bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Contrasting the genetic architecture of 30 complex traits from summary association data

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

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, 90024
[2]Dept of Pathology and Laboratory Medicine
[3]Dept of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, 90024

Abstract

Variance components methods that estimate the aggregate contribution of large sets of variants to the heritability of complex traits have yielded important insights into the disease architecture of common diseases. Here, we introduce new methods that estimate the total variance in trait explained by a single locus in the genome (local heritability) from summary GWAS data while accounting for linkage disequilibrium (LD) among variants. We apply our new estimator to ultra large-scale GWAS summary data of 30 common traits and diseases to gain insights into their local genetic architecture. First, we find that common SNPs have a high contribution to the heritability of all studied traits. Second, we identify traits for which the majority of the SNP heritability can be confined to a small percentage of the genome. Third, we identify GWAS risk loci where the entire locus explains significantly more variance in the trait than the GWAS reported variants. Finally, we identify 55 loci that explain a large proportion of heritability across multiple traits.

Introduction

Large-scale genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphisms (SNPs) associated with hundreds of traits[1,2,3,4]. However, only a

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fraction of the variance in trait can be explained by the risk SNPs reported by GWAS. The so-called "missing heritability problem" is in part due to the stringent significance threshold imposed in GWAS, which neglects causal variants that fail to reach the genome-wide significance level at current sample sizes. As an alternative, variance component (heritability) analysis aggregates the effect of all SNPs regardless of their significance to increase accuracy[5] and has yielded important insights into the genetic architecture of complex traits[6,7,8,9,10,11].

Standard approaches for heritability estimation and partitioning rely on estimating the genetic relationships between pairs of individuals (estimated genome-wide or across a subset of the genome)[8,12,13]. Therefore, these analyses require individual-level genotype data which prohibits their applicability to ultra-large GWAS that, due to privacy concerns, is typically available only at the summary level. To solve this bottleneck, recent methods have shown that heritability, both genome-wide as well as for different functional categories in the genome, can be accurately estimated using only summary GWAS data[6,7]. Although these methods have enabled powerful analyses making insights into genetic basis of complex traits, they rely on the infinitesimal model assumption (i.e. all SNPs contribute to the trait) which is invalid at most risk loci[6,7]. To overcome this drawback, alternative approaches have proposed to impose a prior on the sparsity of effect sizes to further increase heritability estimation accuracy[14]. 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 heritability estimation procedure. Indeed, recent works have shown that heritability estimation can be performed under maximum-likelihood from polygenic scores under a fixed-effect model assuming no LD among SNPs[11].

Here, we introduce Heritability Estimator from Summary Statistics (HESS), an approach to estimate the variance in trait explained by a single locus while accounting for linkage disequilibrium (LD) among SNPs. We build upon recent works[11,15] that treat 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

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

disease architecture (i.e. number of causals and distribution of effect sizes). We extend our method to use LD estimated from reference panels[16] and show that a principal components based regularization of the LD matrix[17] yields approximately unbiased and more consistent estimates of local heritability as compared to existing methods[6].

We applied HESS to partition common SNP heritability at each locus in the genome using GWAS summary data for 30 traits spanning over 10 million SNPs and 2.4 million phenotype measurements. First, we show tht common SNPs explain a large fraction of the total familial heritability, ranging anywhere from 20% to 70% across the studied quantitative traits. Second, we showcase the utility of local heritability estimates in finding loci that explain more variance in trait 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 highest local heritability. We find that most of the 30 selected traits are highly polygenic with a small number of traits driven by a small number of loci. Finally, we report 55 "heritability hotspots" – regions of genome that have a high contribution to the heritability of multiple traits. Taken together, our results give insights into traits where further GWAS and/or fine-mapping studies are likely to recover a significant amount of the missing heritability.

Materials and methods

Overview of methods

We introduce estimators for the variance in trait explained by typed variants at a single locus (local heritability, $h^2_{g,local}$) from summary GWAS data (i.e. Z-scores, effect sizes and their standard errors). We derive our estimator under the assumption that effect sizes at causal variants are fixed and genotypes are drawn from a distribution with a pre-specified covariance structure. The covariance, (i.e. pairwise correlation between any variants at a locus, LD) can be estimated in-sample, from the genotype data in GWAS, or from external reference panels (e.g. 1000 Genomes Project[16]). The finite sample size of the GWAS studies as well as the reference panels used to estimate LD induces statistical noise that needs to be

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

accounted for to obtain an accurate estimation. Since the top projections make up the bulk of the summation, truncated-SVD lends itself as the 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.

Estimating 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 genotypes of individual i at the p 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. Assuming that $\boldsymbol{\beta}$ is fixed and $\mathbf{X}$ is random, the variance is

$$\mathrm{Var}[\mathbf{y}] = \mathrm{Var}[\mathbf{X}\boldsymbol{\beta}] + \sigma_e^2 = \boldsymbol{\beta}^T\mathrm{Cov}[\mathbf{X}]\boldsymbol{\beta} + \sigma_e^2 = \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta} + \sigma_e^2 \tag{1}$$

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

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

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. Extending to p SNPs at the locus, if follows that $\hat{\boldsymbol{\beta}}_{gwas} = \mathbf{V}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}^T\boldsymbol{\epsilon}$ where $\mathbf{V}$ is the LD matrix. With $\boldsymbol{\beta}$ fixed and ϵ random, $\hat{\boldsymbol{\beta}}_{gwas}$ is a random variable with $\mathrm{E}[\hat{\boldsymbol{\beta}}_{gwas}] = \mathrm{E}[\mathbf{V}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}^T\boldsymbol{\epsilon}] = \mathbf{V}\boldsymbol{\beta}$, and $\mathrm{Cov}[\hat{\boldsymbol{\beta}}_{gwas}] = \mathrm{Var}[\mathbf{V}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}^T\boldsymbol{\epsilon}] = \frac{1}{n^2}\mathbf{X}^T\mathrm{Cov}[\boldsymbol{\epsilon}]\mathbf{X} = \frac{\sigma_e^2}{n}\mathbf{V} = \frac{1-h^2_{g,local}}{n}\mathbf{V}$. By central limit theorem, $\hat{\boldsymbol{\beta}}_{gwas} \sim N\left(\mathbf{V}\boldsymbol{\beta}, \frac{1-h^2_{g,local}}{n}\mathbf{V}\right)$.

As GWAS sample size (n) increases $\hat{\boldsymbol{\beta}}_{gwas}$ converges to $\boldsymbol{\beta}_{gwas} = \mathbf{V}\boldsymbol{\beta}$. By simple substi-

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tution in Equation (1) it follows that an estimator for $h^2_{g,local}$ is

$$(\boldsymbol{\beta}^T_{gwas}\mathbf{V}^{-1})\mathbf{V}(\mathbf{V}^{-1}\boldsymbol{\beta}_{gwas}) = \boldsymbol{\beta}^T_{gwas}\mathbf{V}^{-1}\boldsymbol{\beta}_{gwas} \tag{3}$$

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

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

$$\mathrm{E}[\hat{\boldsymbol{\beta}}^T_{gwas}\mathbf{V}^{-1}\hat{\boldsymbol{\beta}}_{gwas}] = \mathrm{tr}\left(\frac{1-h^2_{g,local}}{n}\mathbf{V}^{-1}\mathbf{V}\right) + h^2_{g,local} = \frac{1-\mathrm{E}[\hat{h}^2_{g,local}]}{n}p + \mathrm{E}[\hat{h}^2_{g,local}]. \tag{4}$$

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

$$\hat{h}^2_{g,local} = \frac{n\hat{\boldsymbol{\beta}}^T_{gwas}\mathbf{V}^{-1}\hat{\boldsymbol{\beta}}_{gwas} - p}{n-p}. \tag{5}$$

Following quadratic form theory[18], the variance of $\hat{h}^2_{g,local}$ is

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

Since $h^2_{g,local}$, the true local heritability, is unknown, we use $\hat{h}^2_{g,local}$ instead. For $h^2_{g,local}$ near 0, $\mathrm{Var}[\hat{h}^2_{g,local}] \approx \frac{4}{(n-p)^2}h^2_{g,local} + \frac{2p}{(n-p)^2}$ through Taylor expansion around 0. Thus, the plug in principle yields an estimation of $\mathrm{Var}[\hat{h}^2_{g,local}]$ approximately equal to the truth in the expectation.

Accounting for rank deficiencies in the LD

In the above derivation we made the assumption that the inverse of the LD matrix $\mathbf{V}$ exists. In practice, however, due to pairs of SNPs in perfect LD, $\mathbf{V}$ is usually rank

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

deficient, and thus $\mathbf{V}^{-1}$ does not exist. In such cases we use the Moore-Penrose pseudoinverse $\mathbf{V}^{\dagger}$. 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 rank-deficient LD matrix, we obtain an unbiased estimator, $\hat{h}^2_{g,local} = \frac{n\hat{\boldsymbol{\beta}}_{gwas}^T\mathbf{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas}-q}{n-q}$. We make the same adjustment (replacing p with q) in the variance estimator for $\hat{h}^2_{g,local}$.

Adjusting for noise in external reference LD

When genotype data of GWAS samples is not available, we substitute the in-sample LD matrix $\mathbf{V}$ with external reference LD matrix $\hat{\mathbf{V}}$ estimated from the 1000 Genomes Project[16] using a population that matches the GWAS samples. However, due to limited sample size, external reference LD matrices contain statistical noise that biases our estimate. We apply truncated SVD regularization to remove noise from external reference LD matrix as follows.

First note that $\hat{\boldsymbol{\beta}}_{gwas}^T\mathbf{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas} = \sum_{i=1}^q s_i = \sum_{i=1}^q \frac{1}{w_i}(\hat{\boldsymbol{\beta}}_{gwas}^T\mathbf{u}_i)^2$, where w_i and $\mathbf{u}_i$ are the eigenvalues and eigenvectors of the LD matrix $\mathbf{V}$, and $q = \text{rank}(\mathbf{V})$. For external reference LD matrix $\hat{\mathbf{V}}$ with eigenvalues and eigenvectors $\hat{w}_i$ and $\hat{\mathbf{u}}_i$, we have the same decomposition except that s_i is replaced by $\hat{s}_i = \frac{1}{\hat{w}_i}(\hat{\boldsymbol{\beta}}_{gwas}^T\hat{\mathbf{u}}_i)^2$. Through simulations, we show that most of the signal in $\hat{\boldsymbol{\beta}}_{gwas}^T\mathbf{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas}$ comes from s_i where $i \ll q$ and that $\hat{s}_i \approx s_i$ for $i \ll q$ (see Supplementary Figure 1). In our previous works[19,20], we propose to regularize $\hat{\mathbf{V}}$ using ridge regression penalty. This regularization method is equivalent to replacing $\hat{w}_i$ with $\hat{w}_i + \lambda$, where λ is the penalty. We notice that a large λ is needed to drive down the noise ($\hat{s}_i$ for large i), which diminishes the true signal at the same time. These results motivate us to apply truncated-SVD to remove noise in $\hat{\mathbf{V}}$, i.e. we estimate $\hat{\boldsymbol{\beta}}_{gwas}^T\mathbf{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas}$ by $\sum_{i=1}^k 1/\hat{w}_i(\hat{\boldsymbol{\beta}}_{gwas}^T\hat{\mathbf{u}}_i)^2$, where $k \ll q$. Let $g(\hat{\boldsymbol{\beta}}_{gwas}, k) = \sum_{i=1}^k \frac{1}{\hat{w}_i}(\hat{\boldsymbol{\beta}}_{gwas}^T\hat{\mathbf{u}}_i)^2$, through eigen-decomposition of $\hat{\mathbf{V}}$, it can be shown that

$$\text{E}[g(\hat{\boldsymbol{\beta}}_{gwas}, k)] = \frac{k(1-h^2_{g,local})}{n} + \sum_{i=1}^k \hat{w}_i(\hat{\mathbf{u}}_i^T\boldsymbol{\beta})^2. \tag{7}$$

Since the true local heritability is $h^2_{g,local} = \sum_{i=1}^q w_i(\mathbf{u}_i^T\boldsymbol{\beta})^2$, assuming $\hat{\mathbf{u}}_i = \mathbf{u}_i$ for $i \ll q$, Equation (7) is an approximation of $h^2_{g,local}$ with bias $\frac{k(1-h^2_{g,local})}{n}$. Correcting for this bias

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

yields the estimator for the single-locus case

$$\tilde{h}^2_{g,local} = \frac{ng(\hat{\boldsymbol{\beta}}_{gwas}, k) - k}{n - k}. \tag{8}$$

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

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. Based on the revised model, we decompose $\mathrm{Var}[\mathbf{y}]$ into

$$\mathrm{Var}[\mathbf{y}] = \mathrm{Var}[\mathbf{X}_1\boldsymbol{\beta}_1] + \cdots + \mathrm{Var}[\mathbf{X}_m\boldsymbol{\beta}_m] + \sigma^2_e = h^2_{g,local,1} + \cdots + h^2_{g,local,m} + \sigma^2_e, \tag{9}$$

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

$$\tilde{h}^2_{g,local,i} = \frac{ng(\hat{\boldsymbol{\beta}}_{gwas,i}, k_i) - (1 - \sum_{j=1,j\neq i}^m \hat{h}^2_{g,local,j})k_i}{n - k_i}, \tag{10}$$

which defines a system of linear equations involving m variables ($\hat{h}^2_{g,local,i}$) and m equations. A similar system of linear equations can be solved to obtain the variance estimator,

$$\mathrm{Var}[\hat{h}^2_{g,i}] = \left(\frac{n}{n - q_i}\right)^2 \left(2q_i\frac{\hat{\sigma}^2_e}{n} + 4\hat{h}^2_{g,local,i}\right)\frac{\hat{\sigma}^2_e}{n} + \left(\frac{q_i}{n - q_i}\right)^2 \sum_{j=1,j\neq i}^{m} \mathrm{Var}[\hat{h}^2_{g,local,j}], \tag{11}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Simulation framework

We use HAPGEN2[21] to simulate genotypes for 50,000 individuals starting from half of the 505 European (EUR) individuals in the 1000 Genomes Project[16] for SNPs with minor allele frequency (MAF) greater than 5% in one randomly selected region (1 Mb) on chromosome 1. We reserve the other half of the EUR individuals as external reference panel. From the simulated genotypes of the 50,000 individuals, we then simulate phenotypes based on the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is the standardized genotype matrix with mean 0 and variance 1 at each column. To choose, $\boldsymbol{\beta}$, the true effect size vector, we first select a subset C of $|C|$ causal SNPs at random, such that $\boldsymbol{\beta}_C \sim N(\mathbf{0}, \frac{h^2}{|C|}\mathbf{I}_{|C|})$, where h^2 is the heritability to be simulated, and $\boldsymbol{\beta}_{\bar{C}} = \mathbf{0}$. We draw $\boldsymbol{\epsilon}$ from $N(\mathbf{0}, (1-h^2)\mathbf{I}_n)$ such that $\mathrm{E}[\mathbf{y}] = 0$, $\mathrm{Var}[\mathbf{y}] = 1$, and that the SNP heritability for this locus is h^2. For fixed $\boldsymbol{\beta}$, we then generate replications of trait values $\mathbf{y}$ by varying $\boldsymbol{\epsilon}$. Finally, we compute summary statistics, $\hat{\boldsymbol{\beta}}_{gwas}$, following procedures outlined in previous sections.

Empirical data sets

We obtained publicly available GWAS summary over European ancestry data for 30 traits from the websites of 11 GWAS consortia (see Table 1). For quality control, we restricted our analysis to GWAS studies involving at least 20,000 samples. We used the definition of independent loci as defined in[22] (1.6 Mb on the average). To reduce statistical noise in LD matrix, we focused on estimating heritability attributable to common SNPs, i.e. SNPs with MAF greater than 5% in the European population. Prior to estimating heritability, we also removed SNPs with ambiguous alleles as compared to the reference panel (Supplementary Table 1) and applied our estimator as defined in Equation 11.

Most GWAS apply the genomic control factor (λ_{gc}) to the χ^2 statistics to correct for inflation due to population structure[23]. However, recent works[6,24] show that λ_{gc} can not distinguish between inflation and true polygenicity and overestimates the correction factor needed for population stratification. Although dividing the χ^2 statistics by λ_{gc} does not affect computing the ratios between local and genome-wide heritability[7], it can result in underestimation of the total heritability across the entire genome (by scaling the local estimates of

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

heritability in our framework). We notice all the summary GWAS data has at least one round of genomic correction. To account for this, we first re-estimate the λ_{gc} from summary GWAS and re-inflate the effect sizes with the estimated λ_{gc} before estimating local heritabilities. To estimate λ_{gc} we conservatively regress $\hat{\boldsymbol{\beta}}_{gwas}^T \mathbf{V}^\dagger \hat{\boldsymbol{\beta}}_{gwas}$ against $\hat{\boldsymbol{\beta}}_{gwas,gc}^T \mathbf{V}^\dagger \hat{\boldsymbol{\beta}}_{gwas,gc}$ for the bottom 30% of loci with the smallest estimated local heritability. This is based on the observation that at loci where $h^2_{g,local} = 0$, $\mathrm{E}[\hat{\boldsymbol{\beta}}_{gwas}^T \mathbf{V}^\dagger \hat{\boldsymbol{\beta}}_{gwas}] = \frac{q}{n}$ and $\mathrm{E}[\hat{\boldsymbol{\beta}}_{gwas,gc}^T \mathbf{V}^\dagger \hat{\boldsymbol{\beta}}_{gwas,gc}] = \frac{1}{\lambda_{gc}} \frac{q}{n}$, where $\hat{\boldsymbol{\beta}}_{gwas,gc} = \frac{\hat{\boldsymbol{\beta}}_{gwas}}{\sqrt{\lambda_{gc}}}$ denotes GC corrected effect size vector. We report estimated λ_{gc} for all 30 traits in Supplementary Table 1.

We define GWAS hits as SNPs with p-values less than 5×10^{-8}. For FI and HOMA-IR that do not have any GWAS hits at this threshold, we relax the threshold to 10^{-7}. 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, $\hat{h}^2_{gwas}$, is then estimated as $\sum_{i=1}^{I} \hat{\beta}_i^2$, where $\hat{\beta}_i$ is effect size of the i-th index SNP, and I the number of index SNPs. We estimate the variance of $\hat{h}^2_{gwas}$ as $\mathrm{Var}[\hat{h}^2_{gwas}] = \sum_{i=1}^{I} \mathrm{Var}[\hat{\beta}_i^2] = \sum_{i=1}^{I} \mathrm{Var}[(Z_i/\sqrt{n})^2] = \sum_{i=1}^{I} \mathrm{Var}[\frac{1}{n}\chi_i^2] = 2I/n^2$.

For case-control traits, an adjustment factor is needed to correct for ascertainment[25]. We note that this adjustment factor is derived based on the infinitesimal model, and does not apply to our method, which assumes a fixed effect model. Therefore, we only report unadjusted heritability estimates for case-control traits. However, we note that ratio between local to 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 obtain unbiased estimation whereas the approach that uses the number of SNPs to correct for bias generated by the quadratic form[15] leads to a severe under-estimation of heritability (Supplementary Figure 2). Second, we find that using the top 10-30 eigenvectors

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of the LD matrix (see Methods) provides a good approximation for the estimated heritability when LD is estimated from external reference panels (Supplementary Figure 1).

Next, we compared HESS to the recently proposed LD-score regression (LDSC)[6,7] method that provides estimates of heritability from GWAS summary data. Although LDSC is not designed for local analyses due to model assumptions on polygenicity, it is able to estimate the variance in trait attributable to any sets of SNPs. As expected, in our simulations we find that LDSC is sensitive to the underlying polygenicity and, in general, yields biased estimation of heritability. In contrast, HESS provides an unbiased estimation of heritability across all simulated disease architectures when in-sample LD is available. For example, in simulations where 20% of the variants at the locus are causal explaining 0.05% heritability, HESS yields an estimate of 0.054% (s.e. 0.004%) as compared to 0.025% (s.e. 0.0009%) for LDSC (Figure 1). We attribute this to the fact that HESS does not make any assumption on the distribution of effect sizes at causal variants by treating them as fixed effects in the model. When LD from the sample is unavailable and has to be estimated from reference panels, both methods are biased with HESS yielding results closer to simulated heritability than LDSC (Figure 1 and Supplementary Figure 3).

Finally, unlike LDSC that employs a jack-knife approach to estimate variance in the estimated heritability (thus requiring multiple loci), HESS provides a variance estimator following quadratic form theory (see Methods). We find that the variance estimator yields unbiased estimates when in-sample LD is available and under-estimates the theoretical variance when LD is estimated from reference data (Supplementary Figure 4). We attribute this to the fact that the rank of in-sample LD, which is computed based on many more samples, is in general larger than that of external reference LD.

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 Methods, Table 1, Supplementary Table 1). We estimated the local heritability at 1,703 approximately-independent loci[22] using European individuals of the 1000 Genomes to estimate LD[16]. We first investigated the

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 find a broad consistency between our estimate and the existing estimates from the literature (with differences likely due to different quality controls steps that retained slightly different SNP sets) (see Table 1). For example, HESS estimates a genome-wide SNP heritability (h_g^2) of 19.4% (s.e. 0.5%) for BMI and 39.0% (s.e. 0.4%) for height as compared to previously reported estimates of 21.6% (2.2%) for BMI[1] and 49.8% (4.4%) for height[2]. Most importantly, we find that common SNPs explain a large fraction of the previously reported narrow sense heritability for all quantitative traits we interrogated ranging from 21% for Forearm BMD to 76% for HDL(Table 1). Although we observe a very high contribution of common SNPs to case-control traits as well, we note that our estimator can be biased due to ascertainment in this case (see Methods).

Hidden heritability at known risk loci

Recent work[10,26] has shown that the total heritability explained by all variants at the GWAS risk loci ($h^2_{g,local,gwas}$) 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 find several traits for which $h^2_{g,local,gwas}$ is significantly larger than h^2_{gwas}. For example, $h^2_{g,local,gwas}$ is almost two fold higher (24.2%, s.e. 0.2%) than h^2_{gwas} (14.3%, s.e. 0.01%) for height (Table 1). The difference can be accounted by incomplete tagging of hidden causal variant(s) or allelic heterogeneity (i.e., multiple causal variants). Indeed, conditional analysis identified 36 GWAS loci that contain multiple signals of associations (for a total of 87 GWAS risk SNPs at these loci) for height[27]. Restricting to the 43 loci that contain these multiple causal loci we estimate $h^2_{g,local,gwas}$=3.7%(s.e. 0.09%), an almost 3-fold increase over h^2_{gwas}=1.2%(s.e. 0.003%). These loci, 8.7% of all GWAS loci for height, contribute to 38.1% of the difference between $h^2_{g,local,gwas}$ and h^2_{gwas} across all loci, thus suggesting that the difference is likely due to multiple signals of association. We observe similar results suggesting multiple signals of

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

associations for HDL, TG, RA, SCZ (Table 1).

In contrast, the majority of traits show similar $h^2_{g,local,gwas}$ and h^2_{gwas} (see Table 1) suggesting a single causal variant at these loci very well tagged by the index GWAS variant. For example, it is known that LDL is strongly regulated by single non-coding functional variant at the SORT1 locus[3,28] and that bone mineral density traits (FN) are strongly regulated by WNT16[29,30]. Finally, we note that reference panel LD approximation makes HESS is intrinsically conservative (see Methods) which can also lead to $h^2_{local,GWAS}$ being less than h^2_{gwas}.

Contrasting polygenicity across multiple complex traits

Most studied common traits exhibit a strong polygenic architecture (i.e. an abundance of loci of small effect contributing to trait)[1,2,3,9] . We recapitulate this observation using the HESS analysis and find a strong correlation between chromosome length and the fraction of heritability it explains for most traits we analyze here (Figures 2, 3). We also observe, consistent with previous findings[31], regions such as FTO on chromosome 16 and HLA on chromosome 6 contributing disproportionately to the fraction of heritability for HDL, BMI, and RA, respectively.

Next, we sought to quantify the variability in polygenicity across traits. We rank order loci based on their estimated local heritability, sum their contribution and plot it versus the percentage of genome they occupy (Figure 4). For highly polygenic traits, we expect the cumulative fraction of total SNP heritability to be proportional to the fraction of genome covered, whereas for less polygenic traits, we expect to see a small fraction of the genome accounting for a large fraction of total SNP heritability. For example, in schizophrenia and height the top 1% of the loci with the highest local heritability contribute to 3.9%(s.e. 0.94%) and 8.2%(s.e. 1.8%) of the total SNP heritability of these traits, respectively. 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, TG) have a lower degree of polygenicity, with the top 1% of loci accounting for 15-25% of the total SNP heritability. We note that the different degrees of polygenic signals across traits reflect both a difference in disease architecture (i.e. distribution of effect sizes) as well as a difference in the sample sizes for

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the GWAS summary data.

A different perspective of polygenicity is to restrict to GWAS risk loci (as they clearly contain risk variants) and contrast the proportion of explained variance with the proportion of the genome they occupy. We observe a wide distribution across traits reflecting diverse genetic architectures as well as different sample sizes for the GWAS performed for these traits. For example, approximately 30% of loci across the genome harbor a risk variant for height and account for 50% to the total SNP heritability (an 1.6-fold enrichment). On the other hand, while only 5% of the loci contain GWAS risk variants for HDL, these loci collectively explain 25% of the SNP heritability of HDL (a 5-fold enrichment) (Figure 5).

Loci that contribute to heritability of multiple traits

The 30 traits studied in this work share genetic correlations[6] and we recapitulate this correlation using HESS estimates (Supplementary Figure 5). Motivated by this, we searched for pleiotropic loci which we define as loci that contribute more than 0.3% of the total SNP heritability for at least 3 out of the 30 analyzed traits. In total, we identified 55 such loci distributed genome-wide (see Figure 7 Supplementary Figure 6). As expected, the HLA region (chr16:26-34M), displays strong pleiotropic signal, particularly for immunologically relevant phenotypes (see Figure 7). For instance, locus chr16:32-33M contributes more than 0.3% of total SNP heritability for over 10 traits, with exceptionally strong signals for RA, UC, and IBD (see Figure 7). We also observe several pleiotropic loci on chromosome 10 and 16, including the loci chr10:63-66M, which contributes to PLT, IBD, CD, and UC, and the loci chr16:49-52M, which contributes to CD, IBD, and FN. Previous research has revealed the association between platelet count and inflammatory bowel diseases[32] and the association between IBD traits and bone mineral density traits[33]. Our results suggest that the associations may be caused in part by the pleiotropic effect of the regions chr10:63-66M and chr16:49-52M, respectively. We note that the selection of traits can bias the identification of pleitropic loci towards over-represented traits. Nevertheless, local heritability analysis is still a useful tool to quantify the fraction of total SNP heritability contributed by a single loci and provide valuable insights into identifying pleiotropic loci.

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

We have presented HESS, an unbiased estimator of local heritability from GWAS summary data. Our approach extends existing works that treat SNP effects as fixed and model genotypes as random variables[15] – we proposed a method to regularize LD matrix estimated from external reference panel and generalized the estimator to multiple independent loci. Through extensive simulations, we demonstrated that HESS yields more consistent and less biased local heritability estimates than another recently proposed method that models SNP effects as random variables. We applied HESS on GWAS summary data of 30 traits from 11 GWAS consortia and showed that our results recapitulate previous findings. We then used the local estimate to contrast polygenicity of these traits, find loci with multiple causal variants, and identify heritability hot spots. We note that the enrichment of heritability at GWAS risk loci could be leveraged into prioritizing GWAS or fine-mapping; for example, traits with small enrichment of heritability at GWAS risk loci are more suitable for larger GWAS, whereas traits with large enrichment of heritability at known risk loci could be investigated further through fine-mapping.

Our focus in this work is estimating local heritability attributable to common SNPs (MAF $> 5\%$). Thus, we did not explore the issues of effect size and LD at rare variants. We also note that methods that adjust heritability estimation for case-control traits are currently lacking. Our reported heritability estimation for case-control traits can be biased due to ascertainment in GWAS for case-control traits. Nevertheless, our method still demonstrates its utility in studying and comparing genetic architectures of complex traits. We conjecture that future work that addresses local heritability estimation including both common and rare variants as well as adjustment of local heritability under the fixed-effect model will further improve the utility of our approach.

Acknowledgement

We are very grateful to Alkes Price, Yakir Reshef, Brielin Brown, and Nicholas Mancuso for their helpful discussions and feedback that greatly improved the quality of this

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

manuscript. The authors declare no conflict of interest.

References

[1] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538): 197–206, 2015.

[2] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[3] Global Lipids Genetics Consortium et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274–1283, 2013.

[4] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.

[5] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[6] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[7] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

statistics. *Nature genetics*, 47(11):1228–1235, 2015.

[8] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1): 76–82, 2011.

[9] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 2015.

[10] Alexander Gusev, Gaurav Bhatia, Noah Zaitlen, Bjarni J Vilhjalmsson, Dorothée Diogo, Eli A Stahl, Peter K Gregersen, Jane Worthington, Lars Klareskog, Soumya Raychaudhuri, et al. Quantifying missing heritability at known gwas loci. *Plos Genetics*, 2013.

[11] Luigi Palla and Frank Dudbridge. 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. *The American Journal of Human Genetics*, 97(2):250–259, 2015.

[12] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics*, 2(1):3–19, 1972.

[13] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[14] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, 2013.

[15] Eric R Gamazon, Nancy J Cox, and Lea K Davis. Structural architecture of snp effects on complex traits. *The American Journal of Human Genetics*, 95(5):477–489, 2014.

[16] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[17] Per Christian Hansen. The truncatedsvd as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.

[18] Richard S Elman, Nikita Karpenko, and Alexander Merkurjev. *The algebraic and geometric theory of quadratic forms*, volume 56. American Mathematical Soc., 2008.

[19] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev,

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, page btu416, 2014.

[20] Gleb Kichaev and Bogdan Pasaniuc. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *The American Journal of Human Genetics*, 97(2):260–271, 2015.

[21] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011.

[22] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, page btv546, 2015.

[23] Stephen Turner, Loren L Armstrong, Yuki Bradford, Christopher S Carlson, Dana C Crawford, Andrew T Crenshaw, Mariza Andrade, Kimberly F Doheny, Jonathan L Haines, Geoffrey Hayes, et al. Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, pages 1–19, 2011.

[24] Jian Yang, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'Connell, Massimo Mangino, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011.

[25] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[26] Nicholas Mancuso, Nadin Rohland, Kristin A Rand, Arti Tandon, Alexander Allen, Dominique Quinque, Swapan Mallick, Heng Li, Alex Stram, Xin Sheng, et al. The contribution of rare variation to prostate cancer heritability. *Nature genetics*, 2015.

[27] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

[28] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt,

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Katherine V Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M Ruda, et al. From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature*, 466 (7307):714–719, 2010.

[29] Hou-Feng Zheng, Jon H Tobias, Emma Duncan, David M Evans, Joel Eriksson, Lavinia Paternoster, Laura M Yerges-Armstrong, Terho Lehtimäki, Ulrica Bergström, Mika Kähönen, et al. Wnt16 influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. *Plos Genetics*, 2012.

[30] Daniel L Koller, Hou-Feng Zheng, David Karasik, Laura Yerges-Armstrong, Ching-Ti Liu, Fiona McGuigan, John P Kemp, Sylvie Giroux, Dongbing Lai, Howard J Edenberg, et al. Meta-analysis of genome-wide studies identifies wnt16 and esr1 snps associated with bone mineral density in premenopausal women. *Journal of Bone and Mineral Research*, 28(3):547–558, 2013.

[31] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, et al. Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907, 2015.

[32] Silvio Danese, Carol de la Motte, and Claudio Fiocchi. Platelets in inflammatory bowel disease: clinical, pathogenic, and therapeutic implications. *The American journal of gastroenterology*, 99(5):938–945, 2004.

[33] Laura E Targownik, Charles N Bernstein, and William D Leslie. Inflammatory bowel disease and the risk of osteoporosis and fracture. *Maturitas*, 76(4):315–319, 2013.

[34] Gibran Hemani, Jian Yang, Anna Vinkhuyzen, Joseph E Powell, Gonneke Willemsen, Jouke-Jan Hottenga, Abdel Abdellaoui, Massimo Mangino, Ana M Valdes, Sarah E Medland, et al. Inference of the genetic architecture underlying bmi and height with the use of 20,240 sibling pairs. *The American Journal of Human Genetics*, 93(5):865–875, 2013.

[35] Pim van der Harst, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S Paul, Ulrich Elling, Hooman Allayee, Xinzhong Li, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429):369–375, 2012.

[36] C Garner, T Tatu, JE Reittie, T Littlewood, J Darley, S Cervino, M Farrall, P Kelly,

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TD Spector, and SL Thein. Genetic influences on f cells and other hematologic variables: a twin heritability study. *Blood*, 95(1):342–346, 2000.

[37] Jing-Ping Lin, Christopher J O'Donnell, Li Jin, Caroline Fox, Qiong Yang, and L Adrienne Cupples. Evidence for linkage of red blood cell size and count: Genome-wide scans in the framingham heart study. *American journal of hematology*, 82(7):605–610, 2007.

[38] Jesse D Hinckley, Diana Abbott, Trudy L Burns, Meadow Heiman, Amy D Shapiro, Kai Wang, and Jorge Di Paola. Quantitative trait locus linkage analysis in a large amish pedigree identifies novel candidate loci for erythrocyte traits. *Molecular genetics & genomic medicine*, 1(3):131–141, 2013.

[39] Christian Gieger, Aparna Radhakrishnan, Ana Cvejic, Weihong Tang, Eleonora Porcu, Giorgio Pistis, Jovana Serbanovic-Canic, Ulrich Elling, Alison H Goodall, Yann Labrune, et al. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, 2011.

[40] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, Anna L Gloyn, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–116, 2010.

[41] Annemarie Simonis-Bik, Elisabeth MW Eekhoff, Michaela Diamant, Dorret I Boomsma, Rob J Heine, Jacqueline M Dekker, Gonneke Willemsen, Marieke Van Leeuwen, and Eco JC De Geus. The heritability of hba1c and fasting blood glucose in different measurement settings. *Twin Research and Human Genetics*, 11(06):597–602, 2008.

[42] LJ Rasmussen-Torvik, JS Pankow, DR Jacobs, LM Steffen, AM Moran, J Steinberger, and AR Sinaiko. Heritability and genetic correlations of insulin sensitivity measured by the euglycaemic clamp. *Diabetic Medicine*, 24(11):1286–1289, 2007.

[43] Nicole Soranzo, Serena Sanna, Eleanor Wheeler, Christian Gieger, Dörte Radke, Josée Dupuis, Nabila Bouatia-Naji, Claudia Langenberg, Inga Prokopenko, Elliot Stolerman, et al. Common variants at 10 genomic loci influence hemoglobin a1c levels via glycemic and nonglycemic pathways. *Diabetes*, 59(12):3229–3239, 2010.

[44] GW Mills, PJ Avery, MI McCarthy, AT Hattersley, JC Levy, GA Hitman, M Sampson, and M Walker. Heritability estimates for beta cell function and features of the insulin

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

resistance syndrome in uk families with an increased susceptibility to type 2 diabetes. *Diabetologia*, 47(4):732–738, 2004.

[45] Noah Zaitlen, Bogdan Pasaniuc, Sriram Sankararaman, Gaurav Bhatia, Jianqi Zhang, Alexander Gusev, Taylor Young, Arti Tandon, Samuela Pollack, Bjarni J Vilhjálmsson, et al. Leveraging population admixture to characterize the heritability of complex traits. *Nature genetics*, 46(12):1356–1362, 2014.

[46] SV de Miranda Chagas, S Kanaan, H Chung Kang, M Cagy, RE de Abreu, LA da Silva, RC Garcia, and ML Garcia Rosa. 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(6):329–337, 2011.

[47] SC Elbein and SJ Hasstedt. Quantitative trait linkage analysis of lipid-related traits in familial type 2 diabetes evidence for linkage of triglyceride levels to chromosome 19q. *Diabetes*, 51(2):528–535, 2002.

[48] Cornelius A Rietveld, Sarah E Medland, Jaime Derringer, Jian Yang, Tonu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, Arpana Agrawal, et al. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *science*, 340(6139):1467–1471, 2013.

[49] Hou-Feng Zheng, Vincenzo Forgetta, Yi-Hsiang Hsu, Karol Estrada, Alberto Rosello-Diez, Paul J Leo, Chitra L Dahia, Kyung Hyun Park-Min, Jonathan H Tobias, Charles Kooperberg, et al. Whole [hyphen] genome sequencing identifies en1 as a determinant of bone density and fracture. *Nature*, 526(7571):112–117, 2015.

[50] NK Arden, J Baker, C Hogg, K Baan, and TD Spector. The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. *Journal of Bone and Mineral Research*, 11(4):530–534, 1996.

[51] John RB Perry, Felix Day, Cathy E Elks, Patrick Sulem, Deborah J Thompson, Teresa Ferreira, Chunyan He, Daniel I Chasman, Tõnu Esko, Gudmar Thorleifsson, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514(7520):92–97, 2014.

[52] B Towne. Heritability of age at menarche in girls from the fels longitudinal study.

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

American journal of physical anthropology, 128(1):210, 2005.

[53] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

[54] SMJ Harney, C Vilariño-Güell, IE Adamopoulos, A-M Sims, RW Lawrence, LR Cardon, JL Newton, C Meisel, JJ Pointon, C Darke, et al. Fine mapping of the mhc class iii region demonstrates association of aif1 and rheumatoid arthritis. *Rheumatology*, 47(12): 1761–1767, 2008.

[55] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[56] Patrick F Sullivan, Kenneth S Kendler, and Michael C Neale. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of general psychiatry*, 60(12):1187–1192, 2003.

[57] Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986, 2015.

[58] C Tysk, E Lindberg, G Järnerot, and B Floderus-Myrhed. 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(7):990–996, 1988.

[59] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9): 981, 2012.

[60] P Poulsen, K Ohm Kyvik, A Vaag, and H Beck-Nielsen. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145, 1999.

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Trait	h^2_g	h^2_{pub}	h^2_g/h^2_{pub}	h^2_{gwas}	$h^2_{g,local,gwas}$	$h^{2*}_{g,local,gwas}$	**Enrichment**[a]
BMI (Body Mass Index)[1]	19.4(0.5)	42[34]	0.46	2.2(0.01)	3.6(0.1)	3.5(0.1)	3.4(0.5)
Height (Height)[2]	39.0(0.4)	69[34]	0.57	13.9(2.32)	23.5(0.2)	23.2(0.2)	1.9(0.1)
HB (Haemoglobin)[35]	20.9(2.5)	37[36]	0.56	2.6(0.06)	2.1(0.3)	2.1(0.3)	7.2(1.4)
MCH (Mean Cell Haemoglobin)[35]	25.5(3.1)	52[37]	0.49	8.0(0.92)	5.3(0.5)	5.2(0.5)	9.1(2.1)
MCHC (MCH Concentration)[35]	13.7(3.1)	48[38]	0.29	0.5(0.00)	0.5(0.3)	0.4(0.3)	3.6(1.1)
MCV (Mean Cell Volume)[35]	23.3(2.7)	52[37]	0.44	7.2(0.79)	5.0(0.5)	4.9(0.5)	6.9(1.3)
PCV (Packed Cell Volume)[35]	20.8(3.1)	30[36]	0.69	1.6(0.02)	1.1(0.3)	1.1(0.3)	5.6(1.1)
RBC (Red Blood Cell Count)[35]	23.6(2.9)	56[37]	0.42	4.1(0.22)	2.6(0.4)	2.5(0.4)	5.0(1.4)
PLT (Number of Platelets)[39]	23.2(1.9)	57[36]	0.41	4.1(0.16)	3.5(0.3)	3.5(0.3)	5.9(1.0)
FG (Fasting Glucose)[40]	20.0(3.0)	66[41]	0.30	2.9(0.07)	1.9(0.3)	1.9(0.3)	9.6(3.6)
FI (Fasting Insulin)[40]	19.5(3.1)	36[42]	0.54	0.1(0.00)	0.1(0.1)	0.1(0.1)	11.5(0.0)
HBA1C (HBA1C)[43]	20.0(3.1)	75[41]	0.27	2.1(0.03)	0.8(0.3)	0.8(0.3)	6.8(1.4)
HOMA-B (HOMA-B)[40]	19.1(3.1)	72[44]	0.27	0.7(0.00)	0.5(0.2)	0.5(0.2)	10.3(3.5)
HOMA-IR (HOMA-IR)[40]	20.3(3.0)	38[42]	0.53	0.1(0.00)	0.1(0.1)	0.1(0.1)	12.3(0.0)
HDL (High Density Lipoprotein)[3]	32.0(1.1)	42[45]	0.76	6.4(0.27)	11.6(0.3)	11.3(0.3)	5.6(1.7)
LDL (Low Density Lipoprotein)[3]	26.5(1.2)	40[45]	0.66	8.8(0.49)	8.3(0.3)	8.1(0.3)	6.3(1.1)
TC (Total Cholesterol)[3]	25.8(1.2)	50[46]	0.52	9.0(0.52)	9.3(0.3)	9.1(0.3)	5.9(0.8)
TG (Triglycerides)[3]	28.8(1.2)	40[47]	0.72	9.3(0.52)	9.2(0.3)	9.1(0.3)	8.0(2.1)
EY (Education Years)[48]	15.3(1.0)	40[48]	0.38	0.1(0.00)	0.1(0.1)	0.1(0.1)	3.1(0.9)
COL (College)[48]	15.5(1.0)	40[48]	0.39	0.1(0.00)	0.1(0.1)	0.1(0.1)	4.2(1.0)
FA (Forearm BMD)[49]	17.4(3.8)	84[50]	0.21	0.4(0.00)	0.4(0.1)	0.4(0.1)	12.8(4.6)
FN (Femoral Neck BMD)[49]	28.4(3.3)	84[50]	0.34	2.2(0.05)	2.1(0.4)	2.1(0.4)	6.0(0.7)
LS (Lumbar Spine)[49]	30.1(3.2)	84[50]	0.36	2.3(0.06)	2.5(0.4)	2.5(0.4)	5.9(0.7)
AM (Age at Menarche)[51]	19.6(0.9)	49[52]	0.40	2.7(0.06)	3.2(0.2)	3.2(0.2)	3.5(0.2)
RA (Rheumatoid Arthritis)[53]	64.9(1.4)	55[54]	1.18	12.2(2.12)	21.7(0.3)	21.6(0.3)	10.9(4.9)
SCZ (Schizophrenia)[55]	64.6(1.0)	81[56]	0.80	6.2(0.74)	8.9(0.3)	8.8(0.3)	2.0(0.1)
CD (Crohn's Disease)[57]	29.4(1.6)	53[58]	0.55	3.9(0.17)	4.6(0.3)	4.5(0.3)	4.4(0.6)
IBD[b] (Inflammatory Bowel Disease)[57]	33.5(1.5)	–	–	5.0(0.34)	5.9(0.3)	5.8(0.3)	4.2(0.4)
UC (Ulcerative Colitis)[57]	26.2(1.7)	58[58]	0.45	2.8(0.08)	3.0(0.3)	3.0(0.3)	4.7(0.7)
T2D (Type 2 Diabetes)[59]	22.2(2.0)	26[60]	0.85	1.4(0.01)	1.4(0.2)	1.3(0.2)	6.0(1.1)

Table 1: Total SNP heritability estimates and the amount of h^2_g attributable to loci containing GWAS index SNPs ($h^2_{g,local,gwas}$) and index SNPs only (h^2_{gwas}). $h^{2*}_{g,local,gwas}$ is the same as $h^2_{g,local,gwas}$ except that GWAS index SNPs are excluded in the computation. In supplementary, we report $h^{2\dagger}_{g,local,gwas}$, obtained by excluding all GWAS hits. We also report narrow sense heritability (h^2_{pub}) estimates obtained from twin or family studies. We list case-control traits where our estimate of h^2_g is biased due to ascertainment at the bottom of the table. [a]Similar to[7], we define enrichment as the ratio between the fraction of h^2_g attributable to $h^{2*}_{g,local,gwas}$ and the fraction of genome covered by these loci. We obtain standard errors by jackknife over the loci. [b]IBD refers to the union of CD and UC.

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

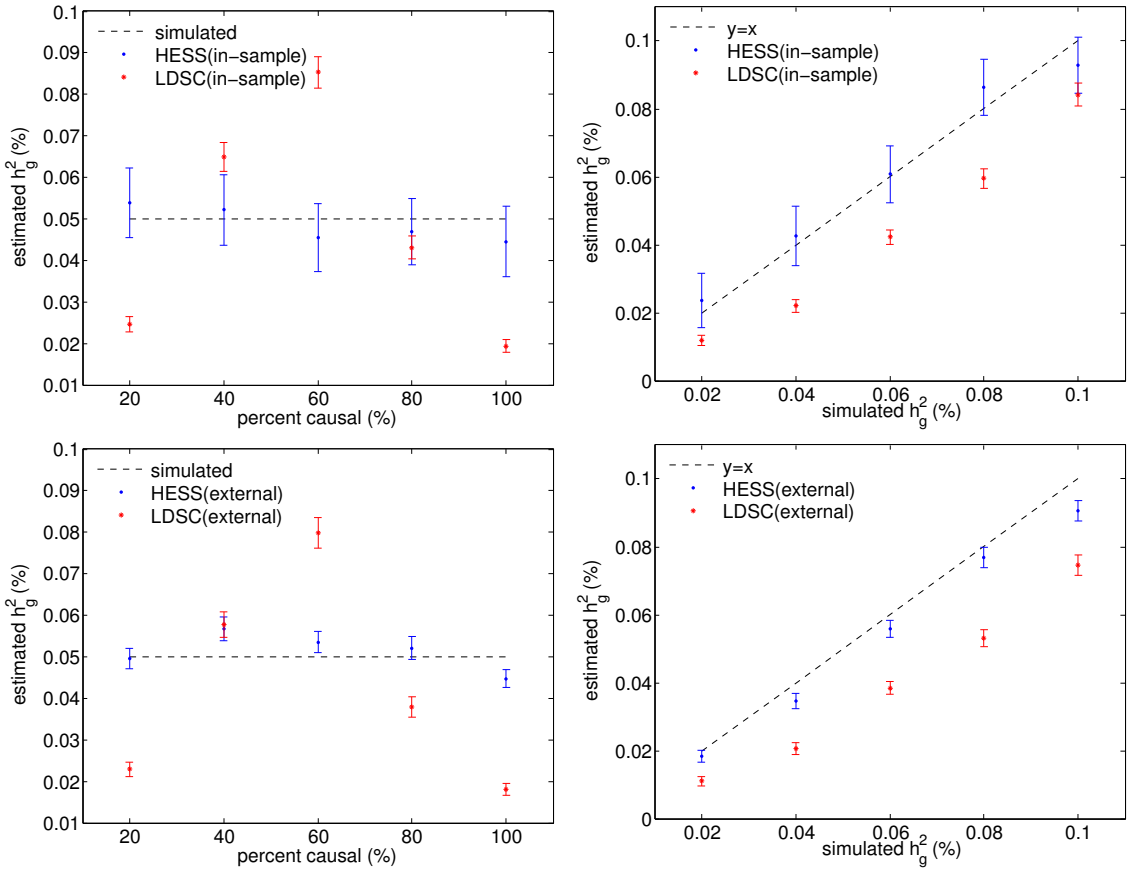

Figure 1: HESS provides superior accuracy over LDSC in estimating local heritability. HESS attains unbiased estimates when in-sample LD is used (top) and approximately unbiased estimates when reference LD is used (bottom). Mean and standard errors in these figures are computed based on 500 simulations, each involving 50,000 simulated GWAS data sets.

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

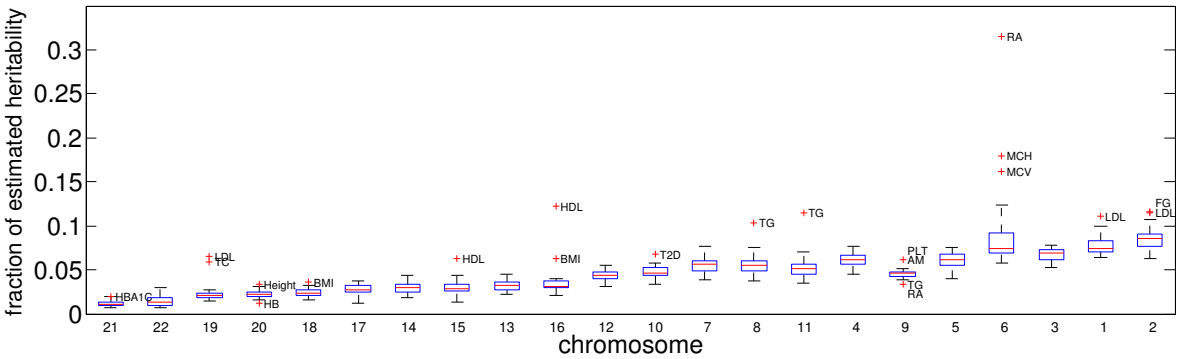

Figure 2: Fraction of h_g^2 per chromosome across the 30 traits studied. Here, the chromosomal heritability is obtained by summing local heritability at loci within the chromosome. For each chromosome we plot the box plots of estimates at the 30 considered traits. Chromosomes are ordered by size. With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

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

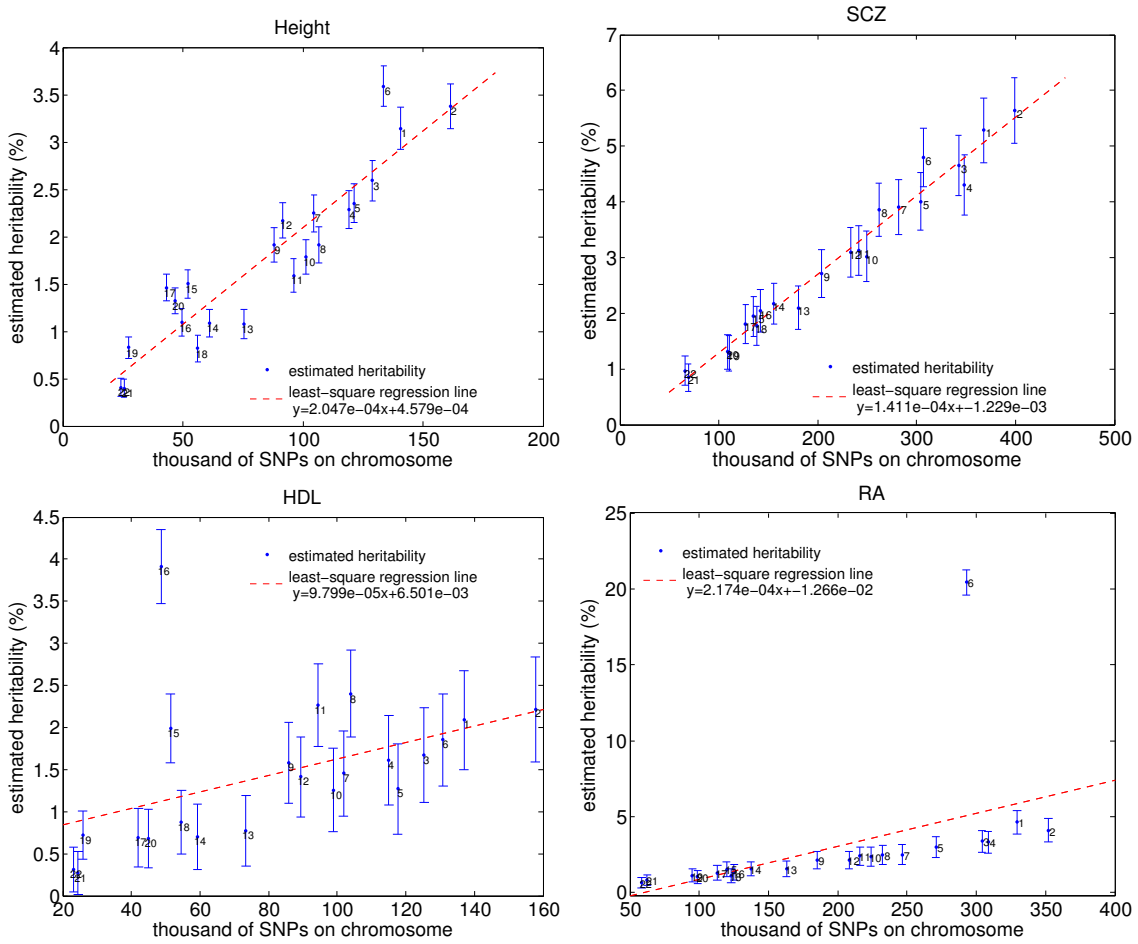

Figure 3: Heritability attributable to each chromosome for four example traits. The chromosomal heritability is obtained by summing local heritability at loci within the chromosome. Standard error is obtained by taking the square root of the sum of variance estimation. See Supplementary Note for results across all traits.

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

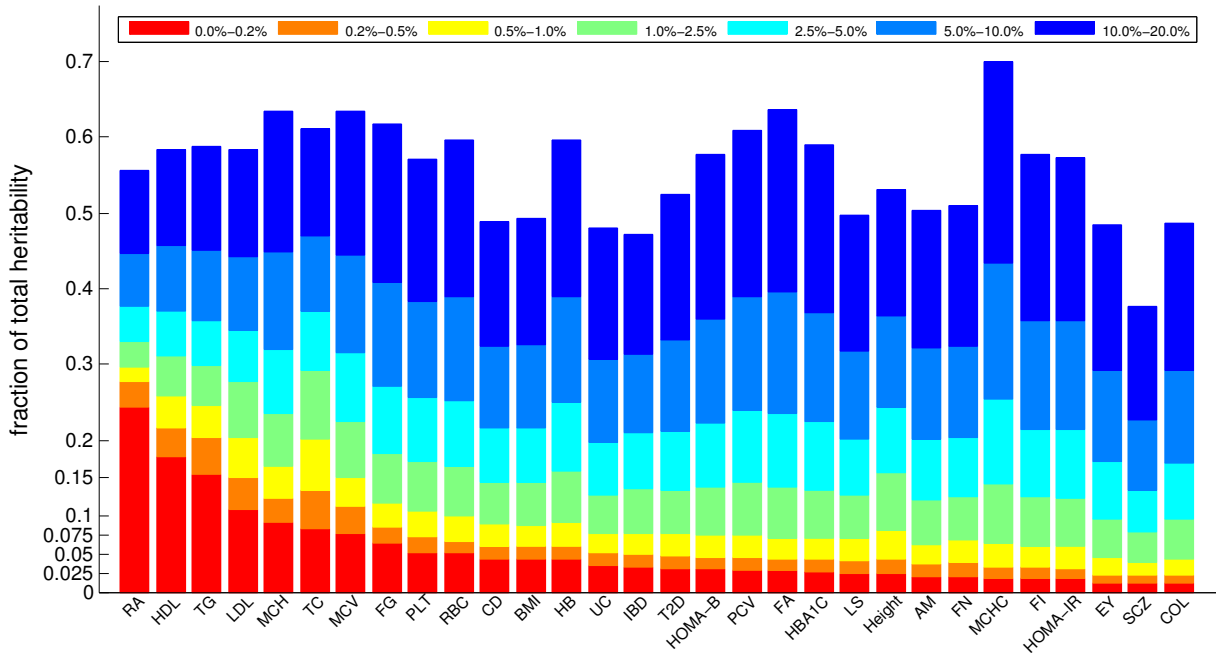

Figure 4: Stacked bar plot showing the percentage of total heritability attributable to different fractions of genome. We rank ordered all genomic loci by their explained heritability and quantified the fraction of total heritability attributable to different percentile ranges. Traits with high polygenicity tend to have bars with height proportional to bin size (e.g. Height and SCZ), whereas less polygenic traits tend to have bars much larger than bin size (e.g. RA and HDL).

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

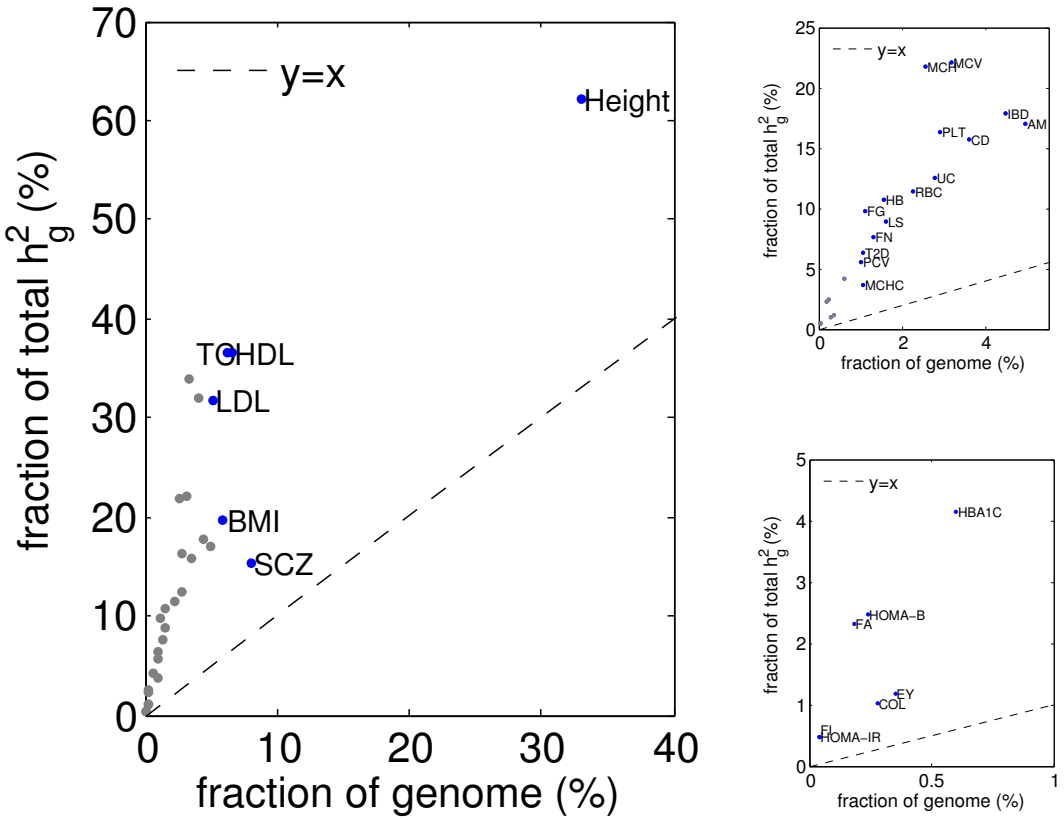

Figure 5: Fraction of h_g^2 explained by all loci that contain a GWAS hit versus the fraction of genome covered by these loci. Images on the right focus successively on the traits near the bottom left.

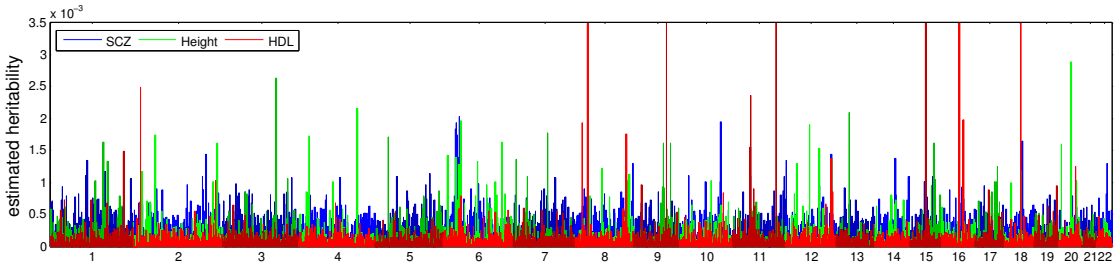

Figure 6: Manhattan-style plots of regional heritability across the genome for the trait Height, HDL, and SCZ. See Supplementary Note for results across all traits.

bioRxiv preprint doi: https://doi.org/10.1101/035907; this version posted January 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

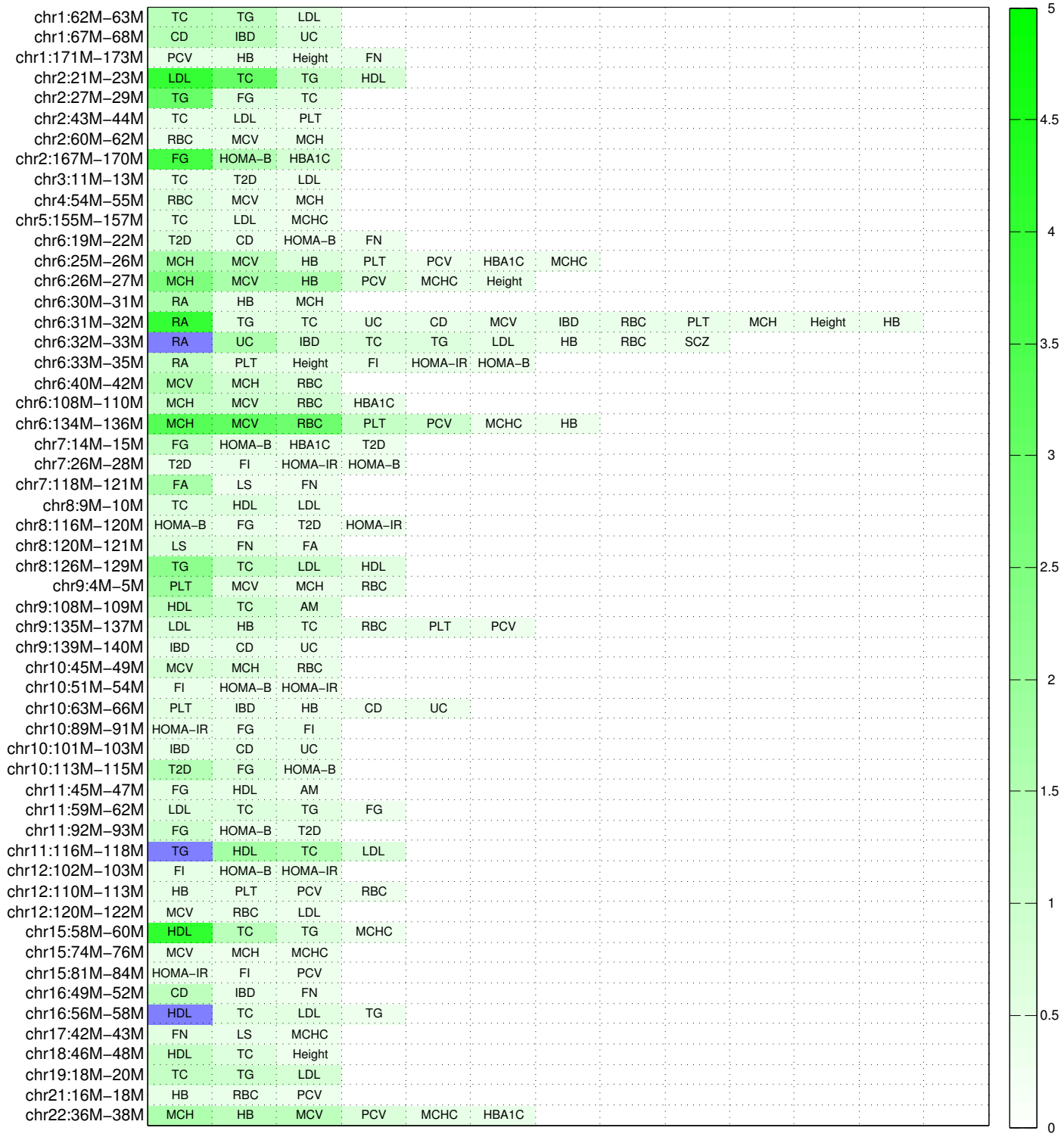

Figure 7: Heat map showing the fraction of total SNP heritability (h_g^2) contributed by each of the 55 "pleiotropic" loci. For each locus, we only display the traits to which the locus contributes more than 0.3% of total SNP heritability. We mark traits to which the locus contributes more than 5% of the total SNP heritability in dark blue.