A novel framework for analysis of the shared genetic background of correlated traits

Gulnara R. Svishcheva[1,2], Evgeny S. Tiys[1,3], Elizaveta E. Elgaeva[1,3], Sofia G. Feoktistova[1,3], Paul R. H. J. Timmers[4,5], Sodbo Zh. Sharapov[1,3], Tatiana I. Axenovich[1,3], Yakov A. Tsepilov[1,3*]

[1] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

[2] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

[3] Novosibirsk State University, Novosibirsk, Russia

[4] MRC Human Genetics Unit, MRC Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

[5] Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK

*correspondence to: tsepilov@bionet.nsc.ru

Keywords: GWAS, shared genetic component, linear combination of traits, shared heritability, proportion of heritability explained by SGF

Abstract

We propose a novel effective framework for analysis of the shared genetic background for a set of genetically correlated traits using SNP-level GWAS summary statistics. This framework called SHAHER is based on the construction of a linear combination of traits by maximizing the proportion of its genetic variance explained by the shared genetic factors. SHAHER requires only full GWAS summary statistics and matrices of genetic and phenotypic correlations between traits as inputs. Our framework allows both shared and unshared genetic factors to be to effectively analyzed. We tested our framework using simulation studies, compared it with previous developments, and assessed its performance using three real datasets: anthropometric traits, psychiatric conditions and lipid concentrations. SHAHER is versatile and applicable to summary statistics from GWASs with arbitrary sample sizes and sample overlaps, allows incorporation of different GWAS models (Cox, linear and logistic) and is computationally fast.

Introduction

There is a growing interest in studying the shared genetic background between genetically correlated traits[1-5] (see, for example, the number of PubMed search results by year for keywords related to "shared genetic background"). Studying this shared genetics between traits can help discover pleiotropic interactions, common genes and pathways, and identify genetic effects that are unique for each trait.

The problem of the decomposition of the variance of several traits into the shared/unshared genetic and environment components were first formulated by S. Write in 1921 [6]. There are widely used classic twin designs to have this problem solved. They are based on structural equation modelling, in particular, multivariate pathway models assuming the existence of the genetic influences common for all traits and unique for each trait [7]. These designs are implemented only for the variance decomposition, but not for the identification of the genetic factors that determine these genetic impacts.

There are several terms for these common and unique genetic impacts. Hereafter we will call them the 'shared genetic impact' (SGI) and 'unshared genetic impacts' (UGI). The genetic factors that determine these impacts will be called 'shared genetic factors' (SGF) and 'unshared genetic factors' (UGF), respectively. The heritability of each trait explained by SGF and UGF will be called 'shared heritability' and 'unshared heritability', respectively.

The application of different methods of multivariate analysis in genome-wide association studies (GWAS) allows the problem of SGF and UGF identification to be partially solved [8-13]. The multivariate methods involve complicated genetic or/and phenotypic correlation structures of traits in the analysis. In most cases, this increases the power of detection of the loci associated with several traits due to pleiotropic effects. If the detected locus has a pleotropic effect on all studied traits, the locus could potentially be attributed to SGF, and if not, to UGF. However, a pleiotropic effect of the locus on all studied traits is necessary but insufficient for inclusion of this locus in SGF (at least effects should be also collinear between traits, see the model description below). Also, if a locus belonging in fact to SGF was not identified as having pleotropic effects on all traits due to a limited statistical power of the analysis, then the locus can be erroneously assigned to UGF. Moreover, this approach of SGF identification assumes a manual classification of loci, which prevents the use of more sophisticated modern in-silico approaches for genetic analysis, for example, the ones that rely on GWAS summary statistics [14]. To our knowledge, there is no specific method that could be good for both variance component decomposition and identification of SGF and UGF.

We had previously developed a method for obtaining genetically independent phenotypes (GIPs) [2]. This method is based on the calculation of the principal components using genetic rather than phenotypic correlations. We applied this method to genetically correlated pain phenotypes and aging related phenotypes and showed that the first GIP component, GIP1, that explains the largest proportion of the genetic variance probably could be interpreted as SGI [2, 15]. This makes GIP promising for identification of loci attributed to SGF. However, this method was not designed specifically for SGI analysis. In addition, no specific experiments have been performed to validate the approach or to estimate its statistical properties.

Here, we present a novel general framework for the estimation of shared and unshared heritability and identification of the shared and unshared genetic factors using the summary statistics of original traits. The essence of our approach is to find the optimum linear combination of traits which has the maximum proportion of its genetic variance explained by the SGF. We validated our framework using simulation studies under different scenarios, by comparing it with the developed GIP approach, and assessed its performance using three real datasets: anthropometric indices, psychiatric disorders and conditions, and lipid concentrations.

Results

Abbreviations and terms

SHAHER: a framework for the estimation of the shared and unshared heritability of studied traits and identification of the shared and unshared genetic factors using the summary statistics of original traits.

SGI: shared genetic impact.

UGI: unshared genetic impact.

SGF (shared genetic factors): genetic factors involved in the control of all studied traits and whose effects are collinear between all studied traits; SGI is due to SGF.

Shared heritability: the proportion of the trait variance explained by SGF.

SGIT (shared genetic impact trait): a trait defined as a linear combination of original traits maximizing its shared heritability.

α: the coefficients of an optimum linear combination of original traits for building the SGIT.

UGF (unshared genetic factors): the residual genetic factors of an original trait after exclusion of the SGF; UGI is due to UGF.

Unshared heritability: the proportion of the trait variance explained by UGFs.

UGIT (unshared genetic impact trait): an original trait after adjustment for the SGIT.

γ: the coefficients of a linear combination of original traits for building the UGIT.

MaxSH (MAXimization of Shared Heritability): a method for estimating the shared and unshared heritability of each trait and calculating the coefficients of the linear combination of the original traits: α, to build the SGIT, and γ, to build the UGITs.

sumCOT (summary-level GWAS for linear Combination of Traits): a method to compute GWAS summary statistics for the linear combination of the original traits using their summary statistics.

Shared heredity model

We adopted a commonly used multivariate pathway model [7] in terms of SGF and UGF. We call it the 'shared heredity model'. For simplicity, we consider SGF and UGF as biallelic SNPs and consider a sample of N unrelated individuals measured for K traits and genotyped for M SNPs. For a standardized normal trait, y ($N\times 1$), the traditional polygenic (null) model takes the form: $y = G\beta + \varepsilon$, where G is an ($N\times M$) matrix of standardized genotypes; β ($M\times 1$) and ε ($N\times 1$) are genetic and non-genetic random effects, respectively; $\beta \sim N(\mathbf{0}, h^2 I_M)$ and $\varepsilon \sim N(\mathbf{0}, (1-h^2)I_N)$, where $\mathbf{0}$ is a null mean vector, h^2 is the trait heritability, and I is an identity matrix of the given dimension. For unrelated individuals, we expect $y \sim N(\mathbf{0}, I_N)$.

We propose to divide M SNPs into two non-overlapping SNP sets with sizes M_0 and M_1 ($M_0 + M_1 = M$). The set of M_0 SNPs called 'SGF' includes only those SNPs whose effects on all traits are collinear. The set of M_1 SNPs consists of the other SNPs, which do not have shared joint influence on all traits at once, this set being called 'UGF'. In accordance with M, G is divided into two matrices, G_0 ($N\times M_0$) and G_1 ($N\times M_1$). To decompose every trait into components explained by the SGF and UGF, we rewrote the traditional polygenic model in terms of G_0 and G_1

$$y_i = \underbrace{G_0 b_{0_i}}_{due\ to\ SGF} + \underbrace{G_1 b_{1_i}}_{due\ to\ UGF} + \varepsilon_i. \quad (1)$$

Here, the first and second terms are genetic components explained by SGF and UGF, respectively, which are assumed independent. In the first term, b_{0_i} is an (M_0×1) vector of non-zero SGF effects, which can be presented as $\beta_0 w_i\sqrt{h_i^2}$, where β_0 is an (M_0×1) non-zero vector that is the same for all traits, $\beta_0 \sim N\left(0, I_{M_0}\right)$, and $w_i^2 h_i^2$ is the heritability of the i-th trait explained by SGF. Here w_i is a non-zero trait-specific multiplier: w_i^2 denotes the proportion of h_i^2 explained by SGF; the value of w_i can be positive and negative, indicating the direction of the SGF effect on the i-th trait. $G_0\boldsymbol{\beta}_0$ is the so-called shared genetic impact or SGI. In the second term of Model (1), b_{1_i} is an (M_1×1) vector of UGF effects, which can be presented as $b_{1i} = \beta_{1i}\sqrt{(1-w_i^2)h_i^2}$, $\beta_{1i} \sim N\left(0, I_{M_1}\right)$. In contrast to β_0, β_{1_i} are different for different traits, moreover they are not collinear. For illustrative purposes, we rewrote Equation (1) as:

$$y_i = \underbrace{\underbrace{G_0\boldsymbol{\beta}_0}_{SGI}\ w_i\sqrt{h_i^2}}_{due\ to\ SGF} + \underbrace{G_1\boldsymbol{\beta}_{1_i}\sqrt{1-w_i^2}\sqrt{h_i^2}}_{due\ to\ UGF} + \varepsilon_i.$$

Overview of the SHAHER framework

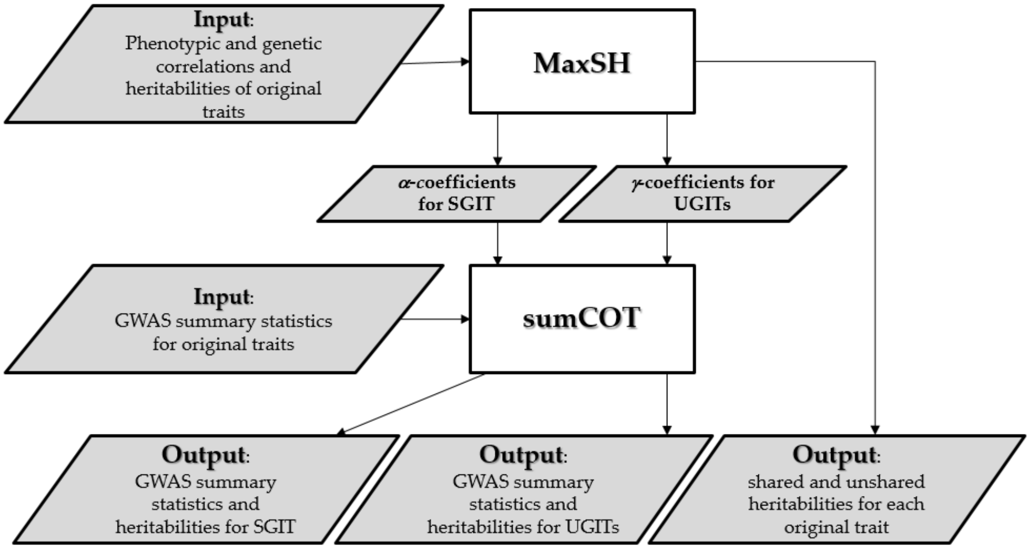
Input:
Phenotypic and genetic correlations and heritabilities of original traits
MaxSH
α-coefficients for SGIT
γ-coefficients for UGITs
Input:
GWAS summary statistics for original traits
sumCOT
Output:
GWAS summary statistics and heritabilities for SGIT
Output:
GWAS summary statistics and heritabilities for UGITs
Output:
shared and unshared heritabilities for each original trait

Figure 1. Flowchart of the SHAHER framework. Details are given in the text.

For analyses of the SGI and UGI on a set of correlated traits, we propose an effective multi-stage framework named SHAHER (see Figure 1). The concept of the framework is first to partition the genetic basis of each original trait into two components: one shared by all the original traits and one shared not by all the original traits, and then to identify the SNPs that contribute to these genetic components. To do this, we propose to construct new traits: (1) an SGIT as a linear combination of original traits, which has the maximum possible heritability explained by the SGF, and (2) UGITs as linear combinations of the original traits, which are obtained by adjusting the original traits for the SGIT. This means that the genetic basis of the UGITs is predominantly determined by the UGF.

Our framework requires matrices of phenotypic correlations (U_{phen}) between the original traits, the matrices of genetic correlations (U_{gen}) between the original traits, the heritabilities of the original traits and GWAS summary statistics of the original traits as inputs. It is worth noting that U_{phen}, U_{gen} and heritabilities could be estimated using GWAS summary statistics of the original traits, for example, by the LD score regression method [16].

SHAHER starts with a preliminary stage, which verifies the presence of SGI in a given set of traits. This is achieved by checking the following requirements for U_{gen}: it must be positive definite; the absolute values of its elements must be significantly more than a given threshold, and the rank of the correlation matrix derived from U_{gen} by rounding its elements to extremal correlation values, either -1 or 1, must be equal to one. If the requirements are met, we turn to the basic stages of SHAHER.

*The MaxSH stage***.** To determine the α and γ coefficients for the linear combinations of the original traits to build the SGIT and UGITs, we developed the MaxSH method, which is based on the correlation component model given below. This model partitions the phenotypic correlation matrix, U_{phen}, into environmental and genetic components, U_{env} and U_{gen}, respectively, the latter being further subdivided into two components caused by the SGF and UGF:

$$\begin{gathered} U_{phen} = \underbrace{\sqrt{H^2}U_{gen}\sqrt{H^2}}_{genetic\ component} + \underbrace{\sqrt{I-H^2}U_{env}\sqrt{I-H^2}}_{environmental\ component} \\ U_{gen} = \underbrace{W\mathbf{1}\mathbf{1}^{\mathrm{T}}W}_{due\ to\ SGF} + \underbrace{\sqrt{I-W^2}U_{unsh}\sqrt{I-W^2}}_{due\ to\ UGF} \end{gathered} \tag{2}$$

Here W is a diagonal matrix, whose i-th diagonal element is w_i; U_{unsh} is a matrix of genetic correlations explained by UGF; H^2 is a diagonal matrix, whose i-th diagonal element is h_i^2, and $\mathbf{1}$ is a (k×1) vector of units. Using this model, MaxSH solves several tasks.

First of all, using only the genetic correlation matrix, U_{gen}, we estimate the proportion of heritability of every trait explained by SGF (W). To do this, we minimize the difference between U_{gen} and the auxiliary matrix V. This matrix is built using formula (2), with the identity matrix used instead of $U_{unsh,}$. The second task is to determine the α-coefficients, which is solved by maximizing the shared heritability of the SGIT. This task is analytically solved as

$$a = \frac{U_{phen}^{-\frac{1}{2}} H W \mathbf{1}}{\sqrt{\mathbf{1}^T W H U_{phen}^{-1} H W \mathbf{1}}}.$$

It requires U_{phen}, H^2 and W as input data.

After determining the α-coefficients and building the SGIT, we build a UGIT for every trait using the residual regression equation $UGIT_i = y_i - SGIT{*}c_i$, where c_i is the impact of the SGIT on the i-th original trait, defined as

$$c_i = cov_{gen}(y_i, SGIT)/h_{SGIT}^2.$$

Here cov_{gen} denotes a genetic covariance. Note that we should use genetic rather than phenotypic covariances, as our goal is to adjust only the genetic components of the original traits.

Since the SGIT is the linear combination of the original traits, the UGITs are linear combinations of the original traits, too. The coefficients of these linear combinations called the γ-coefficients form the matrix of the γ-coefficients $\Gamma = (I_K - \alpha c^T)$, where the i-th column of Γ corresponds to linear combination coefficients for building the i-th UGIT.

The sumCOT stage. This stage is aimed at obtaining GWAS summary statistics for the SGIT and UGITs using the previously determined α and γ coefficients, GWAS summary statistics (Z-scores, allele frequencies and sample sizes for each SNP) for the original traits and the matrix of phenotypic correlations. The method can use Z scores obtained from any regression model and allows for varying sample sizes and sample overlap between traits. This sample overlap is incorporated into the estimation of the matrix of phenotypic correlations. In short, the SNP effects for combined trait are calculated by summing effect estimates from the individual trait GWASes, each multiplied by their corresponding linear coefficient (α or γ), and standardized by the expected variance. The standard errors of the SNP effect are calculated using variance-covariance arithmetic, taking into account the phenotypic covariance between GWAS results to adjust for the sample overlap. Effective sample sizes are then estimated based on the median Z statistic and allele frequencies by solving Equation (1) in [17].

At the final stage, SHAHER checks for the correctness of the output. In particular, we anticipate that UGITs do not have a shared genetic basis. This is verified by applying MaxSH to the matrix of correlations between UGITs.

To summarize, our framework estimates shared and unshared heritabilities for each of the studied original traits and produces GWAS summary statistics for the SGIT and UGITs as outputs.

The full details and mathematical formulae of SHAHER are in *Supplementary Methods*.

Simulation study

To assess the MaxSH performance, we conducted simulation studies. We (1) assessed the accuracy of *w* estimates (using ΔW metrics estimated as $\left(\frac{w_0 - w_{est}}{w_o}\right)^2$, where w_0 and w_{est} are modeled and estimated *w*, respectively) with respect to the loss function given in Fig. 3, (2) assessed the proportion of the shared heritability to the total heritability of the SGIT (the *Q*-value) with respect to the loss function, and (3) compared the analytically predicted total/shared heritabilities of two traits: SGIT and the first component, GIP1, obtained by GIP method [2]. The *Q*-value can be interpreted as the specificity metrics of the SGIT: the closer the *Q*-value to 1, the lower the share of unshared heritability in the total heritability of the SGIT. The simulation scenarios were based on six varying parameters that describe the properties of the genetic and phenotypic correlation matrices. Under each scenario, we considered two situations, where all traits have the same w^2 and different w^2's. To distinguish between these situations, we will hereinafter write either 'w^2' or 'different w^2's'. In total, we performed 10,000 iterations of simulations for each of 288 scenarios.

The results are presented in Supplementary Figures S1-18. For all scenarios, there are few general patterns: (1) the higher simulated *w* values, the higher the accuracy of the *w* estimates, (2) the accuracy of the *w* estimates and the *Q*-value increase with an increasing in the number, *K*, of traits, (3) for all scenarios with w^2>0.8, ΔW was very low (<0.025) and the *Q*-value was more than 90%.

For all scenarios with three traits, the accuracy of the *w* estimates was in general low: ΔW was not higher than 0.7 for scenarios with w^2=0.2 and 0.3, although at w^2 equal to or higher than 0.4 ΔW was less than 0.2. The *Q*-value was higher than 60% for almost all scenarios with $w^2 \geq 0.4$. In almost all cases, the total and shared heritabilities of the SGIT were higher than the corresponding heritabilities of GIP1, except for the scenarios with $h^2=0.8$.

For the scenarios with four and five traits, the accuracy of w estimates was higher: $\Delta W<0.15$ for $w^2 \geq 0.4$ and $\Delta W<0.05$ for $w^2 \geq 0.5$. For the scenarios with $w^2 \geq 0.5$, the Q-value was more than 70% for four traits and more than 80% for five traits. Again, the total and shared heritabilities of the SGIT were higher than the corresponding heritabilities of GIP1 under all scenarios, except for the scenarios with $h^2=0.8$. In the scenarios with $h^2=0.8$, the total and shared heritabilities of the SGIT were higher than those of the GIP1 at $w^2 \geq 0.5$.

In summary, the performance of MaxSH was suitable at $w^2 \geq 0.5$ and when the number of traits being higher than or equal to four. In the case of small w or three traits, the results of MaxSH should be interpreted with caution.

Real data assessment

We applied SHAHER to three datasets: anthropometric (five traits), psychometric (four traits) and lipid traits (three traits). We should note that the performance of SHAHER applied to three traits is limited (see simulation results), yet still passable, although the results should be interpreted with caution. We present SHAHER results for anthropometric traits in the main text as an example. The full results for the psychometric and lipid traits are presented in Supplementary Results.

At the first step, we confirmed that SGI exists for five traits. At the second step, we determined the α and γ coefficients and their CI (see Supplementary Table 1a). At the third step, we applied sumCOT and obtained GWAS results for the SGIT and UGITs (see Supplementary Table 2a for heritability estimates and LD score regression intercepts). SHAHER results are presented in Figure 2.

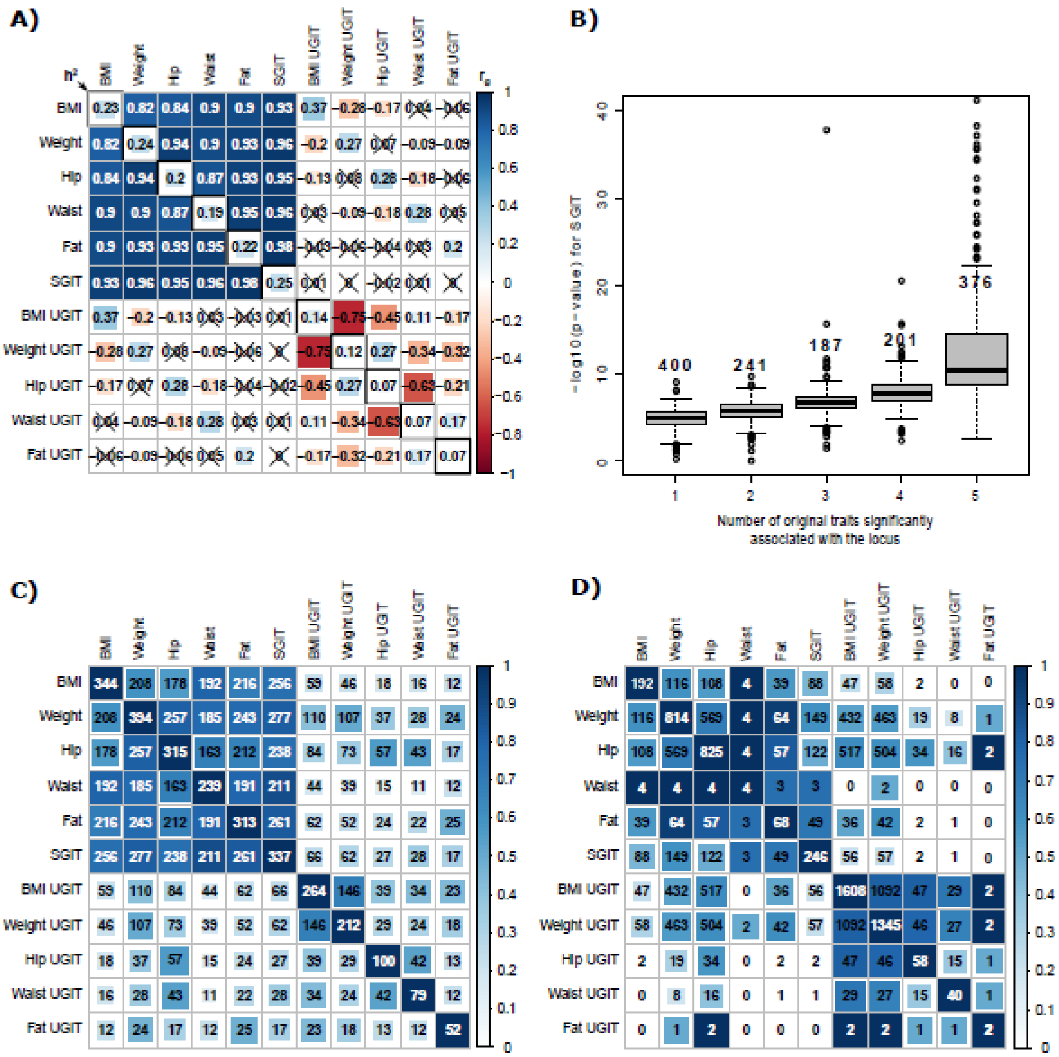

Figure 2. Results of the application of SHAHER to anthropometric traits. A) The heatmap of genetic correlations between the original, SGI and UGI traits. The number, color strength and size of the squares in the matrix show the values of the correlation coefficients between the traits. The diagonal elements represent heritabilities. Crossed out values indicate insignificant correlations. B) Boxplots of –$\log_{10}$(p-value) for the SGIT with respect to the number of the original traits significantly associated with the locus. Two outliers for loci with –$\log_{10}$(p-value) > 40 are omitted. The number at the top of the boxplot corresponds to the number of significant SNPs. C) The heatmap of the numbers of overlapping loci between traits. The numbers in the cells represent the absolute numbers of overlapping loci. The color strength and size of the squares in the cells show the relative scaled number of overlapping loci (on the scale from 0 to 1). The diagonal elements represent the number of loci found for every trait. D) The heatmap of the numbers of overlapping gene sets between traits. The color strength and size of the squares in the cells show the relative scaled number of overlapping gene sets (on the scale from 0 to 1). The diagonal elements represent the number of gene sets found for every trait.

Figure 2A demonstrates genetic correlations between all pairs of the original anthropometric traits, SGIT and UGITs. All the original traits were positively correlated with r > 0.82. We did not observe any significant genetic correlation between the SGIT and the UGITs. Moreover, we did not observe additional SGI among UGITs, which was up to expectation. The heritabilities of the UGITs varied from 0.07 to 0.14.

We revealed a dependence of the SGIT p-value from the number of the original traits significantly associated with the locus (Figure 2B). It clearly shows that the loci associated with all the original traits have lower SGIT p-values than the other loci.

Joint clumping of 11 traits (five original traits, five UGITs and SGIT) resulted in 820 genome-wide significantly associated loci (p-value < 5×10^{-8}, Supplementary Table 3a). If a locus was not significantly associated with any of the original traits, it was considered new. SGIT was significantly associated with 337 SNPs. We detected no new loci among SGIT loci. The clumping of UGITs revealed 422 loci, of which 199 were new. At the same time, the clumping of only original traits allowed 621 loci to be detected, of which 161 could not be detected by analyzing SGIT or UGITs. Thus, the joint analysis of SGIT and UGITs increased the number of associated loci by more than 32%. Figure 2C reflects the overlapping between significantly associated loci for 11 analyzed traits. There is a weak albeit non-zero overlap between loci for UGITs and SGIT, although the genetic correlation between them is zero. It could be due to the conservative settings of the clumping procedure, which tends to clump together closely located loci, and due to some level of unspecificity of the SHAHER.

Next, we checked how enriched gene sets overlap between the SGIT, UGITs and original traits (see Figure 2D). Significant results (FDR<5%) of enriched gene sets and tissue enrichment analyses are presented in Supplementary Table 4. As expected, the heatmap of the overlapping gene sets looks similar to the heatmap of genetic correlations and the heatmap of the overlapping loci. Moreover, there were almost no overlap between SGIT and any UGIT. For the original traits, the number of enriched gene sets varied a lot: from 4 for the waist to 825 for the hip circumference. It should be noted that we observed a high number of enriched gene sets for the BMI UGIT, 1608, which was almost ten times the value for BMI (192).

Finally, we obtained GIP1 GWAS statistics and calculated the genetic correlations between the SGIT and GIP1. The genetic correlation was higher than 0.97.

Discussion

We developed a new fast and efficient framework, which allows us to decompose the heritability of each trait from a given set of traits into two components. One of them is explained by shared genetic factors common to all traits. Another one is explained by unshared genetic factors specific for each trait. The framework not only decomposes heritability, but also identifies SNPs associated with the shared and unshared genetic impacts. To our knowledge, this framework is unparalleled. It has an additional advantage: it uses GWAS summary statistics obtained for original traits and does not require raw genotype or phenotype data.

We compared the performances of MaxSH and GIP in identifying the shared genetic components. GIP calculates the linear combination coefficients via the eigenvalues of the genetic covariance matrix and can be considered a close approximation to MaxSH. In our simulations, GIP and MaxSH were similar in almost all scenarios, with MaxSH being somewhat superior in terms of the power (total heritability) and quality (shared heritability). If obtaining genetically independent phenotypes is not the aim, we suggest using SHAHER, because it is more robust and gives additional metrics like SGI contributions to the heritability of the original traits.

The framework is computationally effective. The stage using sumCOT is the most time consuming. However, it takes only several minutes for an average computer to conduct a GWAS of a linear combination of traits with 6M SNPs using a C++ implementation of the sumCOT. MaxSH, based on numerical optimization procedures, and the other parts of the framework take seconds.

The proposed sumCOT method can be applied as an independent tool to address additional tasks. One of them is making a summary-level adjustment of traits by other traits using the same scheme as was used for obtaining the UGIT GWAS statistics. This can be helpful, for example, for ridding the studied trait's genetic component of the genetic component that was caused by the confounding or unaccounted effects of assortative mating or family effects, which is quite a problem in GWAS at the biobank scale [15, 18]. Another task is a GWAS for the trait that appears as a linear combination of the original traits. The sumCOT method is robust to differences in sample sizes used for GWASs of original traits and is applicable to different GWAS models (Cox, linear or logistic).

The main interest in the application of the SHAHER framework lies in the possibility of obtaining novel biological insights into a trait’s heritability composition. This can be achieved by the application of a huge variety of *in-silico* follow-up techniques to the SGIT and UGITs. The SGIT is of interest by itself, but we also emphasize the importance of the comparison of shared and unshared impacts for each trait. In our real data application, the most remarkable case is BMI in the set of anthropometric traits (see Figure 3C). We found 246 and 1608 significantly enriched gene sets for the SGIT and UGIT of BMI, respectively, with negligible overlapping between them of size 56. By analyzing only BMI, we would have detected only 192 enriched gene sets. By analyzing each of the impacts separately, we dramatically increased the number of observed unique gene sets (1798 in total for both SGI and UGI). It means that each sub-phenotype controlled by the SGF and UGF is less heterogeneous than the original trait. According to the significant gene sets, the UGIT of BMI (see Supplementary Tables 4) controls some structural changes in body compositions and bone formation, while the SGIT is involved in some general signaling pathways and pathways related to nervous system development and probably to general psycho-social aspects of BMI, obesity and other anthropometric traits [19].

Although SHAHER is effective, it has several limitations. First, when trait-trait genetic correlations are weak, it is expected that the contributions of these traits to the shared heritability will be small, too. In this case, MaxSH may overestimate these contributions. Secondly, the framework is applicable only if the number of traits is no less than three. In the case of three traits, the performance is limited and the SHAHER results should be interpreted with caution. We have shown in simulations and real dataset examples that MaxSH works better at higher numbers of genetically correlated traits being analyzed. However, an increase in the number of weakly correlated traits leads to a decrease in the proportion of SNPs associated with all traits simultaneously and to a decrease in the efficiency of the framework. Thirdly, although the set of SNPs identified by the SGIT GWAS is enriched for the SGF, each SNP should be interpreted with caution for whether it is shared or not, because SHAHER has some level of unspecificity. Finally, if any confounding effects were included in the GWAS of the original traits, these effects are amplified in the SGIT [15]. The confounding effects can be controlled easily using special methods like LD score regression [20], although this method fails to distinguish a polygenic component if the trait was measured in the sample with the assortative mating or family effects. Thus, we suggest a thorough check of the original GWAS for the presence of any effects of possible confounders before proceeding to SHAHER.

In conclusion, we propose a novel effective framework for analysis of the shared genetic background for a set of genetically correlated traits using GWAS summary statistics. The framework allows us to obtain novel biological insights into the trait's genetic impact composition. By analyzing shared and unshared genetic impacts separately, we increased the number of identified loci and observed unique gene sets, identified genetic mechanisms being common for all traits or specific for every trait. Of note, sumCOT can be used as a stand-alone method for obtaining GWAS results of the linear combination of the traits using their summary statistics.

Materials and Methods

Simulation study

Under different scenarios, we designed simulations to assess the performance of MaxSH. We (1) assessed the accuracy of *w* estimates, (2) assessed the proportion of SGIT heritability explained by the SGF to the total heritability of the SGIT (the *Q*-value), and (3) compared the analytically predicted total and shared heritabilities of the SGIT and GIP1 with respect to the loss function. The design of our simulation experiment is shown in Figure 3. To generate the input for the MaxSH and GIP approaches, we used a six-parameter simulation model, in which K is the number of traits; W_0^2 is a (K×K) diagonal matrix, where the i-th diagonal element is w_i^2 (the proportion of heritability explained by SGF); s is the proportion of zeros in the matrix U_{unsh}; d_1 is the amplitude of the uniform distribution for non-zero values of U_{unsh} and d_2 is the amplitude of the uniform distribution for U_{env}; H_2 is the diagonal matrix with diagonal elements equal to the trait heritabilities. The parameters values used are given in Figure 3.

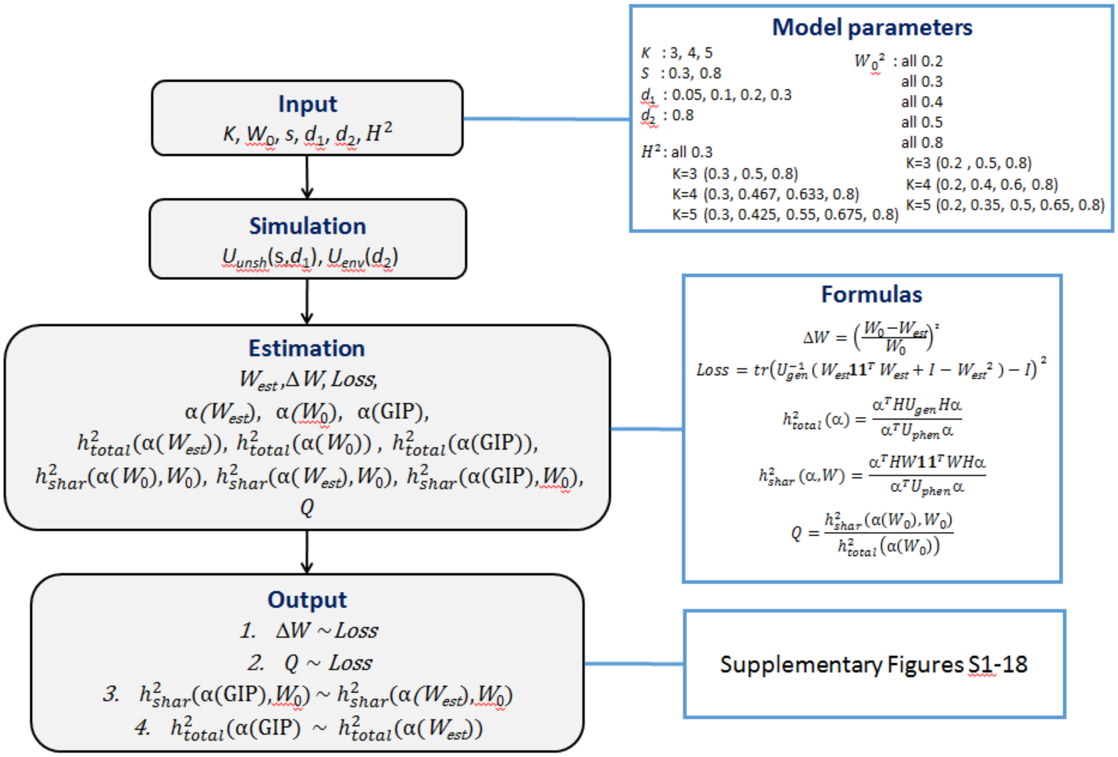
Input
K, W_0, s, d_1, d_2, H^2
Simulation
$U_{unsh}(s, d_1), U_{env}(d_2)$
Estimation
$W_{est}, \Delta W, Loss,$
$\alpha(W_{est}), \alpha(W_0), \alpha(\mathrm{GIP}),$
$h^2_{total}(\alpha(W_{est})), h^2_{total}(\alpha(W_0)), h^2_{total}(\alpha(\mathrm{GIP})),$
$h^2_{shar}(\alpha(W_0), W_0), h^2_{shar}(\alpha(W_{est}), W_0), h^2_{shar}(\alpha(\mathrm{GIP}), W_0),$
Q
Output
1. $\Delta W \sim Loss$
2. $Q \sim Loss$
3. $h^2_{shar}(\alpha(\mathrm{GIP}), W_0) \sim h^2_{shar}(\alpha(W_{est}), W_0)$
4. $h^2_{total}(\alpha(\mathrm{GIP})) \sim h^2_{total}(\alpha(W_{est}))$
Model parameters
K : 3, 4, 5
S : 0.3, 0.8
d_1 : 0.05, 0.1, 0.2, 0.3
d_2 : 0.8
H^2 : all 0.3
K=3 (0.3 , 0.5, 0.8)
K=4 (0.3, 0.467, 0.633, 0.8)
K=5 (0.3, 0.425, 0.55, 0.675, 0.8)
$W_0{}^2$: all 0.2
all 0.3
all 0.4
all 0.5
all 0.8
K=3 (0.2 , 0.5, 0.8)
K=4 (0.2, 0.4, 0.6, 0.8)
K=5 (0.2, 0.35, 0.5, 0.65, 0.8)
Formulas
$\Delta W = \left(\frac{W_0 - W_{est}}{W_0}\right)^2$
$Loss = tr\left(U_{gen}^{-1}(W_{est}\mathbf{11}^T W_{est} + I - W_{est}{}^2) - I\right)^2$
$h^2_{total}(\alpha) = \frac{\alpha^T H U_{gen} H \alpha}{\alpha^T U_{phen} \alpha}$
$h^2_{shar}(\alpha, W) = \frac{\alpha^T H W \mathbf{11}^T W H \alpha}{\alpha^T U_{phen} \alpha}$
$Q = \frac{h^2_{shar}(\alpha(W_0), W_0)}{h^2_{total}(\alpha(W_0))}$
Supplementary Figures S1-18

Figure 3. A schematic depicting the overall workflow of a simulation study. All details are given in the text.

For each fixed number, K, of the original traits and fixed heritability, h_i^2 (i=1,…,K), of each trait, we simulated U_{gen}. To do this, we separately modelled two its components caused by SGF and UGF as $W\mathbf{11}^TW$ and $\sqrt{I-W^2}U_{unsh}\sqrt{I-W^2}$, respectively (see the 'Model' box in Figure M1 of Supplementary Methods). Here **1** is a (K×1) vector of units, and U_{unsh} is a (K×K) matrix randomly generated using the parameters s and d_1 (see Supplementary Methods). Then we randomly generated the trait-trait correlation matrix U_{env} explained by the environmental factors, by giving the parameter d_2 (see Supplementary Methods). Finally, we modeled a matrix of phenotypic correlations by using Model (2) with regard to simulated values $W_{0.}$.

Using simulation data, U_{phen}, U_{gen} and H^2, we estimated W_{est} and calculated its squared relative difference with the simulated values of W_0 (ΔW). We revealed a dependence of ΔW on the loss function ($Loss$). The $Loss$ value characterizes the difference between U_{gen} and the auxiliary matrix V.

Then we estimated α in three ways: (1) using MaxSH and W_0, (2) using MaxSH and W_{est}, and (3) using the GIP method [2]. On the basis of these estimates, we formed three traits being the linear combinations of the original traits. For these combined traits, we calculated the total heritability and the heritability explained by SGF.

The simulated experiments were repeated 10,000 times for each set of parameters. The model parameters and formulas for all calculated values are shown in Figure 3.

Application to real data

Data sets

We used three publicly available real data sets: anthropometric traits, psychiatric conditions and lipid concentrations, which contain five, four and three traits respectively.

The group of anthropometric traits consisted of UK Biobank GWAS summary statistics obtained from the Neale lab (http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank) for people of European ancestry: BMI (N = 336,107), weight (N = 336,227), hip (N = 336,601), waist circumference (N = 336,639) and whole body fat mass (N = 330,762).

The second dataset reflecting psychometric traits was constructed from GWAS results provided by the Psychiatric Genomics Consortium (https://www.med.unc.edu/pgc/download-results/) for bipolar disorder, BIP (N cases = 20,352; N controls = 31,358) [21], major depressive disorder, MDD (N cases = 43,204; N controls = 95,680; without UK Biobank and 23andMe data) [22] and schizophrenia, SCZ (N cases = 36,989; N controls = 113,075). Summary statistics for the fourth trait – subjective well-being (N = 110,935) – were derived from UK Biobank data from the Neale lab. All the psychometric trait GWASs were conducted using samples of white Europeans.

The last dataset corresponding to lipid traits was formed using GWAS data for European participants from the Global Lipid Genetics Consortium (http://csg.sph.umich.edu/willer/public/lipids2013/) for LDL cholesterol (N = 173,082), triglycerides (N = 177,860) and total cholesterol (N = 187,365).

Summary statistics for the three data sets were integrated and quality controlled by the GWAS-MAP platform developed by our group [23]. The GWAS-MAP database contains implemented software for quality control of GWAS results, estimation of phenotypic correlations and LD Score regression (LDSC) [20].

We conducted the quality control of all data and unified them within the GWAS-MAP platform [23]. We filtered all summary statistics by minor allele frequencies ≥ 0.01. Additionally, we filtered GWAS results for BIP by imputation qualities ≥ 0.9. We did not apply this filter to the other traits due to the absence of imputation quality in summary statistics data. Finally, using GWAS-MAP, we performed a correction for genomic control for all traits (including the original traits, SGIT and UGITs) with an LDSC intercept greater than 1 [20]. Thus, we corrected all traits from the psychometric dataset apart from MDD, all original anthropometric traits and their SGIT and lipid SGIT as their LDSC intercept exceeded 1 (see Supplementary Tables 2a-c). Moreover, all SNPs with the p-value equal to 0 were excluded from analysis.

Genetic analysis

Pairwise phenotypic correlations between traits were computed using the GWAS-MAP platform described above. The used method is based on correlations between insignificant *z*-statistics for independent SNPs as previously described in [9]. SNP-based heritability and genetic correlation coefficients were estimated using the LD Score regression software [16] embedded in the GWAS-MAP platform. The significance threshold for genetic correlations was set at 4.5×10^{-4} (0.05/112, where 112 is the number of pairwise combinations between all original traits, their SGIT and UGITs in each dataset - between 11, 9 and 7 traits for anthropometry, psychometric and lipid traits respectively).

SHAHER analysis included checking if there was an SGI or not, the application of MaxSH and conducting SGIT and UGIT GWASs. The threshold for confirming the existence of an SGI at the first stage was empirically set to 0.2.

For each dataset, we visualized the full genetic correlation matrices using the *corrplot*() function from the corrplot R package (v.0.84) [24]. We also placed the SNP-based heritability estimates on the diagonal and crossed out non-significant values.

Finally, we compared the GWAS results obtained for the SGIT by MaxSH and GIP (the principal component analysis on the matrix of genetic covariances)[2].

Gene set and tissue/cell type enrichment analyses

We performed a gene set enrichment analysis and a tissue/cell type enrichment analysis combined with a gene prioritization using the Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) tool v.1.1, release 194 [25]. We selected genome-wide significant SNPs (p-value < 5×10^{-8}) from summary statistics before the genomic control and applied the DEPICT software with default parameters (https://data.broadinstitute.org/mpg/depict/). The MHC region was excluded from analyses.

Next, for the gene set enrichment results, we calculated the number of significant enriched gene sets (FDR < 5%) and constructed an overlapping matrix, in which each cell represents the number of overlapping gene sets for each pair of traits. For each pair of traits, we scaled the number of overlapping gene sets by the minimum number of significant gene sets for this pair of traits. The resulting matrix was visualized using the corrplot R-package as descried above.

The number of original traits associated with SGIT loci

We performed a clumping procedure to search for loci associated with each of the original traits, SGIT and UGITs at a genome-wide significance level of 5×10^{-8}. The associated locus was defined as a genomic region spanning 500 kb in either direction of the lead SNP. Those loci that were significantly associated with SGIT, but not with the original traits, were assumed to be new loci.

We expected that the loci associated with all the original traits used to obtain SGIT are likely to be SGF. To test this expectation, for each dataset we selected all independent loci that were significantly associated with at least one of the original traits and calculated the number of the original traits significantly associated with these loci. For the original anthropometric and lipid traits, we empirically set the significance threshold at p-value = 1×10^{-5}. For the psychometric traits, it was set at 1×10^{-3}. We then analyzed the SGIT p-values for the selected loci and constructed boxplots of $-\log_{10}$ for them with regard to the number of the original traits significantly associated with these loci.

Data Availability

The SHAHER framework is implemented as a set of R/C++ scripts and is freely available at https://github.com/Sodbo/shared_heredity.

Acknowledgements

Funding

The work of GRS was supported by the Russian Foundation for Basic Research (project 20-04-00464). The work of YAT and EEE was supported by the Russian

Foundation for Basic Research (project 19-015-00151). The work of SZS was supported by the Russian Ministry of Education and Science under the 5-100 Excellence Programme and by the Federal Agency of Scientific Organizations via the Institute of Cytology and Genetics (project 0259-2021-0009/AAAA-A17-117092070032-4). The work of PRHJT was supported by the Medical Research Council Human Genetics Unit (MC_UU_00007/10).

Author contributions

YAT, GRS and TIA conceived and oversaw the study. YAT, GRS, SZS and TIA contributed to the design of the study and interpretation of the results. GRS developed the MaxSH method, including the algorithm and program, and conducted simulation studies. PT developed the C++ version of sumCOMB and tested the developed framework. EST, EEE, SGF, SZS and YAT wrote the source code for the framework and performed real data analyses. All co-authors discussed the results and contributed to preparing the draft and final version of the manuscript.

Conflict of interest

P.R.H.J.T. is an employee of BioAge Labs. The remaining authors declare no conflict of interest.

Supplementary Information legend

Supplementary Tables 1a-c: linear combination coefficients and CIs of SGIT for real data sets

Supplementary Tables 2a-c: results of LD score regression for original traits from real data sets

Supplementary Tables 3a-c: clumping results for real data sets

Supplementary Tables 4-6: DEPICT results for real data sets

References

1. Jiang X, Finucane HK, Schumacher FR, Schmit SL, Tyrer JP, Han Y, et al. Shared heritability and functional enrichment across six solid cancers. Nature communications. 2019;10(1):1-23.

2. Tsepilov YA, Freidin MB, Shadrina AS, Sharapov SZ, Elgaeva EE, van Zundert J, et al. Analysis of genetically independent phenotypes identifies shared genetic factors associated with chronic musculoskeletal pain conditions. Communications biology. 2020;3(1):1-13.

3. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, et al. Analysis of heritability and shared heritability based on genome-wide association studies for 13 cancer types. JNCI: Journal of the National Cancer Institute. 2015;107(12).

4. Brainstorm C, Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, et al. Analysis of shared heritability in common disorders of the brain. Science. 2018 Jun 22;360(6395). PubMed PMID: 29930110. Pubmed Central PMCID: PMC6097237.

5. Yang Y, Zhao H, Heath AC, Madden PA, Martin NG, Nyholt DR. Shared genetic factors underlie migraine and depression. Twin Research and Human Genetics. 2016;19(4):341-50.

6. Wright S. Correlation and Causation. JouMal of. Agricultural Research. 1921.

7. Rijsdijk FV, Sham PC. Analytic approaches to twin data using structural equation models. Briefings in bioinformatics. 2002;3(2):119-33.

8. Galesloot TE, Van Steen K, Kiemeney LA, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. PloS one. 2014;9(4):e95923.

9. Stephens M. A unified framework for association analysis with multiple related phenotypes. PloS one. 2013;8(7):e65245.

10. Yang X, Zhang S, Sha Q. Joint analysis of multiple phenotypes in association studies based on cross-validation prediction error. Scientific reports. 2019;9(1):1-10.

11. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature genetics. 2018;50(2):229-37.

12. Fatumo S, Carstensen T, Nashiru O, Gurdasani D, Sandhu M, Kaleebu P. Complimentary methods for multivariate genome-wide association study identify new susceptibility genes for blood cell traits. Frontiers in genetics. 2019;10:334.

13. Ning Z, Tsepilov YA, Sharapov SZ, Grishenko AK, Feng X, Shirali M, et al. Beyond power: multivariate discovery, replication, and interpretation of pleiotropic loci using summary association statistics. bioRxiv. 2019:022269.

14. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics. 2017;18(2):117-27.

15. Timmers PRHJ, Tiys ES, Sakaue S, Akiyama M, Kiiskinen TTJ, Zhou W, et al. Genetically independent phenotype analysis identifies LPA and VCAM1 as drug targets for human ageing. bioRxiv. 2021:2021.01.22.427837.

16. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015 Nov;47(11):1236-41. PubMed PMID: 26414676. Pubmed Central PMCID: PMC4797329.

17. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. Quality control and conduct of genome-wide association meta-analyses. Nature protocols. 2014;9(5):1192-212.

18. Howe LJ, Lawson DJ, Davies NM, Pourcain BS, Lewis SJ, Smith GD, et al. Genetic evidence for assortative mating on alcohol consumption in the UK Biobank. Nature communications. 2019;10(1):1-10.

19. Marcellini F, Giuli C, Papa R, Tirabassi G, Faloia E, Boscaro M, et al. Obesity and body mass index (BMI) in relation to life-style and psycho-social aspects. Archives of gerontology and geriatrics. 2009;49:195-206.

20. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. An atlas of genetic correlations across human diseases and traits. Nature genetics. 2015;47(11):1236.

21. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nature genetics. 2019;51(5):793-803.

22. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature genetics. 2018;50(5):668-81.

23. Gorev D, Shashkova T, Pakhomov E, Torgasheva A, Klaric L, Severinov A, et al., editors. GWAS-MAP: a platform for storage and analysis of the results of thousands of genome-wide association scans. Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS\SB-2018); 2018.

24. Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. corrplot: visualization of a correlation matrix. R package v. 0.84. 2017.

25. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nature communications. 2015;6(1):1-9.