Documentation

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

This document accompanies "Turley_et_al_(2018)_PGS_WLS.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 WLS respondents. 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:

FID	Family identifier
IID	Individual identifier
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

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 [2]. Here, we use LDpred [3], 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 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.

Genotype data and imputation. Genotype data from The Illumina Human Omni Express Bead Chip were available for 9,109 individuals and 713,014 variants. We imputed these genotypes against the HRC v1.1 European reference panel [4] using the Michigan Imputation Server [5]. Prior to imputation, we identified the non-European individuals by plotting the principal components (PCs) of the covariance matrix of the individuals' genotype data[9] together with the PCs of 1000 Genomes populations and visually inspecting the plots. We dropped the identified non-European individuals from the sample. Additionally, we excluded individuals that do not satisfy the following criteria: (i) genotype missingness rate is less than 0.05 in all chromosomes, (ii) there is no mismatch between surveyed sex and genetic sex, (iii) there is no mismatch between surveyed relationship data and genetic relatedness, (iv) the individual is not an outlier in terms of heterozygosity/homozygosity, and (v) the individual is not an ancestral outlier. We also dropped SNPs that have a call rate less than 0.95, Hardy-Weinberg exact test P-value less than 10^{-5}, or minor allele frequency less than 0.01. 607,469 autosomal SNPs and 8,527 individuals remained in the data.

Next, we checked the data against the HRC reference panel[1] for consistency of strand, id names, positions, alleles, reference/alternative allele assignment, and allele frequency differences using version 4.2.5 of the HRC-1000G-check-bim.pl[2] program. The program updates strand, position and reference/alternative allele assignment when possible. It removes a SNP if it has any of the following properties: (i) A/T or G/C alleles and a minor allele frequency greater than 0.4, (ii) alleles that do not match the HRC data, (iii) minor allele frequency discrepancy with the HRC data greater than 0.2, (iv) not available in the HRC data. After all checks, 604,710 SNPs remained which were taken forward for imputation. Genotype probabilities were imputed for 39,127,657 variants and 8,527 individuals.

Estimation of LD patterns. We estimated LD patterns using the imputed WLS genotype data. To obtain the LD reference data, we first converted the genotype probabilities for 38,909,200 biallelic SNPs to hard calls using Plink v1.9 [6]. We restricted the set of genetic variants to 1,211,685 HapMap3 [7] SNPs, because these SNPs are generally well-imputed and provide good coverage of the genome in European-ancestry individuals. Next, we estimated a genetic relatedness matrix, restricting further to SNPs with minor allele frequency greater than 0.01. We dropped one individual from each of the 2,842 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 5,691 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

[1] Site list was downloaded from http://www.haplotype-reference-consortium.org/site
[2] Script available at http://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim.v4.2.5.zip

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 6 outliers, which were then dropped from the data. In the final reference data, there were 5,685 individuals and 1,211,685 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). All weights were obtained from GWAS discovery samples that did not contain the WLS.

We adjusted the weights for linkage disequilibrium using the LDpred software tool [3] 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 1,078,091 weights for all three traits. 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 HRC imputation and the LD-adjusted weights described above.

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 [8]. To reduce such concerns, we recommend controlling for the top 10 principal components (PCs) of the covariance matrix of the individuals' genotypic data[9], which are included in "Turley_et_al_(2018)_PGS_WLS.txt". The principal components were calculated 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 0.6 on the set of conventionally unrelated individuals and then projected onto the rest of the sample. 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).

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. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–515 (2013).
3. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).
4. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–1283 (2016).
5. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48,** 1284–1287 (2016).
6. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).
7. Altshuler, D. M., Gibbs, R. A. & Peltonen, L. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
8. Hamer, D. & Sirota, L. Beware the chopsticks gene. *Mol. Psychiatry* **5,** 11–13 (2000).
9. 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).