GWIS: Genome Wide Inferred Statistics for non-linear functions of multiple phenotypes

H. A. Nieuwboer, R. Pool, C. V. Dolan, D. I. Boomsma & M. G. Nivard

Corresponding author

M. G. Nivard m.g.nivard@vu.nl

Here we present a method of genome wide inferred study (GWIS) that provides an approximation of genome wide association study (GWAS) summary statistics for a variable that is a function of phenotypes for which GWAS summary statistics, phenotypic means and covariances are available. GWIS can be performed regardless of sample overlap between the GWAS of the phenotypes on which the function depends. As GWIS provides association estimates and their standard errors for each SNP, GWIS can form the basis for polygenic risk scoring, LD score regression[1], Mendelian randomization studies, biological annotation and other analyses. Here, we replicate a body mass index (BMI) GWAS using GWIS based on a height GWAS and a weight GWAS. We proceed to use a GWIS to further our understanding of the genetic architecture of schizophrenia and bipolar disorder.

An example of a well known variable that is a (non-linear) function of multiple phenotypes is BMI, which is defined as weight over height squared. We demonstrate the accuracy of GWIS by reconstructing a body mass index (BMI) GWAS based on publicly available height and weight GWAS summary statistics[2] (see **URLs**). For each single nucleotide polymorphism (SNP) included in the height and weight GWAS with a minor allele frequency (MAF) larger than 0.05 (as obtained from the HAPMAP Consortium[3], see **URLs**), we infer estimates and standard errors of these estimates for the association between the SNP and BMI. In a GWAS, BMI must be ascertained for all participants, whereas in a GWIS, we rely on population parameters which reflect the genetic effects on height and weight. Furthermore, the original GWAS for height and weight do not have to be performed in a common set of individuals.

Based on the summary statistics of GWAS for standardized male height and weight[2], our GWIS replicated 310 out of 356 genome wide hits (an 87.1% replication rate), and found three false positive results (see **Supplementary Table 1**), when compared to a true BMI GWAS performed in the same sample. To demonstrate the method when the constituent phenotypes (i.e., weight and height) are measured independently, we substituted the male height GWAS results for the female height results. Here we assumed the male and female genetic architecture for height in males and females are identical[4]. The GWIS based on independent samples replicated 135 out of 356 genome wide significant signals (a 37.9% replication rate) and yielded eight false positive associations (see **Supplementary Table 2**). All false positives that arise in the female GWIS occured for SNPs which were measured in a small subset of participants (N = 1666, where the total sample included up to 73137 women). The Manhattan plots in **Figure 1** revealed that even though there is a loss of power, both forms of GWIS and the original BMI GWAS implicate associations in the same genomic regions.

Using LD score regression[1], we computed the genetic correlations between BMI based on the GWAS summary statistics, the GWIS using male height data and the GWIS using female height data. As LD score regression requires information on the number of participants available per SNP, we assume the sample size for the BMI GWIS to be the lowest per-SNP sample size of either the height or weight GWAS used. As expected, the genetic correlation between BMI as measured in GWAS, BMI as approximated in GWIS using male height data and BMI as approximated in GWIS using female height data is close to unity (see **Table 1**). Next, we estimated genetic correlations between BMI based on the GWAS, BMI based on GWIS using male height data, BMI based on

GWIS using female height data and educational attainment[5], LDL cholesterol[6], age at menarche[7], rheumatoid arthritis[8] and coronary artery disease[9]. Inference made on the genetic correlates of BMI based on GWIS closely mirror the inference made based on BMI GWAS summary statistics.

[Table 1]

Ruderfer et al.[10] performed GWA studies of bipolar disorder (BIP), schizophrenia (SCZ), the pooled bipolar and schizophrenia cases versus the pooled controls (BIP + SCZ) and a GWAS in which the bipolar cases featured as controls and the schizophrenia cases as cases (SCZ - BIP) (see **URLs**). The latter two studies can be reproduced with a GWIS. However, the primary interest of these studies is identifying overlap and contrast between SCZ and BIP. SCZ and BIP are two psychiatric disorders with a substantially correlated genetic underlying liabilities[11]. This correlation prohibits the investigation of genetic variants that are specifically linked to either SCZ or BIP, as well as the investigation of genetic overlap between tertiary traits and SCZ or BIP. As a more exotic application of GWIS, we determine whether the genetic correlation between SCZ or BIP and a tertiary trait is specific to either SCZ or BIP. To this end, we defined a function that decomposes the genetic SCZ liability into a part shared with the genetic liability of BIP and a residual, referred to as unique genetic SCZ liability (unique SCZ). In a similar manner, we defined a function that decomposes the genetic BIP liability into a part shared with the genetic liability of SCZ and a

residual, referred to as unique genetic BIP liability (unique BIP). These functions are given by

$$\text{Unique SCZ} := (1+c)\text{SCZ} - (1-c)\text{BIP}$$

$$\text{Unique BIP} := (1+d)\text{BIP} - (1-d)\text{SCZ}$$

where

$$c = \frac{h^2_{\text{BIP}} - \text{Coh}(\text{BIP}, \text{SCZ})}{h^2_{\text{BIP}} + \text{Coh}(\text{BIP}, \text{SCZ})} \qquad d = \frac{h^2_{\text{SCZ}} - \text{Coh}(\text{BIP}, \text{SCZ})}{h^2_{\text{SCZ}} + \text{Coh}(\text{BIP}, \text{SCZ})}.$$

Here, $\text{Coh}(\text{BIP}, \text{SCZ})$ denotes the coheritability between BIP and SCZ (i.e., $h_{\text{SCZ}} \cdot r_{\text{BIP,SCZ}} \cdot h_{\text{BIP}}$ with $r_{\text{BIP,SCZ}}$ the latent phenotypical correlation between bipolar disorder and schizophrenia) and $h^2_{\text{BIP}}, h^2_{\text{SCZ}}$ denote the heritabilities of BIP and SCZ, respectively. For the derivation of c and d, see the online methods. Note that we do not measure unique SCZ or unique BIP in individuals. Furthermore, note that the functions themselves depend on estimated heritability and coheritabilities, which leads to less accurate estimates of genetic effects on unique SCZ and unique BIP. As effect sizes for SCZ and BIP are reported in terms of odds ratios, we take their logarithms to obtain effect sizes on the liabilities.

We performed a GWIS of unique SCZ and a GWIS of unique BIP. For our analysis of unique SCZ and unique BIP in a GWIS, we include SNPs with information values between 0.9 and 1.1 as reported by Ruderfer et al., and minor allele frequencies larger than 0.05 (as obtained from the HAPMAP Consortium[3]), both inclusion criteria reflect common practice in GWA studies[12]. LD score regression[1] was used to estimate genetic correlations between unique SCZ, unique BIP and educational attainment. We validated the absence of the genetic correlations between unique BIP and SCZ, and unique SCZ and BIP, by applying LD score regression (**Table 2**). Further

investigation revealed that unique SCZ does not genetically correlate with educational attainment, whereas unique BIP genetically correlated with educational attainment. This suggests that the observed genetic correlation between schizophrenia liability and educational attainment is fully explained by its genetic correlation with bipolar disorder liability.

[Table 2]

As shown above, GWIS can yield significant novel insight in variables that can be expressed as (non-linear) function of phenotypes. A GWIS can, for example, be performed for equations describing the steady state kinetics of (bio-)chemical reactions involving metabolites of which the concentrations have been analyzed in GWAS, or for equations describing (active) membrane transport of proteins or metabolites given that GWAS summary statistics are available for their concentrations on both sides of the barrier. Other applications of GWIS include converting a GWAS performed for a phenotype as expressed on a logarithmic scale, to a GWIS of the phenotype on its original scale. This transformation can be of use if polygenic risk scoring on the original scale is preferred.

Successful application of GWIS depends on the availability of sufficiently accurate GWAS summary statistics, the number of phenotypes involved in the function, as well as the degree of approximation. The accuracy of the summary statistics of each of the individual GWAS affects the accuracy of the GWIS results. Furthermore, the error of the GWIS statistics increase as more phenotypes are included, due to accumulation of the error in the GWAS results of each of these

phenotypes. The degree of approximation used also affects the GWIS results, as the quadratic approximation of a function generally fits better than a linear approximation (see **Supplemental Note 1** for a quadratic approximation of BMI). As the sample sizes used in GWA studies increases, GWIS becomes applicable to a broader domain of functions and yields more accurate results. Related to this last observation, all false positive associations found in the BMI GWIS based on female height data were attributable to a limited sample size for these particular SNPs. We recommend removing SNPs with low allele frequencies, poor imputation quality and SNPs available for a limited number of participants in the original GWA studies before performing GWIS.

With these points of care in mind, however, our method provides a means of obtaining the GWAS summary statistics of a variable that is a function of phenotypes when GWAS summary statistics for these phenotypes are available in (not necessarily overlapping) samples, as outlined in **Figure 2**. This remains possible even when this variable is difficult or impossible to measure in individual participants.

Online methods

Let $V = f(P_1, \ldots, P_k)$ be a function of the k phenotypes $P_1, \ldots, P_k$. Furthermore, let $S \sim \text{bin}(n = 2, p = \text{effect allele frequency (EAF)})$ be a binomially distributed variable corresponding to the number of effect alleles (EA) of a biallelic SNP. Let N denote the sample size. We assume we have a multivariate linear regression model

$$\begin{bmatrix} P_{11} & P_{21} & \ldots & P_{k1} \\ P_{12} & P_{22} & \ldots & P_{k2} \\ \vdots & \vdots & & \vdots \\ P_{1N} & P_{2N} & \ldots & P_{kN} \end{bmatrix} = \begin{bmatrix} 1 & S_1 \\ 1 & S_2 \\ \vdots & \vdots \\ 1 & S_N \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \ldots & \beta_{0k} \\ \beta_{11} & \beta_{12} & \ldots & \beta_{1k} \end{bmatrix} + \epsilon \tag{1}$$

which we write as

$$P = S\beta + \epsilon. \tag{2}$$

P is a $N \times k$ matrix, S is a $N \times 2$ matrix, β is a $2 \times k$ matrix and ϵ is a $N \times k$ matrix. We assume ϵ is a matrix where the columns are normally distributed with zero mean and covariance matrix Σ. Only an estimate for the matrix β called $\widehat{\beta}$ is known, along with the standard errors of each of the $\widehat{\beta_{1j}}$, the covariance matrix between the phenotypes $P_1, \ldots, P_k$ and the mean of each phenotype. This is equivalent to having the summary statistics of the GWA studies of each of the k phenotypes and their phenotypic covariances.

The goal is to estimate λ_0, λ_1 in

$$f(P_{1i}, \ldots, P_{ki}) =: V_i = \lambda_0 + \lambda_1 S_i + e \tag{3}$$

with e normally distributed with zero mean. This is equivalent to performing a GWAS of V. To do

this, we use a first-order Taylor approximation of V around the point

$$\mathcal{E}(s) := (\mathbb{E}[P_{1i}|S_i = s], \ldots, \mathbb{E}[P_{ki}|S_i = s]).$$

The point $\mathcal{E}(s)$ corresponds to the mean of the phenotypes of the individuals that have s effect alleles on this SNP. The first-order Taylor approximation is of the form

$$L_i := f(\mathcal{E}(s)) + \sum_{l=1}^{k} \frac{\partial f(\mathcal{E}(s))}{\partial P_l}(P_{li} - \mathbb{E}[P_{li}|S_i = s])$$

where $\partial f(\mathcal{E}(s))/\partial P_l$ denotes the partial derivative of f with respect to P_l, evaluated in the point $\mathcal{E}(s)$. Then, it follows that

$$\mathbb{E}[L_i|S_i = s] = \mathbb{E}[f(\mathcal{E}(s))|S_i = s] = f(\mathcal{E}(s)) \tag{4}$$

since for each l in $1, \ldots, k$,

$$\mathbb{E}\big[P_{li} - \mathbb{E}[P_{li}|S_i = s]\,|S_i = s\big] = 0$$

by the linearity of the expectation operator. Equation (4) shows that the mean of the linear approximation is equal to the function evaluated in the phenotypic mean of individuals that have s effect alleles. The error incurred in the linearization process takes the form

$$\frac{1}{2}\sum_{j=1}^{k}\sum_{l=1}^{k} \frac{\partial^2 f(\widetilde{\mathcal{E}})}{\partial P_j \partial P_l}(P_{ji} - \mathbb{E}[P_{ji}|S_i = s])(P_{li} - \mathbb{E}[P_{li}|S_i = s])$$

for some $\widetilde{\mathcal{E}}$ inbetween the two points $(P_{1i}, \ldots, P_{ki})$ and $\mathcal{E}(s)$.

Note that the linearization is only possible if f satisfies certain regularity conditions on the relevant space of phenotype values. For example, division by 0 is not allowed. This can be avoided by linearly transforming the observed phenotypes, along with their associated parts of the β-matrix.

We now attempt to derive a linear model for our approximate expression for $\mathbb{E}[V_i|S_i = s]$. We write

$$\mathbb{E}[L_i|S_i = s] = \lambda_0 + \lambda_1 s$$

and note that if s is 0, we have a direct approximation for λ_0:

$$\widehat{\lambda_0} = \mathbb{E}[L_i|S_i = 0]\,.$$

However, as we have shown, $\mathbb{E}[L_i|S_i = s] = f(\mathcal{E}(s))$, so our approximation for λ_0 becomes

$$\begin{aligned}\widehat{\lambda_0} &= f(\mathcal{E}(0)) \\ &= f(\beta_{01}, \beta_{02}, \ldots, \beta_{0k})\end{aligned}$$

i.e., the function f evaluated at the intercepts of our linear regression model. We can also estimate $\lambda_1 = (\mathbb{E}[L_i|S_i = s] - \lambda_0)/s$ by evaluating this expression for $s = 1, 2$ and weighing the results by their (estimated) relative population frequencies. The expression for $\widehat{\lambda_1}$ is given by

$$\begin{aligned}\widehat{\lambda_1} =& \frac{2\mathrm{EAF}(1-\mathrm{EAF})}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}(f(\mathcal{E}(1)) - \widehat{\lambda_0}) + \frac{\mathrm{EAF}^2}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}\frac{f(\mathcal{E}(2)) - \widehat{\lambda_0}}{2} \\ =& \frac{2\mathrm{EAF}(1-\mathrm{EAF})}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}(f(\beta_{01} + \beta_{11}, \beta_{02} + \beta_{12}, \ldots, \beta_{0k} + \beta_{1k}) - \widehat{\lambda_0}) \\ &+ \frac{\mathrm{EAF}^2}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}\frac{f(\beta_{01} + 2\beta_{11}, \beta_{02} + 2\beta_{12}, \ldots, \beta_{0k} + 2\beta_{1k}) - \widehat{\lambda_0}}{2}.\end{aligned}$$

To test our estimates for λ_0 and λ_1, their standard errors must be obtained. However, since we do not have the covariance matrix of $\hat{\beta}$, we must first estimate the covariance between each of the $\widehat{\beta_{ij}}$. With the theory of multivariate linear regression, we know that the least squares solution to the model $P = S\beta + \epsilon$ is given by

$$\hat{\beta} = (S^T S)^{-1} S^T P$$

with corresponding variance-covariance matrix

$$\mathrm{Var}(\hat{\beta}) = (S^T S)^{-1} \otimes \Sigma \tag{5}$$

assuming that columns of ϵ have zero mean and the rows of ϵ are pairwise uncorrelated[13]. This is under the assumption of complete sample overlap. The matrix Σ is a $k \times k$ matrix with the elements $\Sigma_{jl} = \mathrm{Cov}(\epsilon_j, \epsilon_l)$, the covariance between the errors in the linear regressions of the phenotypes P_j and P_l on S. We assume that the effect of each of the individual SNPs is small, so $\mathrm{Var}(\epsilon_j) \approx \mathrm{Var}(P_j)$ and $\mathrm{Cov}(\epsilon_j, \epsilon_l) \approx \mathrm{Cov}(P_j, P_l)$. Expanding $S^T S$ gives

$$S^T S = \begin{bmatrix} N & \sum S_i \\ \sum S_i & \sum (S_i^2) \end{bmatrix}$$

with inverse

$$(S^T S)^{-1} = \frac{1}{N \sum (S_i^2) - (\sum S_i)^2} \begin{bmatrix} \sum (S_i)^2 & -\sum S_i \\ -\sum S_i & N \end{bmatrix}$$

From this, we can infer

$$\begin{aligned} \mathrm{Cov}(\widehat{\beta_{1j}}, \widehat{\beta_{1l}}) &= \frac{N \mathrm{Cov}(P_j, P_l)}{N \sum (S_i^2) - (\sum S_i)^2} \\ &= \frac{\mathrm{Cov}(P_j, P_l)}{N \, \mathrm{Var}\, S_i} \\ &= \mathrm{SE_j} \cdot \mathrm{Cor}(P_j, P_l) \cdot \mathrm{SE_l}. \end{aligned}$$

In case there is only partial sample overlap, $\mathrm{Cov}(\widehat{\beta_{1j}}, \widehat{\beta_{1l}})$ may also be approximated as

$$\mathrm{SE}_j \cdot \mathrm{Cor}(P_j, P_l) \frac{N_{\cap j,l}}{\sqrt{N_j N_l}} \cdot \mathrm{SE}_l. \tag{6}$$

Here, $N_{\cap j,l}$ is the number of individuals that is present in both the GWAS of P_j and the GWAS of P_l, N_j is the number of individuals in the GWAS for P_j and N_l is the number of individuals in

the GWAS for P_l. If one cannot determine $\mathrm{Cor}(P_j, P_l)$ directly or the sample overlap between the GWA studies is unknown, it is possible to use LD score regression based on the summary statistics to estimate $\mathrm{Cor}(P_j, P_l)\frac{N_{\cap j,l}}{\sqrt{N_j N_l}}$. Note that in the absence of sample overlap, $N_{\cap j,l}$ is zero and thus $\mathrm{Cov}(\widehat{\beta_{1j}}, \widehat{\beta_{1l}})$ is zero.

Having obtained the covariance matrix for $\hat{\beta}$, we can apply the Delta-method[14] to find the standard errors of $\widehat{\lambda_0}$ and $\widehat{\lambda_1}$. The derivation above is done in terms of linear regression assuming a continous response variable. However, a link function may be used to apply this to other response variables.

Here, we outline a GWIS as applied to BMI. BMI is defined as weight (in kilograms) divided by height (in meters) squared. Let μ_w, μ_h denote the means of respectively weight and height and let $\alpha_w, \alpha_h, \beta_w, \beta_h$ denote the intercepts of weight and height and the regression coefficients in the regression of weight and height on the SNP respectively. We assume all of these parameters are known. As shown above, the mean of our approximated BMI is equal to

$$\frac{\mu_w}{\mu_h^2}, \tag{7}$$

i.e., BMI calculated for the mean weight and mean height. In our case, the GWA summary statistics were for standardized weight and height, but were destandardized before computing the GWIS. The destandardization is based on information on population averages and standard deviations obtained from the Netherlands Twin Register (NTR)[15]. The destandardization involves multiplying the effect sizes by the standard deviation and using the population mean as a substitute for the intercept. The mean of the appromixation is in general going to be equal to the function evaluated

in the means of the phenotypes. The linear regression of BMI on the number of effect alleles of a given SNP is

$$\mathrm{BMI} = \alpha_{\mathrm{BMI}} + \beta_{\mathrm{BMI}} \cdot \mathrm{SNP} + e \tag{8}$$

where α_{BMI} is the intercept of the linear regression, β_{BMI} is the regression coefficient and e is the error of the linear regression.

Then, the derived values for the intercept and the regression coefficient become

$$\alpha_{\mathrm{BMI}} = \frac{\alpha_w}{\alpha_h^2} \tag{9}$$

and

$$\begin{aligned}\beta_{\mathrm{BMI}} = &\frac{2\mathrm{EAF}(1-\mathrm{EAF})}{2\mathrm{EAF}(1-\mathrm{EAF})+\mathrm{EAF}^2}\left(\frac{\alpha_w+\beta_w}{(\alpha_h+\beta_h)^2}-\alpha_{\mathrm{BMI}}\right)\\ &+\frac{1}{2}\cdot\frac{\mathrm{EAF}^2}{2\mathrm{EAF}(1-\mathrm{EAF})+\mathrm{EAF}^2}\left(\frac{\alpha_w+2\beta_w}{(\alpha_h+2\beta_h)^2}-\alpha_{\mathrm{BMI}}\right)\end{aligned} \tag{10}$$

where EAF is the effect allele frequency of the SNP.

In our examples, we have used a linear approximation to perform the GWIS; however, in **Supplemental Note 1** we outline the second order approximation of BMI, which should be used in conjunction with a second order Delta-rule.

Given two phenotypes A and B, we can use our method to define a new trait as

$$X := (1+c)A - (1-c)B \tag{11}$$

for a specific constant c. This constant is chosen such that a certain type of correlation between X and B becomes zero and the correlation between A and X is nonzero. Note that this correlation

may be genetic, environmental or phenotypic, depending on the application. In terms of linear regression, this can be seen as

$$(1+c)A = (1-c)B + X$$

so that X is the residual of the linear regression (with fixed coefficients) of $(1+c)A$ on $(1-c)B$.

Note that zero correlation does not imply that X and B are independent; rather, they have only become linearly independent. The expression for c is

$$c := \frac{\operatorname{Var} B - \operatorname{Cov}(A,B)}{\operatorname{Var} B + \operatorname{Cov}(A,B)}. \tag{12}$$

Note that Cov and Var here denote the covariances and variances that are specific to the type of correlation that is considered. For example, in the case of genetic correlation, Cov denotes the coheritability and Var denotes the heritability of the traits.

We derive c by solving $\operatorname{Cov}(X,B) = 0$, so

$$\operatorname{Cov}(X,B) = 0$$

$$(1+c)\operatorname{Cov}(A,B) - (1-c)\operatorname{Cov}(B,B) = 0$$

$$(\operatorname{Cov}(A,B) + \operatorname{Cov}(B,B))c = \operatorname{Cov}(B,B) - \operatorname{Cov}(A,B)$$

$$c = \frac{\operatorname{Var}(B) - \operatorname{Cov}(A,B)}{\operatorname{Var}(B) + \operatorname{Cov}(A,B)}$$

which is well-defined if and only if $\operatorname{Var}(B) \neq -\operatorname{Cov}(A,B)$, that is, B is not equal to $-A$.

An equivalent expression for c is

$$c = \frac{1 - \operatorname{Cor}(A,B)\frac{\sigma_A}{\sigma_B}}{1 + \operatorname{Cor}(A,B)\frac{\sigma_A}{\sigma_B}}.$$

The term $\mathrm{Cor}(A, B)\frac{\sigma_A}{\sigma_B}$ corresponds to the slope of the linear regression of B on A. Thus, X is actually the distance between the data points in a 2-dimensional plane and their projection onto the linear regression line of A on B, rather than the vertical distance between the predicted value of A and the data point. This allows for error in the assessment of both A and B, rather than only measurement error in A. This is important since X is analyzed in a GWIS and the estimates for the association between both A and B and a SNP have a certain standard error.

1. Bulik-Sullivan *et al.* An atlas of genetic correlations across human diseases and traits. *bioRxiv* (2015).

2. Randall, J. C. *et al.* Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genetics* **9**, e1003500 (2013).

3. Gibbs, R. A. *et al.* The international hapmap project. *Nature* **426**, 789–796 (2003).

4. Vink, J. M. *et al.* Sex differences in genetic architecture of complex phenotypes? *PLoS ONE* **7**, e47371 (2012).

5. Rietveld, C. A. *et al.* Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).

6. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

7. Perry, J. R. B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014). Letter.

8. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014). Letter.

9. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**, 333–338 (2011).

10. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry* **19**, 1017–1024 (2014).

11. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature genetics* **45**, 984–994 (2013).

12. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols* **9**, 1192–1212 (2014).

13. Finn, J. D. *A general model for multivariate analysis.* (Holt, Rinehart & Winston, 1974).

14. Greene, W. *Econometric Analysis* (Pearson/Prentice Hall, 2008).

15. Willemsen, G. *et al.* The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies. *Twin Research and Human Genetics* **13**, 231–245 (2010).

16. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* **381**, 1371 – 1379 (2013).

Supplemental Note 1. A second order approximation of BMI

The second order Taylor approximation of BMI is given by

$$Q_i = \frac{\mathbb{E}[W_i \mid S_i = s]}{\mathbb{E}[H_i \mid S_i = s]^2} + \frac{W_i - \mathbb{E}[W_i \mid S_i = s]}{\mathbb{E}[H_i \mid S_i = s]^2} + \frac{-2\mathbb{E}[W_i \mid S_i = s]\,(H_i - \mathbb{E}[H_i \mid S_i = s])}{\mathbb{E}[H_i \mid S_i = s]^3}$$
$$+ \frac{1}{2}\left(0 + \frac{-4(W_i - \mathbb{E}[W_i \mid S_i = s])(H_i - \mathbb{E}[H_i \mid S_i = s])}{\mathbb{E}[H_i \mid S_i = s]^3} + \frac{6\mathbb{E}[W_i \mid S_i = s]\,(H_i - \mathbb{E}[H_i \mid S_i = s])^2}{\mathbb{E}[H_i \mid S_i = s]^4}\right)$$

so that

$$\mathbb{E}[Q_i \mid S_i = s] = \frac{\mathbb{E}[W_i \mid S_i = s]}{\mathbb{E}[H_i \mid S_i = s]^2} - \frac{2\,\mathrm{Cov}(W_i, H_i|S_i = s)}{\mathbb{E}[H_i \mid S_i = s]^3} + \frac{3\mathbb{E}[W_i \mid S_i = s]\,\mathrm{Var}(H_i|S_i = s)}{\mathbb{E}[H_i \mid S_i = s]^4}$$

Then, using this for our linear regression

$$\mathbb{E}[Q_i \mid S_i = s] = \lambda_0 + \lambda_1 s$$

together with $s = 0$ gives

$$\widehat{\lambda_0} = \frac{\alpha_w}{\alpha_h^2} - \frac{2\,\mathrm{Cov}(W, H)}{\alpha_h^3} + \frac{3\alpha_w\,\mathrm{Var}(H)}{\alpha_h^4}$$

under the assumption that $\mathrm{Cov}(W_i, H_i|S_i = s) = \mathrm{Cov}(W, H)$ and a similar assumption for the variance of height. Then,

$$\begin{aligned}
\widehat{\lambda_1} =& \frac{2\mathrm{EAF}(1-\mathrm{EAF})}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}(f(\beta_{01} + \beta_{11}, \beta_{02} + \beta_{12}, \ldots, \beta_{0k} + \beta_{1k}) - \widehat{\lambda_0}) \\
&+ \frac{\mathrm{EAF}^2}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}\frac{f(\beta_{01} + 2\beta_{11}, \beta_{02} + 2\beta_{12}, \ldots, \beta_{0k} + 2\beta_{1k}) - \widehat{\lambda_0}}{2} \\
=& \frac{2\mathrm{EAF}(1-\mathrm{EAF})}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}\left(\frac{\alpha_w + \beta_w}{(\alpha_h + \beta_h)^2} - \left(\frac{\alpha_w}{\alpha_h^2} - \frac{2\,\mathrm{Cov}(W, H)}{\alpha_h^3} + \frac{3\alpha_w\,\mathrm{Var}(H)}{\alpha_h^4}\right)\right) \\
&+ \frac{\mathrm{EAF}^2}{2\mathrm{EAF}(1-\mathrm{EAF}) + \mathrm{EAF}^2}\frac{\left(\frac{\alpha_w + 2\beta_w}{(\alpha_h + 2\beta_h)^2} - \left(\frac{\alpha_w}{\alpha_h^2} - \frac{2\,\mathrm{Cov}(W, H)}{\alpha_h^3} + \frac{3\alpha_w\,\mathrm{Var}(H)}{\alpha_h^4}\right)\right)}{2}
\end{aligned}$$

Figure legends and table captions

Figure 1: A is a Manhattan plot of $-\log p$-values for the BMI GWAS as performed by Randall et al., B is a Manhattan plot of $-\log p$-values for the BMI GWIS using male height data, whereas C is a Manhattan plot of $-\log p$-values for the BMI GWIS using female height data. The location on the x-axis corresponds to the genomic location of the SNP. In each figure, the blue line corresponds to $p = 1 \cdot 10^{-5}$, whereas the red line corresponds to $p = 5 \cdot 10^{-8}$.

Figure 2: A schematic representation of the role of GWIS in relation to traditional GWAS. GWIS provides a connection between the several GWA studies of phenotypes and a GWA study of a function of these phenotypes, without requiring access to the actual phenotypical data.

Table 1: The table reports estimated genetic correlations along with their standard errors between BMI based on GWAS, GWIS using male height data, GWIS using female height data and other traits. These correlations are obtained with LD score regression.

Table 2: The table reports estimated genetic correlations along with their standard errors between schizophrenia, bipolar disorder, unique schizophrenia, unique bipolar disorder and educational attainment. These correlations are obtained with LD score regression.

Supplementary Table 1: The table reports associations found in the BMI GWIS based on male height data which are not significant in the original BMI GWAS. Note the differences in effect sizes are minor and p-values are close to the significance threshold in both analyses.

Supplementary Table 2: The table reports associations found in the BMI GWIS based on female height data which are not significant in the original BMI GWAS. Note the strong deviance in effect size, standard error and p-value between the analyses. This deviance is likely caused by the limited number of individuals for which these SNPs were measured in the female height GWAS.

URLs

PGC summary statistics used in the schizophrenia and bipolar disorder analysis[10]:

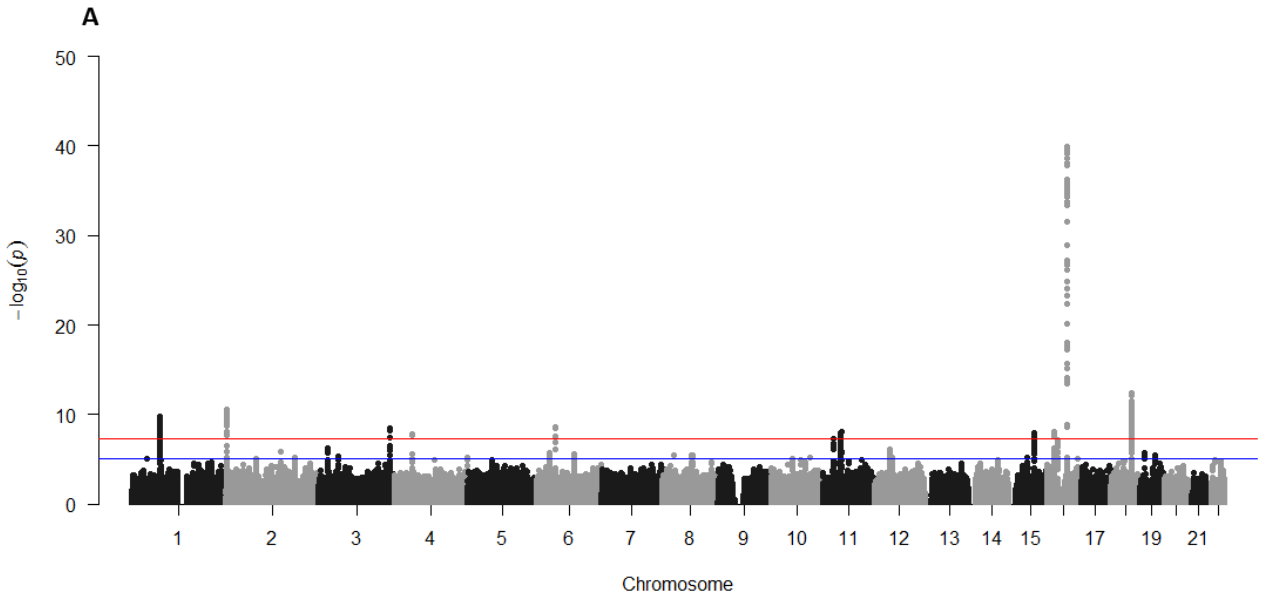
https://www.med.unc.edu/pgc/results

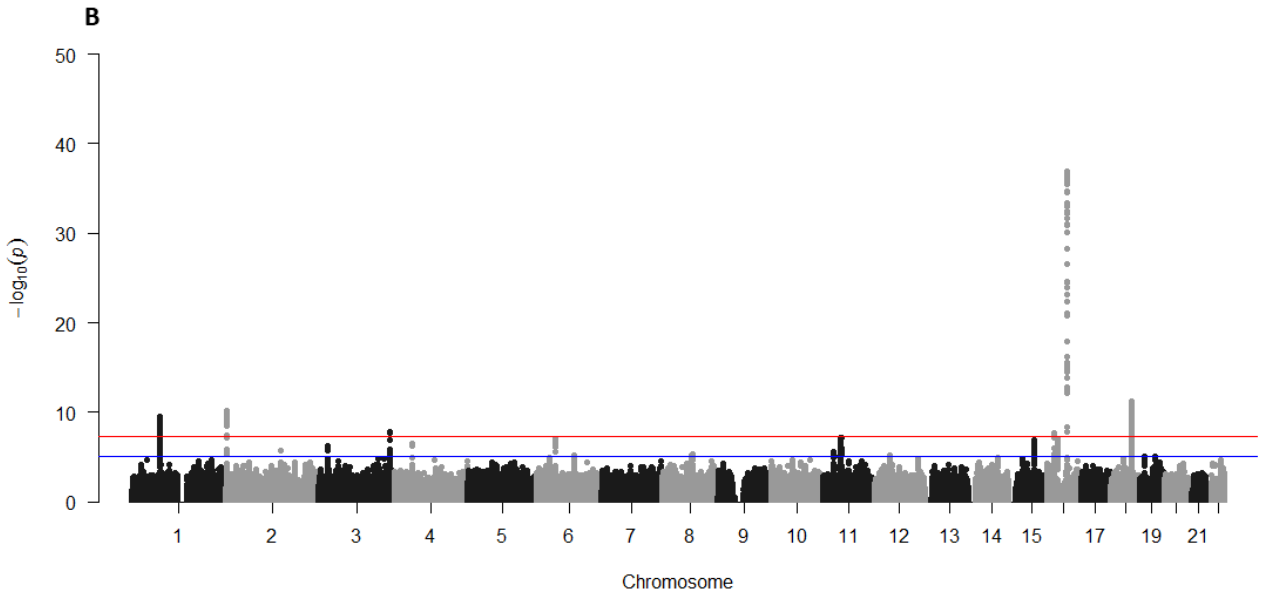
GIANT summary statistics used in the BMI analyses[2]:

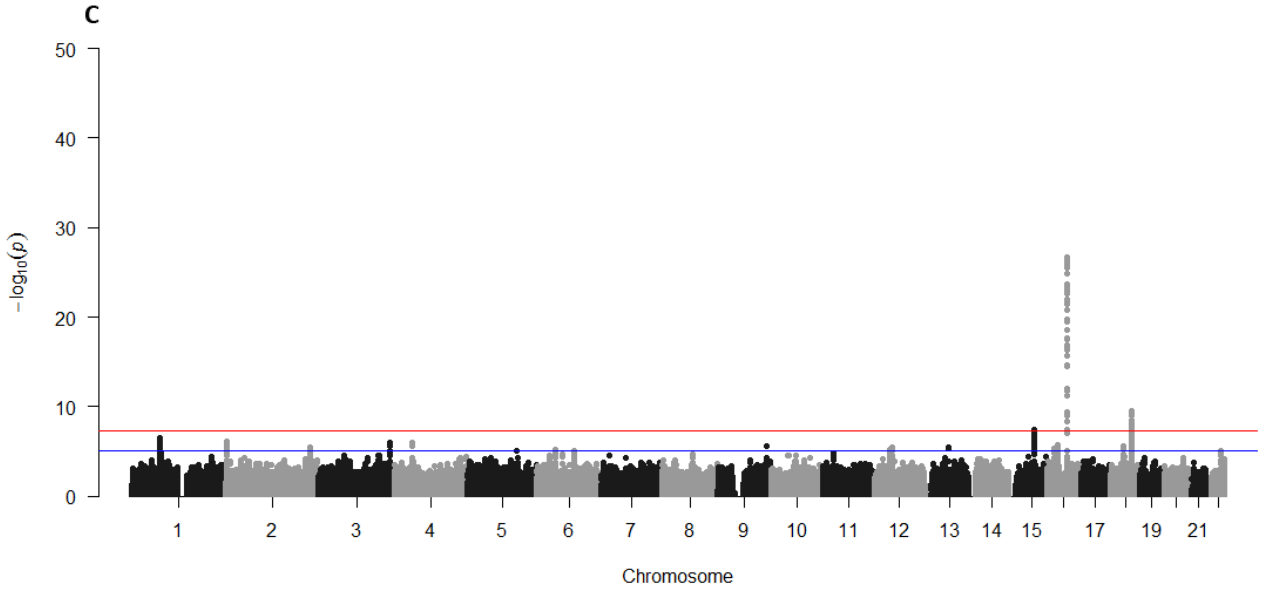
https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

HAPMAP 2 allele frequencies were obtained from the the public webpage of the HAPMAP Consortium[3]:

http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en







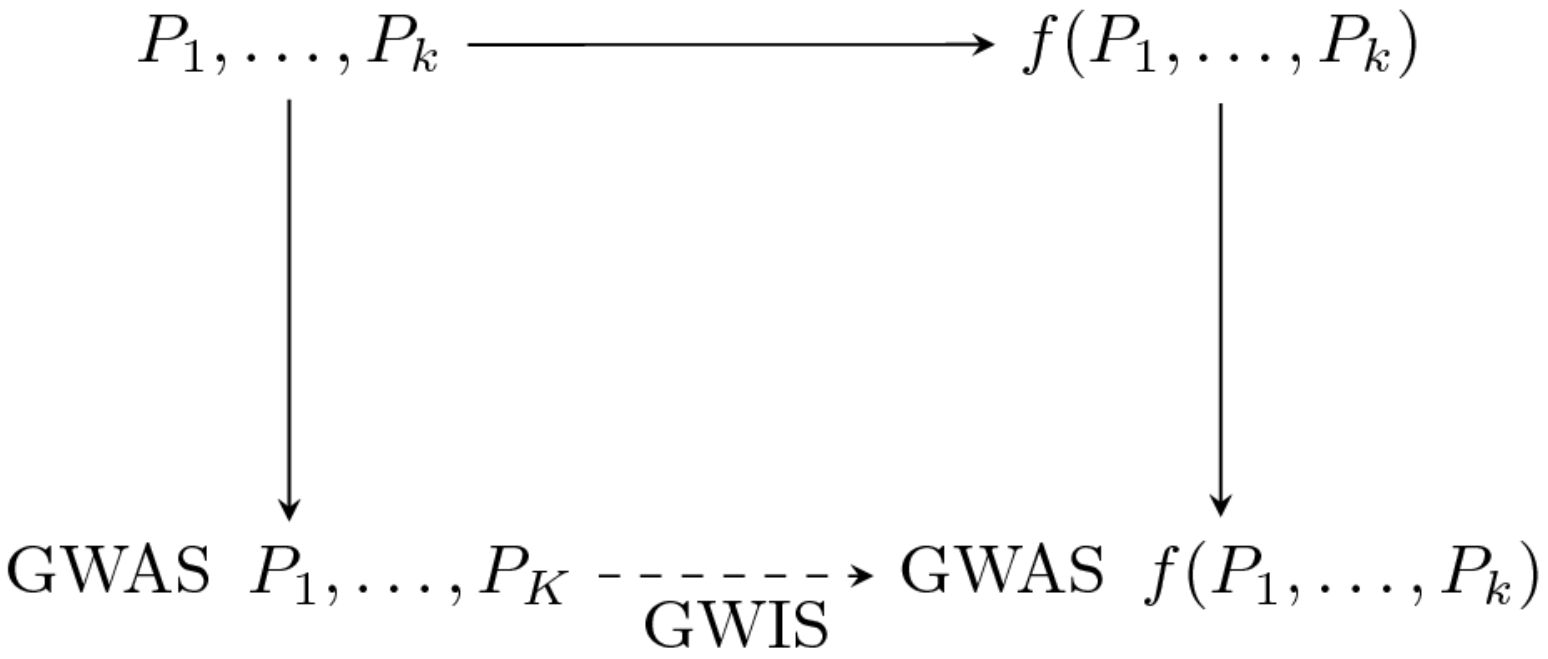


Table 1

	GWIS (female height)	GWAS	Rheumatoid arthritis	Age at menarche	LDL	Educational attainment	Coronary artery disease
GWIS (male height)	0.967 (0.012)	1.007 (0.002)	0.029 (0.045)	-0.338 (0.035)	0.019 (0.058)	-0.145 (0.053)	0.153 (0.063)
GWIS (female height)	-	0.974 (0.013)	0.018 (0.053)	-0.371 (0.041)	-0.003 (0.061)	-0.157 (0.062)	0.150 (0.071)
GWAS	-	-	0.039 (0.042)	-0.332 (0.032)	0.013 (0.050)	-0.160 (0.047)	0.173 (0.061)

Table 2

	Unique SCZ	Unique BIP	SCZ	BIP
Educational attainment	0.041 (0.082)	0.218 (0.102)	0.148 (0.050)	0.273 (0.067)
BIP	0.106 (0.110)	0.816 (0.031)	0.572 (0.063)	-
SCZ	0.882 (0.026)	-0.016 (0.101)	-	-

Supplemental Table 1

Rs-ID	GWIS Beta	GWIS S.E.	GWIS p-value	GWAS Beta	GWAS S.E.	GWAS p-value	N
rs4649957	-0.03823948	0.00687369	2.65E-08	-0.037	0.0068	6.93E-08	58612.4
rs4650139	0.03592988	0.00648556	3.02E-08	0.034	0.0065	1.21E-07	58611.5
rs7530496	-0.03572799	0.00648196	3.55E-08	-0.034	0.0065	1.69E-07	58504.2

Supplemental Table 2

Rs-ID	GWIS Beta	GWIS S.E.	GWIS p-value	GWAS Beta	GWAS S.E.	GWAS p-value	N
rs11206949	-5.71393947	0.010152767	0	0.94	1.1	0.38	1666
rs11206950	-2.91013533	0.00155349	0	-0.94	1.1	0.38	1666
rs11206951	-5.71393947	0.010152767	0	0.94	1.1	0.38	1666
rs11206952	-5.71393947	0.010152767	0	0.94	1.1	0.38	1666
rs12073104	-5.71969538	0.010170506	0	0.94	1.1	0.38	1666
rs12080262	-2.91013533	0.00155349	0	-0.94	1.1	0.38	1666
rs12083335	-2.9083309	0.005892014	0	-0.94	1.1	0.38	1666
rs12087759	-2.91013533	0.00155349	0	-0.94	1.1	0.38	1666