distribution across traits reflecting diverse genetic

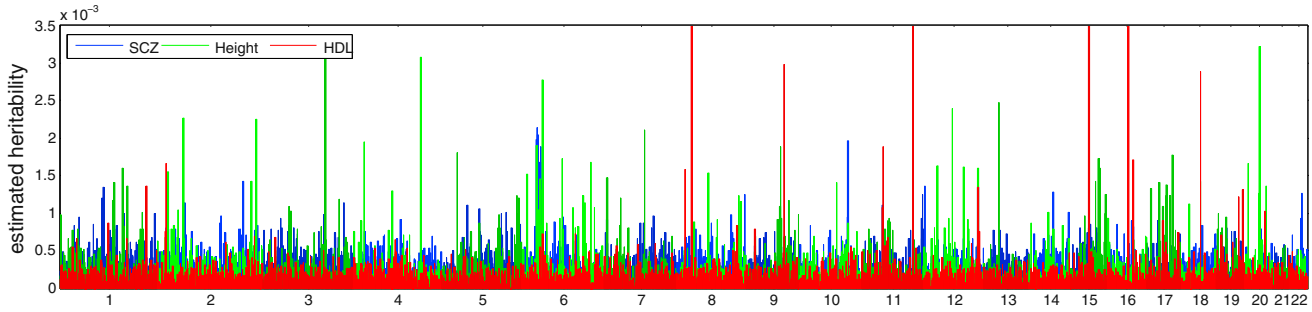

Figure 3. Manhattan-Style Plots of Regional Heritability across the Genome for the Traits Height, HDL, and SCZ
See Figures S16 and S17 for results across all traits.

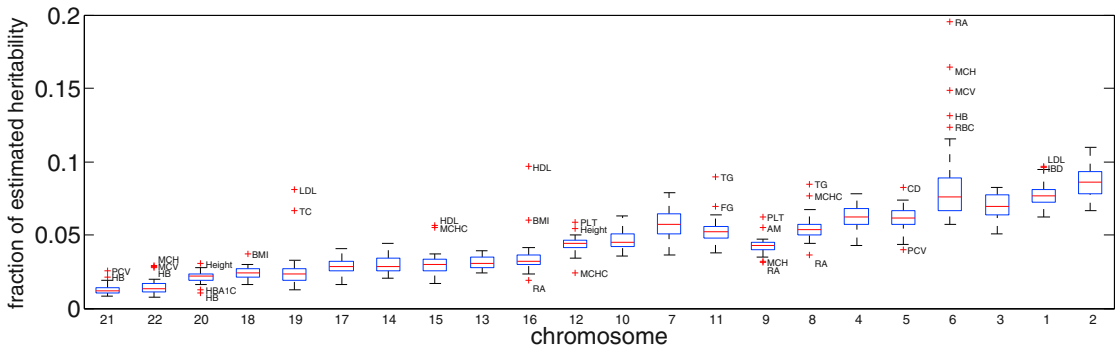

Figure 4. Fraction of h_g^2 per Chromosome across the 30 Traits Studied
Here, we obtained the chromosomal heritability by summing local heritability at loci within the chromosome. For each chromosome, we plot the box plots of estimates at the 30 considered traits. Chromosomes are ordered by size. With some notable exceptions, all traits show a strong polygenic signature of genetic architecture.

architectures, as well as different sample sizes for the GWASs performed for these traits. For example, approximately 30% of loci across the genome harbored a risk variant for height and accounted for 50% of the total SNP heritability (a 1.5-fold enrichment). On the other hand, although only 5% of the loci contained GWAS risk variants for HDL, these loci collectively explained 25% of the SNP heritability of HDL (a 4.6-fold enrichment; Figure 7).

Loci Contributing to Heritability of Multiple Traits

It has been previously established that a number of the 30 traits investigated in this study share a genetic basis.[61] Correlating estimates of local SNP heritability across the entire genome can serve as a proxy for the magnitude of pleiotropy, and we can identify pairs of traits whose genetic components tend to localize within the same regions of the genome (Figure S10). Motivated by this, we searched for specific pleiotropic loci that we define as loci that contribute significant non-zero SNP heritability (one-tailed p value < 0.05, Bonferroni corrected for 1,703 loci) for at least 3 out of the 30 analyzed traits. In total, we identified 36 such loci distributed genome-wide (see Figure 8 and Figure S11).

As expected, the HLA region (chr6: 26–34 Mb) displayed strong pleiotropic signal, particularly for immunologically relevant phenotypes (see Figure 8). For instance, the locus chr6: 32–33 Mb contributed a significant amount of SNP heritability for eight traits and had exceptionally strong signals for RA, ulcerative colitis, and inflammatory bowel disease (see Figure 8). We also observed several other pleiotropic loci, including chr2: 199–202 Mb (contributing to age at menarche, SCZ, and height), chr6: 134–136 Mb (contributing to multiple red blood cell traits), and chr19: 45–46 Mb (contributing to multiple lipid traits). It is well known that genetic correlations exist among red blood cell traits,[24,62,63] as well as among lipid traits.[3,61] Interestingly, previous research has also revealed that early age at menarche is associated with later onset of SCZ.[64] Our results suggest that these genetic correlations and associations might be caused in part by the pleiotropic effect of these loci.

We note that the selection of traits can bias the identification of pleiotropic loci toward over-represented traits such as height and lipid traits. Nevertheless, analyzing local SNP heritability is still a useful tool for quantifying the fraction of total SNP heritability contributed by a single locus and provides valuable insights into identifying pleiotropic loci.

Discussion

We have presented HESS, an unbiased estimator of local SNP heritability from GWAS summary data. We extend existing work[16] that estimates heritability under the fixed-effects model by proposing to regularize the external reference LD matrix via truncated SVD and generalizing the estimator to multiple independent loci. Through extensive simulations, we demonstrate that HESS is unbiased when given in-sample LD and yields more consistent and less biased estimates of local SNP heritability than LDSC given external reference LD. We applied HESS on GWAS summary data of 30 complex traits from 12 GWAS consortia and showed that our results recapitulate previous findings. We then used these estimates of local SNP heritability to contrast polygenicity of complex traits, found loci with multiple causal variants, and identified heritability hotspots. We note that enrichment of heritability at GWAS risk loci could be leveraged into prioritizing GWASs or fine mapping; for example, traits with small enrichment of heritability at GWAS risk loci are more suitable for larger GWASs, whereas traits with large enrichment of heritability at known risk loci could be investigated further through fine mapping.

In this work, we focus on estimating local heritability attributable to common autosomal variants (MAF $> 5\%$) and ignore potential heritability captured by the sex

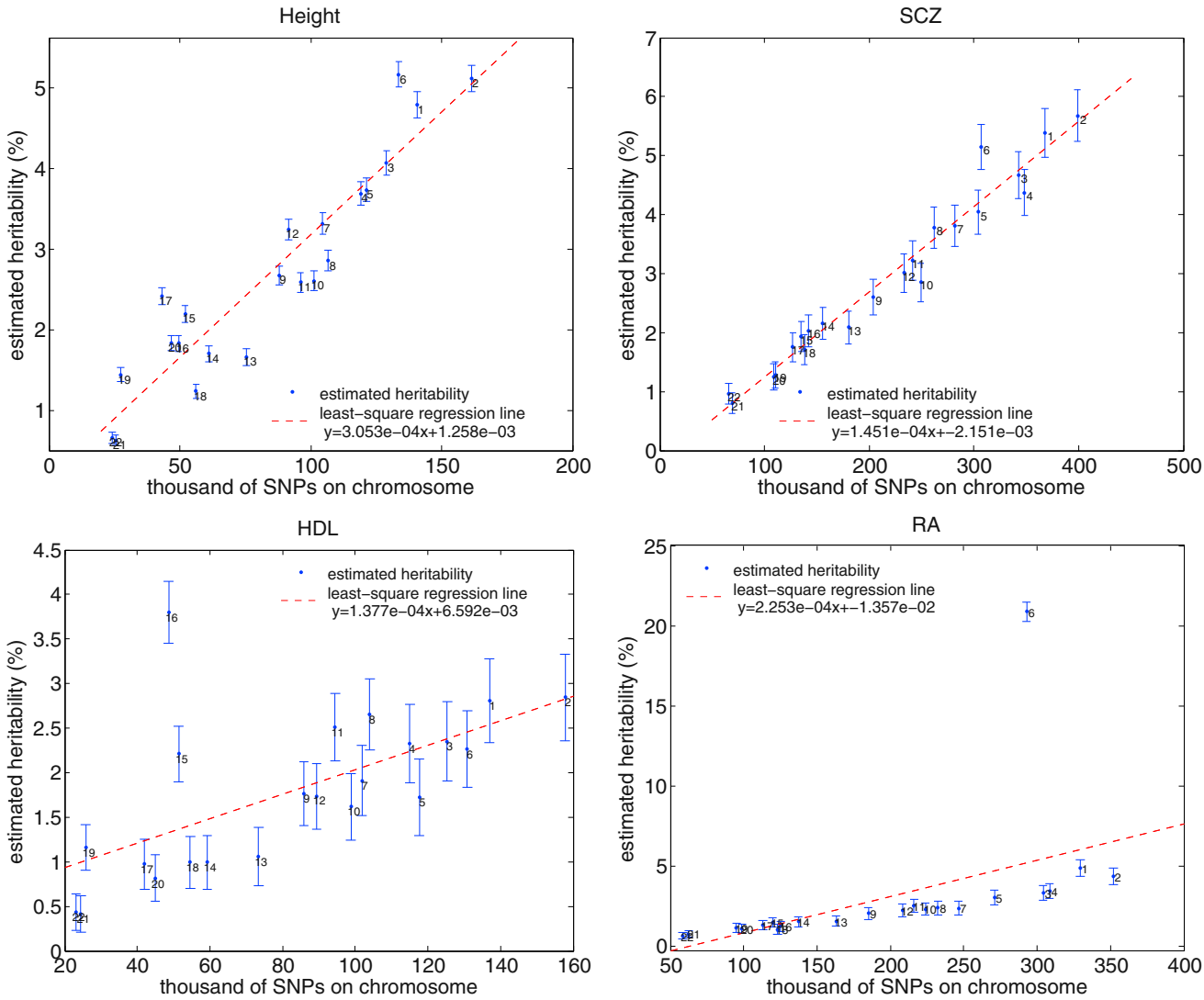

Figure 5. Heritability Attributable to Each Chromosome for Four Example Traits
We obtained the chromosomal heritability by summing local heritability at loci within the chromosome. We obtained SE by taking the square root of the sum of variance estimation. See Figures S13–S15 for results across all traits. Error bars represent twice the SE.

chromosomes and rare variants. We also note that our heritability estimates for case-control traits are not adjusted for ascertainment because it is unclear whether adjustment derived from the infinitesimal model can be directly applied for the fixed-effects model. Thus, our reported heritability estimation for case-control traits can be biased as a result of ascertainment. Future work that addresses estimation of local heritability including both common and rare variants, sex chromosomes, and adjustment of heritability estimates under the fixed-effects model for case-control traits will further improve the utility of our approach.

We conclude with several caveats and limitations of our work. First, our method relies on independent LD blocks, which are often hard to define as a result of LD leakage across multiple loci. In this work, we attempt to minimize LD leakage by using principled approaches to define approximately independent loci. Second, when only external reference LD is available, our method can yield biased estimates because external reference LD usually has a lower rank than its corresponding in-sample LD. Furthermore, cryptic relatedness in the GWAS data could also bias our estimation procedure. This makes hypothesis testing difficult. However, with in-sample LD and larger reference panels, such as the Haplotype Reference Consortium,[65] this bias will be reduced because LD can be inferred more precisely. We also note that our estimated λ_{GC} can be a potential source of bias; thus, our genome-wide estimate should be interpreted with caution. Third, for stable estimation, the number of eigenvectors used (k) in the truncated-SVD regularization should be chosen on the basis of the GWAS sample size—GWASs with large sample sizes can afford a large k, whereas GWASs with small sample sizes should use a small k. We recommend applying our method to summary data obtained from GWASs involving around or above 50,000 samples. For GWASs

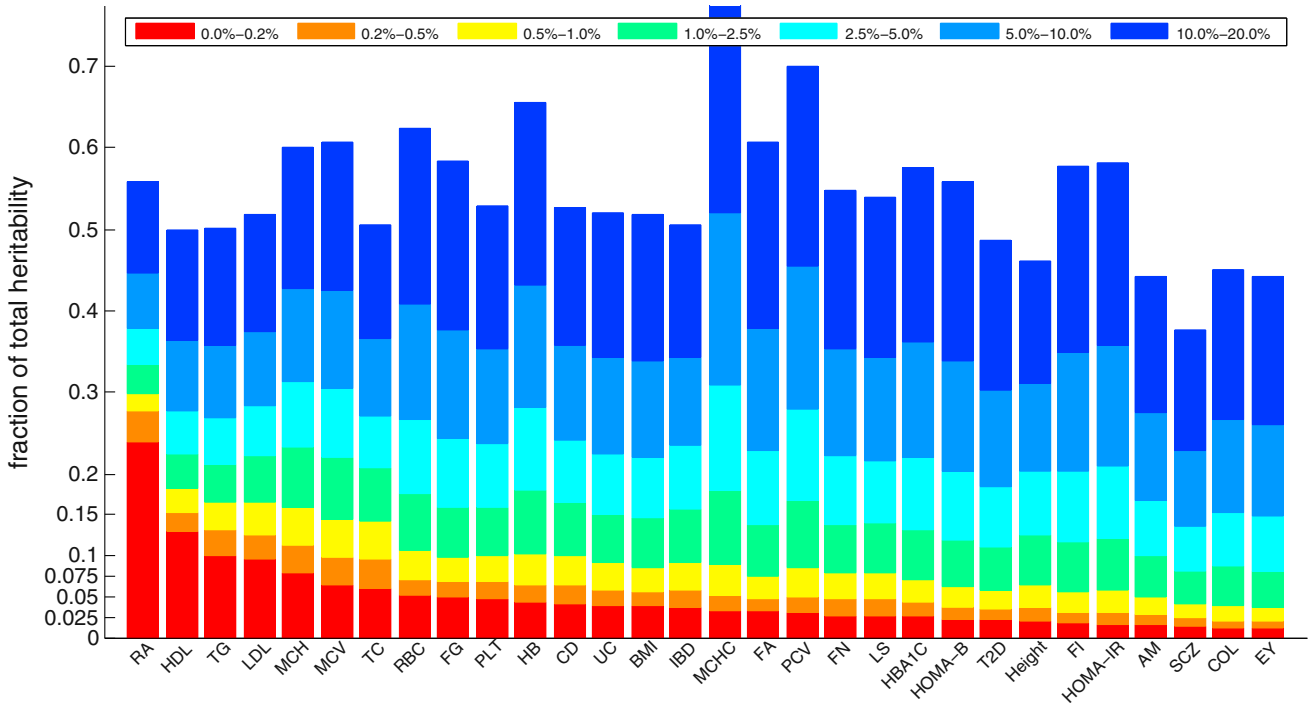

Figure 6. Stacked Bar Plot Showing the Percentage of Total Heritability Attributable to Different Fractions of the Genome
We rank ordered all genomic loci by their explained heritability and quantified the fraction of total heritability attributable to different percentile ranges. Traits with high polygenicity (e.g., height and SCZ) tend to have bars with a height proportional to bin size, whereas less polygenic traits (e.g., RA and HDL) tend to have bars much larger than bin size.

with small sample sizes, when genome-wide SNP heritability is known, one can still apply Equation 14 to obtain stable estimates of local heritability. We also note that although using the same number of eigenvectors for all loci facilitates the study of the statistical properties of our estimator, this approach might not be optimal for all loci. We conjecture that selecting k according to a more principled approach (e.g., on the basis of the distribution of eigenvalues) might reduce bias, and we leave such investigation as future work.

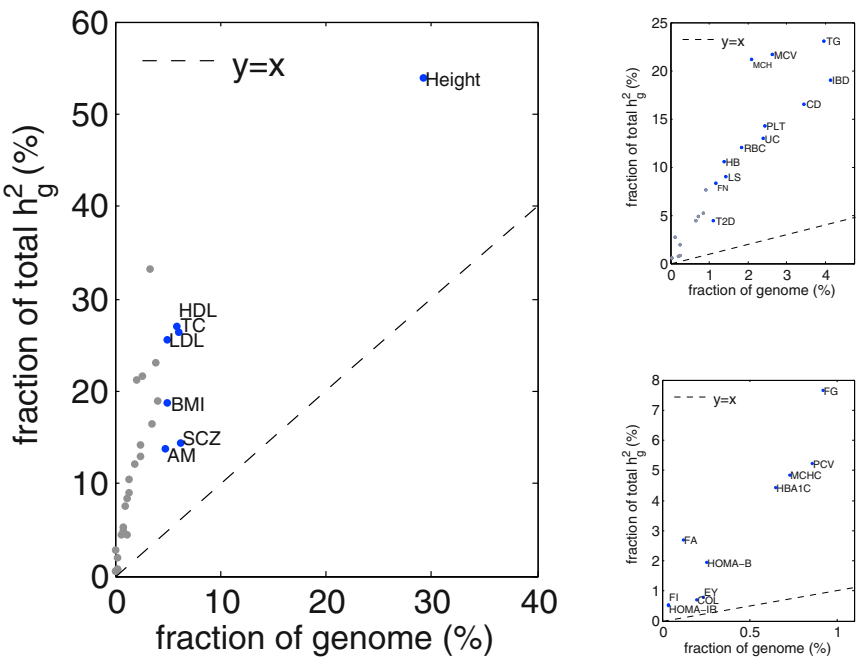

Figure 7. Fraction of h_g^2 Explained by All Loci Containing a GWAS Hit versus the Fraction of Genome Covered by These Loci
Images on the right focus successively on the traits near the bottom left.

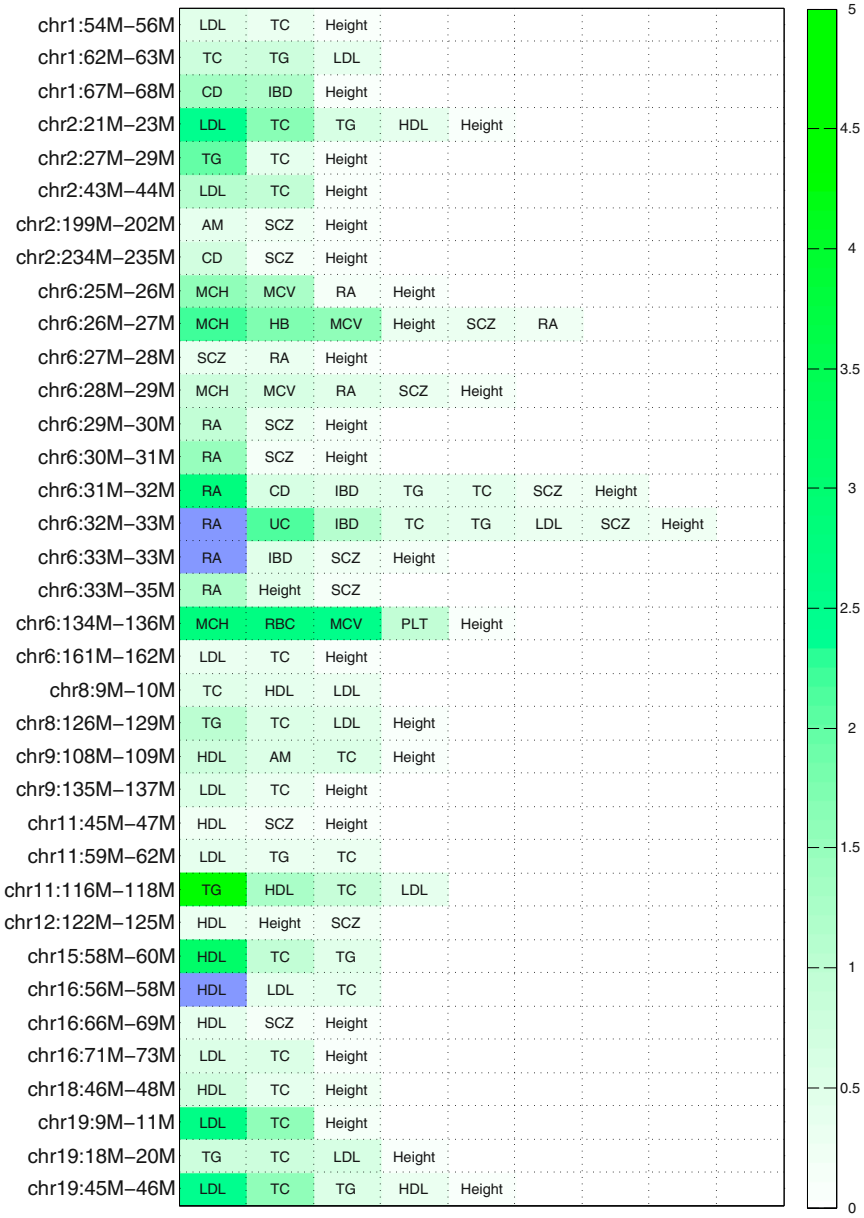

Figure 8. Heatmap Showing the Fraction of Total SNP Heritability, h_g^2, Contributed by Each of the 36 "Pleiotropic" Loci

For each locus, we only display traits to which the locus contributes a significant amount of heritability. We mark traits to which the locus contributes more than 5% of the total SNP heritability in dark blue.

Supplemental Data

Supplemental Data include 20 figures and 2 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.05.013.

Acknowledgments

We are very grateful to Alkes Price, Yakir Reshef, Brielin Brown, and Nicholas Mancuso for their helpful discussions and feedback, which greatly improved the quality of this manuscript. We also would like to thank the anonymous reviewers for their insightful suggestions. We also thank Dr. Nicole Soranzo for kindly sharing summary data for the platelet traits.

Received: December 30, 2015
Accepted: May 9, 2016
Published: June 23, 2016

Web Resources

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

References

1. 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; International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.
2. 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; 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.

3. 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.
4. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42* (Database issue, D1), D1001–D1006.
5. 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.
6. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015a). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
7. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
8. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
9. Loh, P.R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al.; Schizophrenia Working Group of Psychiatric Genomics Consortium (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392.
10. Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., et al. (2013). Quantifying missing heritability at known GWAS loci. PLoS Genet. *9*, e1003993.
11. Palla, L., and Dudbridge, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. Am. J. Hum. Genet. *97*, 250–259.
12. Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. Nat. Rev. Genet. *3*, 872–882.
13. 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.
14. Golan, D., Lander, E.S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. Proc. Natl. Acad. Sci. USA *111*, E5272–E5281.
15. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
16. Gamazon, E.R., Cox, N.J., and Davis, L.K. (2014). Structural architecture of SNP effects on complex traits. Am. J. Hum. Genet. *95*, 477–489.
17. 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
18. Hansen, P.C. (1987). The truncated SVD as a method for regularization. BIT Numerical Mathematics *27*, 534–553.
19. Elman, R., Karpenko, N., and Merkurjev, A. (2008). The algebraic and geometric theory of quadratic forms, *Volume 56* (American Mathematical Society).
20. Ben-Israel, A., and Greville, T.N.E. (2003). Generalized inverses: theory and applications, *Volume 15* (Springer Science & Business Media).
21. 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.
22. Kichaev, G., and Pasaniuc, B. (2015). Leveraging functional-annotation data in trans-ethnic fine-mapping studies. Am. J. Hum. Genet. *97*, 260–271.
23. Hemani, G., Yang, J., Vinkhuyzen, A., Powell, J.E., Willemsen, G., Hottenga, J.J., Abdellaoui, A., Mangino, M., Valdes, A.M., Medland, S.E., et al. (2013). Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. Am. J. Hum. Genet. *93*, 865–875.
24. 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.
25. Garner, C., Tatu, T., Reittie, J.E., Littlewood, T., Darley, J., Cervino, S., Farrall, M., Kelly, P., Spector, T.D., and Thein, S.L. (2000). Genetic influences on F cells and other hematologic variables: a twin heritability study. Blood *95*, 342–346.
26. Lin, J.-P., O'Donnell, C.J., Jin, L., Fox, C., Yang, Q., and Cupples, L.A. (2007). Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study. Am. J. Hematol. *82*, 605–610.
27. Hinckley, J.D., Abbott, D., Burns, T.L., Heiman, M., Shapiro, A.D., Wang, K., and Di Paola, J. (2013). Quantitative trait locus linkage analysis in a large Amish pedigree identifies novel candidate loci for erythrocyte traits. Mol. Genet. Genomic Med. *1*, 131–141.
28. 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.
29. 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; MAGIC investigators (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. *42*, 105–116.
30. Simonis-Bik, A.M., Eekhoff, E.M., Diamant, M., Boomsma, D.I., Heine, R.J., Dekker, J.M., Willemsen, G., van Leeuwen, M., and de Geus, E.J. (2008). The heritability of HbA1c and fasting blood glucose in different measurement settings. Twin Res. Hum. Genet. *11*, 597–602.

31. Rasmussen-Torvik, L.J., Pankow, J.S., Jacobs, D.R., Steffen, L.M., Moran, A.M., Steinberger, J., and Sinaiko, A.R. (2007). Heritability and genetic correlations of insulin sensitivity measured by the euglycaemic clamp. Diabet. Med. *24*, 1286–1289.
32. 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.
33. Mills, G.W., Avery, P.J., McCarthy, M.I., Hattersley, A.T., Levy, J.C., Hitman, G.A., Sampson, M., and Walker, M. (2004). Heritability estimates for beta cell function and features of the insulin resistance syndrome in UK families with an increased susceptibility to type 2 diabetes. Diabetologia *47*, 732–738.
34. Zaitlen, N., Pasaniuc, B., Sankararaman, S., Bhatia, G., Zhang, J., Gusev, A., Young, T., Tandon, A., Pollack, S., Vilhjálmsson, B.J., et al. (2014). Leveraging population admixture to characterize the heritability of complex traits. Nat. Genet. *46*, 1356–1362.
35. de Miranda Chagas, S.V., Kanaan, S., Chung Kang, H., Cagy, M., de Abreu, R.E., da Silva, L.A., Garcia, R.C., and Garcia Rosa, M.L. (2011). Environmental factors, familial aggregation and heritability of total cholesterol, low density lipoprotein-cholesterol and high density lipoprotein-cholesterol in a Brazilian population assisted by the Family Doctor Program. Public Health *125*, 329–337.
36. Elbein, S.C., and Hasstedt, S.J. (2002). Quantitative trait linkage analysis of lipid-related traits in familial type 2 diabetes: evidence for linkage of triglyceride levels to chromosome 19q. Diabetes *51*, 528–535.
37. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al.; LifeLines Cohort Study (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science *340*, 1467–1471.
38. 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; UK10K Consortium (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature *526*, 112–117.
39. Arden, N.K., Baker, J., Hogg, C., Baan, K., and Spector, T.D. (1996). The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. J. Bone Miner. Res. *11*, 530–534.
40. 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; Early Growth Genetics (EGG) Consortium (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature *514*, 92–97.
41. Towne, B., Czerwinski, S.A., Demerath, E.W., Blangero, J., Roche, A.F., and Siervogel, R.M. (2005). Heritability of age at menarche in girls from the Fels Longitudinal Study. Am. J. Phys. Anthropol. *128*, 210–219.
42. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al.; RACI consortium; GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
43. Harney, S.M.J., Vilariño-Güell, C., Adamopoulos, I.E., Sims, A.-M., Lawrence, R.W., Cardon, L.R., Newton, J.L., Meisel, C., Pointon, J.J., Darke, C., et al. (2008). Fine mapping of the MHC Class III region demonstrates association of AIF1 and rheumatoid arthritis. Rheumatology (Oxford) *47*, 1761–1767.
44. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
45. Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch. Gen. Psychiatry *60*, 1187–1192.
46. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.; International Multiple Sclerosis Genetics Consortium; International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.
47. Tysk, C., Lindberg, E., Järnerot, G., and Flodérus-Myrhed, B. (1988). Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. Gut *29*, 990–996.
48. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.
49. Poulsen, P., Kyvik, K.O., Vaag, A., and Beck-Nielsen, H. (1999). Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. Diabetologia *42*, 139–145.
50. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.
51. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.
52. Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G., et al. (2011). Quality control procedures for genome-wide association studies. Curr. Protoc. Hum. Genet. *Chapter 1*, 19.
53. Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M., et al.; GIANT Consortium (2011b). Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. *19*, 807–812.
54. 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.
55. Mancuso, N., Rohland, N., Rand, K.A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., et al.; PRACTICAL consortium (2016). The contribution of rare variation to prostate cancer heritability. Nat. Genet. *48*, 30–35.

56. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
57. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.
58. Zheng, H.F., Tobias, J.H., Duncan, E., Evans, D.M., Eriksson, J., Paternoster, L., Yerges-Armstrong, L.M., Lehtimäki, T., Bergström, U., Kähönen, M., et al. (2012). WNT16 influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. PLoS Genet. *8*, e1002745.
59. Koller, D.L., Zheng, H.F., Karasik, D., Yerges-Armstrong, L., Liu, C.T., McGuigan, F., Kemp, J.P., Giroux, S., Lai, D., Edenberg, H.J., et al. (2013). Meta-analysis of genome-wide studies identifies WNT16 and ESR1 SNPs associated with bone mineral density in premenopausal women. J. Bone Miner. Res. *28*, 547–558.
60. Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). Fto obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. *373*, 895–907.
61. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
62. Ganesh, S.K., Zakai, N.A., van Rooij, F.J., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat. Genet. *41*, 1191–1198.
63. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al.; BioBank Japan Project; CHARGE Consortium (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum. Mol. Genet. *22*, 2529–2538.
64. Cohen, R.Z., Seeman, M.V., Gotowiec, A., and Kopala, L. (1999). Earlier puberty as a predictor of later onset of schizophrenia in women. Am. J. Psychiatry *156*, 1059–1064.
65. McCarthy, S., Das, S., Kretzschmar, W., Durbin, R., Abecasis, G., and Marchini, J. (2015). A reference panel of 64,976 haplotypes for genotype imputation. bioRxiv http://dx.doi.org/10.1101/035170.