SSGAC Depression, Neuroticism and Subjective Well-being: GWAS and MTAG Scores (Ver 1.0)

Authors: Aysu Okbay a.okbay@vu.nl, Patrick Turley paturley@broadinstitute.org
Date: 12 January 2018

This document accompanies “Turley_et_al_(2018)_PGS_HRS.txt” and describes the construction of polygenic scores for depressive symptoms, subjective well-being, and neuroticism based on Turley et al. [1] for European-ancestry HRS respondents who provided salivary DNA between 2006 and 2008 [2]. If you are using these polygenic scores in your study, please cite:

Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* (2018). doi:10.1038/s41588-017-0009-4

Decription of variables:

HHID	HRS Household Identifier
PN	Person number
PGS_DEP_GWAS	Polygenic score for depression, obtained using standard GWAS results
PGS_NEUR_GWAS	Polygenic score for neuroticism, obtained using standard GWAS results
PGS_SWB_GWAS	Polygenic score for subjective well-being, obtained using standard GWAS results
PGS_DEP_MTAG	Polygenic score for depression, obtained using multivariate analysis of depression, neuroticism and subjective well-being
PGS_NEUR_MTAG	Polygenic score for neuroticism, obtained using multivariate analysis of depression, neuroticism and subjective well-being
PGS_SWB_MTAG	Polygenic score for subjective well-being, obtained using multivariate analysis of depression, neuroticism and subjective well-being
PC1 - PC10	Top 10 principal components (PCs) of the covariance matrix of the individuals’ genotypic data. PCs 1-5 and PCs 6-10 were randomly labelled within each PC set to help reduce identifiability.

Methodology. A polygenic score for an individual is defined as a weighted sum of a person’s genotypes at K SNPs,

$$\hat{g}_i = \sum_{j=1}^{K} x_{ij} w_j \qquad (1)$$

Methodologies for PGS construction differ primarily across two dimensions: how to generate the weights w_j, and how to determine which K SNPs to include [3]. Here, we use LDpred [4], a Bayesian method that includes all measured SNPs and weights each SNP by (an approximation) to its conditional effect, given other SNPs. The theory underlying LDpred is derived assuming the variance-covariance matrix of the genotype data in the training sample is known and assuming some prior effect-size distribution. In practice, the matrix is not known but must be approximated using linkage disequilibrium (LD) patterns from a reference sample. LDpred calculates posterior effect-size distributions for the true effect sizes $\boldsymbol{\beta}$ (i.e., that are conditional on all other SNPs, unlike the GWAS estimates), and each SNP's weight is set equal to the mean of its (conditional) posterior effect-size distribution.

Estimation of LD patterns. We estimated LD patterns using HRS genotype data imputed to the March 2012 release of the 1000 Genomes Phase 1 reference panel [5]. To obtain the LD reference sample, we first converted the genotype probabilities for 21,632,048 variants and 12,454 individuals to hard calls using Plink v1.9 [6]. There were 1,399,136 variants which had kgp-numbers as SNP identifiers instead of rs-numbers. We updated the SNP identifiers of these SNPs to rs-numbers using the crosswalk (HRS_1000G_rsid_map.csv) provided by the HRS [2]. Next, we excluded 3,802 individuals that are not in the list of European-ancestry individuals (hwe_eur_keep.txt) provided by the HRS [7], and 10 individuals that have a genotyping missingness rate greater than 0.02. We restricted the set of genetic variants to 1,216,794 HapMap3 [8] SNPs because these SNPs are generally well-imputed and provide good coverage of the genome in European-ancestry individuals. We dropped 72,349 SNPs with a genotyping call rate less than 0.98. We estimated a genetic relatedness matrix with an additional filter excluding SNPs with minor allele frequency less than 0.01. We dropped one individual from each of the 267 pairs of individuals with a genetic relatedness exceeding 0.025.

In order to make sure that there are no genetic outliers in the sample that can bias the LD estimates, we clustered the remaining 8,375 individuals based on identity-by-state distances in Plink v1.9 [6], again restricting to SNPs with minor allele frequency greater than 0.01. Plink reports a Z score for each individual's IBS distance to his/her closest neighbor. We examined these Z scores and marked an individual as genetic outlier if his/her Z-score was smaller than -5. We dropped these individuals and repeated the process, until no more individuals with a Z score less than -5 remained in the data. The algorithm identified 22 outliers, which were then dropped from the reference data. In the final data, there were 8,353 individuals and 1,144,445 SNPs.

Weights. We provide two polygenic scores for each phenotype based on different sets of summary statistics from Turley et al. [1]: (i) a score based on standard GWAS summary statistics, which are the coefficient estimates from univariate GWAS of subjective well-being, depressive symptoms and neuroticism; and (ii) a score based on MTAG summary statistics, which are obtained from a multivariate analysis of the three phenotypes using the MTAG software tool (see below). In order to avoid overfitting, HRS was excluded from all GWAS discovery samples.

We adjusted the weights for linkage disequilibrium using the LDpred software tool [4] and the reference genotype data whose construction is described above. The LD-adjusted weights were obtained for the SNPs that are available in both the reference data and the (standard GWAS or MTAG) summary statistics for all three traits, and that pass the filters imposed by LDpred: (i) the variant has a minor allele frequency (MAF) greater than 1% in the reference data, (ii) the variant does not have ambiguous nucleotides, (iii) there is no mismatch between nucleotides in the summary statistics and reference data, and (iv) there is no high (>0.15) MAF discrepancy between summary statistics and validation sample. This resulted in 1,018,542 weights for depressive symptoms, 1,018,454 weights for neuroticism, and 1,018,543 weights for subjective well-being. The posterior effect sizes were calculated assuming a fraction of causal SNPs equal to one and setting the LD window to $M/3000$, where M is the number of SNPs included in the analysis.

Polygenic scores. We calculated the scores in Plink v1.9 [6], using genotype probabilities obtained from the 1000 Genomes imputation and the LD-adjusted weights described above for 8652 European-ancestry individuals listed in the "hwe_eur_keep.txt" file provided by the HRS.

MTAG-based polygenic scores. MTAG is a method that uses GWAS summary statistics for a primary phenotype and for one or more secondary phenotypes to produce an updated set of summary statistics for the primary phenotype which, under certain assumptions, will be more precisely estimated than the input GWAS summary statistics.

There are costs and benefits to using an MTAG-based polygenic score. For instance, in all cases, MTAG-based polygenic scores will be more predictive of their corresponding phenotype in expectation. In some cases, however, MTAG can have a high false discovery rate (see Supplementary Note section 1.4 of Turley et al. [1]), which may lead to spurious correlations between the MTAG-based polygenic score and other phenotypes.

We therefore offer the following recommendations. If in a regression, the dependent variable and the polygenic score correspond to the same phenotype, we recommend using the MTAG-based score. If the dependent variable and the polygenic score correspond to different phenotypes, but the coefficient of interest in the regression is not the coefficient associated with the polygenic score (e.g., if the polygenic score is only being used as a control variable in an experimental setting), then we also recommend using the MTAG-based polygenic score. Care should be taken when interpreting the coefficient of an MTAG-based polygenic score in this setting, however, since any observed association may be driven through channels involving the secondary phenotypes. This is especially true when the maxFDR is large (see Turley et al [1] , Supplementary Note section 1.4). If researchers are interested in the coefficient on the polygenic score, they should either use GWAS-based scores, or justify why such channels would lead to negligible bias in their particular case.

Principal components. It is important to take a number of steps to minimize the risk that an observed association between the outcome of interest and the polygenic score is due to unaccounted-for population stratification. A score is stratified if its distribution varies across members of different ancestry groups. Absence to control for differences in ancestry can severely bias estimates of effect sizes, since members of different groups may vary in the outcome of interest for environmental reasons[9]. To reduce such concerns, we recommend controlling for the top 10 ancestry-specific principal components (PCs) of the covariance matrix of the individuals' genotypic data[10], which are included in "Turley_et_al_(2018)_PGS_HRS.txt". The principal components were obtained in Plink v1.9[11] using SNPs with call rate greater than 0.99, minor allele frequency greater than 0.01, and imputation accuracy greater than 0.7. Prior to calculating the principal components, we excluded long-range LD regions on chromosomes 5 (44-51.5 Mb), 6 (25-33.5 Mb), 8 (8-12 Mb) and 11 (45-57 Mb). Remaining SNPs were LD-pruned (R^2<0.1 on a 1000kb window). Following a recommendation from HRS, we have randomly labeled PCs 1-5 and PCs 6-10 within each PC set to help reduce identifiability.

References

1. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* (2018). doi:10.1038/s41588-017-0009-4
2. Health and Retirement Study. *Health and Retirement Study: Candidate Gene and SNP Data Description*. (2014).
3. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–515 (2013).
4. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).
5. Health & Retirement Study, C. Imputation Report - 1000 Genomes Project reference panel. (2012).
6. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).
7. Health and Retirement Study. *Quality Control Report for Genotypic Data*. (2012).
8. Altshuler, D. M., Gibbs, R. A. & Peltonen, L. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
9. Hamer, D. & Sirota, L. Beware the chopsticks gene. *Mol. Psychiatry* **5,** 11–13 (2000).
10. Price, A. L. *et al.* The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet.* **5,** e1000505 (2009).
11. Purcell, S. M. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).