

*HRS Polygenic Scores
2006-2010 Genetic Data*

Documentation Report

Erin Ware, University of Michigan
Lauren Schmitz, University of Michigan
Jessica Faul, University of Michigan

Survey Research Center
Institute for Social Research
University of Michigan
Ann Arbor, Michigan

June 2017

Funding

The Health and Retirement Study is funded by a grant from the National Institute on Aging (U01 AG009740) with supplemental support from the Social Security Administration. HRS is conducted by the University of Michigan.

Suggested Citation

Ware, E., Schmitz, L., & Faul, J. (2017). HRS Polygenic Scores 2006-2010 Genetic Data. University of Michigan. <https://hrs.isr.umich.edu/publications/biblio/9201>

Table of Contents

I.	Introduction.....	2
A.	Rationale.....	2
B.	Sample Selection for the Enhanced Face-to-Face Interview.....	2
C.	Consent and Administration Procedures	2
II.	HRS Genomic Data.....	3
A.	PGS Construction.....	3
B.	Sources for SNP weights.....	4
C.	Notes about the use of PGS.....	4
D.	Educational Attainment.....	6
E.	Height	7
F.	Body Mass Index (BMI).....	8
G.	Waist Circumference and Waist-to-Hip Ratio	9
H.	Blood Pressure: Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), Pulse Pressure (PP), and Mean Arterial Pressure (MAP).....	11
I.	Alzheimer’s Disease.....	14
J.	General Cognition.....	15
K.	Schizophrenia	16
L.	Smoking Initiation.....	17
M.	Subjective Wellbeing	18
N.	Neuroticism	19
O.	Depressive Symptoms	20
III.	Setup.....	21
A.	Distribution Set.....	21
	Table 1: Contents of the Distribution Package.....	21
B.	Program Statements.....	22
1.	Using the Files with SAS.....	22
2.	Using the Files with SPSS	22
3.	Using the Files with Stata	22
C.	Non-Windows Environments.....	22
1.	SPSS in an OSX environment	22
2.	STATA in an OS X Environment.....	23
IV.	If You Need to Know More	24
A.	HRS Internet Site	24
B.	Contact Information	24

I. Introduction

This guide describes the construction of polygenic scores (PGSs) for a variety of phenotypes for HRS respondents who provided salivary DNA between 2006 and 2010. These scores serve as an attempt to harmonize research across studies and facilitate use among HRS data users by making these scores available publicly. PGSs for each phenotype are based on a single, replicated genome-wide association study (GWAS). These scores will be updated as sufficiently large GWAS are published for new phenotypes or as meta-analyses for existing phenotypes are updated. This document describes the general method of construction with details on each phenotype included as appendices.

A. Rationale

Complex health outcomes or behaviors of interest to the research community are often highly polygenic, or reflect the aggregate effect of many different genes so the use of single genetic variants or candidate genes may not capture the dynamic nature of more complex phenotypes. A PGS aggregates thousands to millions of individual loci across the human genome and weights them by effect sizes, often derived from a GWAS, as an estimate of the strength of their association to produce a single quantitative measure of genetic risk and to increase power in genetic analyses.

B. Sample Selection for the Enhanced Face-to-Face Interview

In 2006, HRS initiated what is referred to as an Enhanced Face-to-Face Interview (EFTF). In addition to the core interview, the EFTF interview includes a set of physical performance tests, anthropometric measurements, blood and saliva samples, and self-administered questionnaire on psychosocial topics. Approximately fifty percent of households with at least one living respondent were selected for the EFTF interview across all primary sampling units (PSUs). A random one half of the 2006 sample was preselected to the EFTF interview. The other half was selected in 2008. Similarly, new cohort households for 2010 were randomly assigned into one of these two groups with EFTF data collection beginning in 2010 or 2012. The sample was selected at the household-level to ensure that the same request was made to both members of a household. New spouses of respondents flagged to complete an EFTF interview were also asked to do so. Thus, in coupled households, both members of the couple were selected. Some respondents who were selected for the enhanced face-to-face sample were not asked to complete the physical measures or biomarkers. This group included respondents who a) needed to be interviewed by proxy, b) resided in a nursing home, or c) declined a face-to-face interview but agreed to be interviewed by telephone. The preload variable that identifies the EFTF sample is KX090_R for 2006, LX090_R for 2008, and MX090_R for 2010 (located in the respondent preload file for each wave), for which a value of 3 indicates that the respondent was assigned to the EFTF sample for that wave. The variable EFTFASSIGN in the Cross-wave Tracker file indicates the respondents' permanent assignment for enhanced face-to-face rotation for 2006 and beyond.

C. Consent and Administration Procedures

Prior to administering the saliva collection, a consent form was administered by the interviewer. Respondents were asked to read and sign the form. Respondents who did not sign the consent form were not asked to complete the collection. Respondents were instructed not to eat, drink, smoke, chew gum or brush their teeth during this component of the interview (at least 30 mins prior to saliva collection). In 2006, saliva was collected using a mouthwash collection method. In 2008 and beyond, the data collection method switched to the Oragene DNA Collection Kit (OG-250).¹

¹ More information on the saliva collection protocol can be found here:

II. HRS Genomic Data

Genotyping was conducted by the Center for Inherited Disease Research (CIDR) in 2011, 2012, and 2015 (RC2 AG0336495 and RC4 AG039029). Genotype data on over 15,000 HRS participants was obtained using the Illumina HumanOmni2.5 BeadChips (HumanOmni2.5-4v1, HumanOmni2.5-8v1), which measures ~2.4 million SNPs. Individuals with missing call rates >2%, SNPs with call rates <98%, HWE p-value < 0.0001, chromosomal anomalies, and first degree relatives in the HRS were removed. For more information on the genotype data and quality control process see [the QC Report](#).

Imputation to the 1000 Genomes Project cosmopolitan reference panel phase 3 version 5 (initial release on May 2013, haplotypes released Oct 2014) was performed by the University of Washington Genetics Coordinating Center (GCC) using IMPUTE2 (2), with phasing performed using SHAPEIT2. Overall, ~21 million SNPs were imputed from the original 1,905,968 SNPs that were genotyped and passed quality control. Masking of genotyped SNPs to assess the accuracy of imputation was performed to estimate the median concordance between actual and imputed genotypes (median concordance > 0.995), and additional quality control metrics indicate high quality imputation. A copy of the [imputation report](#) is available on the HRS Web site.

Principal component (PC) analysis was performed to identify population group outliers and to provide sample eigenvectors as covariates in the statistical model used for association testing to adjust for possible population stratification. SNPs used for PC analysis were selected by linkage disequilibrium (LD) pruning from an initial pool consisting of all autosomal SNPs with a missing call rate < 5% and minor allele frequency (MAF) > 5%, and excluding any SNPs with a discordance between HapMap controls genotyped along with the study samples and those in the external HapMap data set. In addition, the 2q21 (LCT), HLA, 8p23, and 17q21.31 regions were excluded from the initial pool. Genetic ancestry in HRS was identified through PC analysis on genome-wide SNPs calculated across all participants using the aforementioned filtering criteria.

The final European American sample includes all self-reported non-Hispanic whites that had PC loadings within \pm one standard deviations of the mean for eigenvectors 1 and 2 in the PC analysis of all unrelated study subjects. The final African American sample includes all self-reported African Americans within two standard deviations of the mean of all self-identified African Americans for eigenvector 1 and \pm one standard deviation of the mean for eigenvector 2 in the PC analysis of all unrelated study subjects. Once ancestry-specific analysis samples were identified (n=2,279 non-Hispanic Black, n=9,991 non-Hispanic White), PCA was run again within each sample to create sample eigenvectors for covariates in the statistical model used for association testing to adjust for possible population stratification. These are referred to as “ancestry-specific PCs”.

A. PGS Construction

While conceptually simple, there are numerous ways to estimate PGSs, not all achieving the same end goals. We systematically investigated the impact of four key decisions in the building of PGSs from published genome-wide association meta-analysis results: 1) whether to use single nucleotide polymorphisms (SNPs) assessed by imputation, 2) criteria for selecting which SNPs to include in the score, 3) whether to account for linkage disequilibrium (LD), and 4) if accounting for LD, which type of method best captures the correlation structure

Crimmins EM, Faul JD, Kim JK, et al. *Documentation of Biomarkers in the 2006 and 2008 Health and Retirement Study*. Ann Arbor, Michigan: Institute for Social Research, University of Michigan; 2013.

Crimmins EM, Faul JD, Kim JK, Weir DR. *Documentation of Biomarkers in the 2010 and 2012 Health and Retirement Study*. Ann Arbor, Michigan: Survey Research Center, University of Michigan; 2015.

among SNPs (i.e. clumping vs. pruning). Using the Health and Retirement Study (HRS) we examined the predictive ability as well as the variability and co-variability in PGSs arising from these different estimation approaches.²

Overall, results from these analyses concluded that including all available SNPs in a PGS (i.e. not accounting for any LD or p-value thresholding) either demonstrated the largest predictive power (incremental R^2) of the score or produced a score that did not differ significantly from scores with similar predictive power that employed some degree of LD trimming or p-value thresholding. Thus, the HRS has chosen to provide scores that include all available SNPs in the PGS that overlap between the GWAS meta-analysis and the HRS genetic data.

Weighted sums were chosen to calculate the PGSs. Weights were defined by the odds ratio or beta estimate from the GWAS meta-analysis files corresponding to the phenotype of interest. If the beta value from the GWAS meta-analysis was negative (or the odds ratio (OR) < 1), the beta/OR measures were converted to positive values (OR > 1) and the reference allele flipped to represent phenotype-increasing PGSs. PGSs are calculated using the following formula:

$$PGS_i = \sum_{j=1}^J W_j G_{ij}$$

where i is individual i ($i=1$ to N), j is SNP j ($j=1$ to J), W is the meta-analysis effect size for SNP j and G is the genotype, or the number of reference alleles (zero, one, or two), for individual i at SNP j . Due to the long-range linkage disequilibrium in this region, making linkage equilibrium difficult to obtain, the MHC region on chromosome 6 (26-33Mb) was omitted from all PGSs. Missing data was imputed within ancestry using the expected genotyped given the allele frequency. Scores were similar when not employing the missing data imputation default. PGSs were calculated using PRSice and PLINK.

B. Sources for SNP weights

To incorporate externally valid SNP weights from replicated GWAS, we performed a search of the literature to identify large GWAS meta-analysis studies related to the selected phenotype. Where possible, meta-analyses that did not include the HRS in the discovery analysis were selected to be independent of our data. If the HRS was included in the analyses, we requested that the consortia repeat the analysis with the HRS removed. All base SNP files from GWAS meta-analyses were converted to NCBI build 37 annotation for compatibility with HRS SNP data.

C. Notes about the use of PGS

PGSs are released for both the European ancestry and African ancestry groups, separately. However, it should be noted that the GWAS used to inform the SNP weights come from GWAS on European ancestry groups and as a result, PGSs for other ancestry groups may not have the same predictive capacity (Martin et al. 2017; Ware et al. 2017).

Ancestry specific PCs 1-10 are included for each group. PCs 1-5 and PCs 6-10 are randomly labeled within each PC set to help reduce identifiability. To control for confounding from population stratification, or to account for any ancestry differences in genetic structures within populations that could bias estimates, **we highly recommend that users perform analyses separately by ancestral group and, at the very least, adjust for PCs 1-5** (Price et al., 2006). The PCs control for any genetic aspects of common ancestry that could be spuriously correlated with the PGS and the outcome of interest.

² For additional information on this analysis, see:

Ware EB, Schmitz LL, Faul JD, Gard AM, Smith JA, Mitchell CM, Weir DR, Kardina SLR. (2017) *Method of Construction Affects Polygenic Score Prediction of Common Human Traits*. BiorXiv. 2017 doi: <https://doi.org/10.1101/106062>

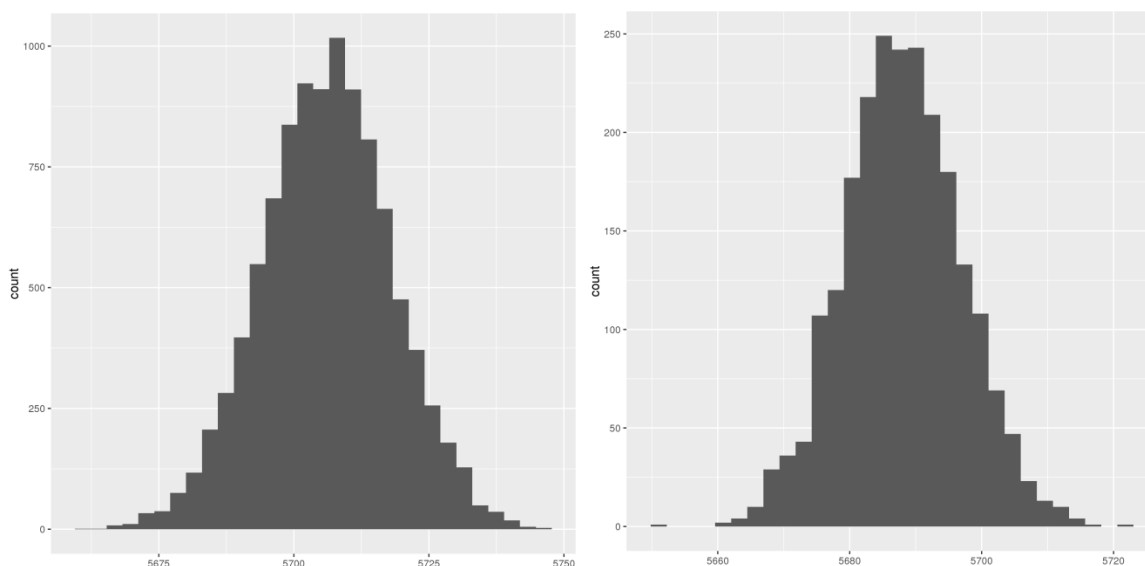
References

- Euesden J, Lewis CM, & O'Reilly PF (2015) PRSice: Polygenic Risk Score software. *Bioinformatics* 31(9):1466-1468.
- Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., ... & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4), 635-649.
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8), 904-909.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559-575.
- Ware EB, Schmitz LL, Faul JD, Gard AM, Smith JA, Mitchell CM, Weir DR, Kardina SLR. (2017) Method of Construction Affects Polygenic Score Prediction of Common Human Traits. BiorXiv. 2017 doi: <https://doi.org/10.1101/106062>
- Ware EB, Schmitz LL, Faul JD, Gard AM, Smith JA, Mitchell CM, Weir DR, Kardina SLR. (2017). Method of Construction Affects Polygenic Score Prediction of Common Human Traits. BiorXiv. 2017 doi: <https://doi.org/10.1101/106062>

D. Educational Attainment

The educational attainment PGSs were created using results from a 2016 study by the Social Science Genetic Association Consortium (SSGAC). The meta-analysis included 293,723 individuals in the discovery sample and 111,349 in replication. Genome-wide significant SNPs were identified in 74 loci (**Supplementary Information section 1.6.1**¹). Approximately 9.3 million SNPs were included in the analyses, with all cohorts utilizing SNPs imputed to the 1000 genomes reference panel (1000G). The original GWAS included the HRS. To compute the PGSs for HRS respondents, the SSGAC provided SNP weights with the HRS and 23andMe results (due to data use agreements) removed (combined discovery + replication sample size without HRS: N=395,109). Study-specific GWASs controlled for the first ten principal components of the genotypic data, a third-order polynomial in age, an indicator for being female, interactions between age and female, and study-specific controls, including dummy variables for major events such as wars or policy changes that may have affected access to education in their specific sample. Mixed-sex cohorts were additionally asked to upload separate regression results for men and women. The European ancestry PGSs contain 1,329,385 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 1,325,831 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	5661.29	5746.66	85.37	5706.31	5706.11	0.12
African ancestry	2279	5650.52	5721.28	70.76	5687.85	5688.00	0.19



European ancestry: Distribution of PGS_EDU2_SSGAC16 African ancestry: Distribution of PGS_EDU2_SSGAC16

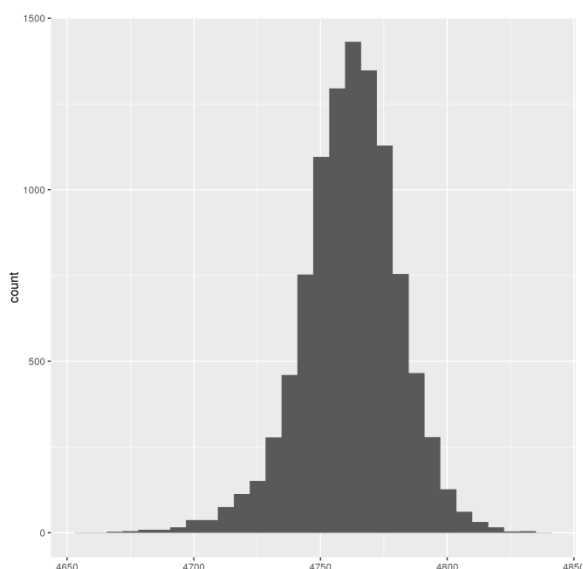
References

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., ... & Oskarsson, S. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604), 539.

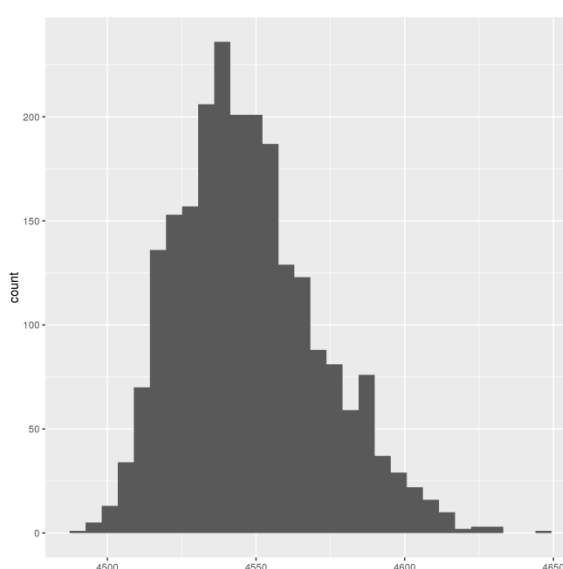
E. Height

PGSs for height were created using results from a 2014 study conducted by the Genetic Investigation of Anthropometric Traits (GIANT) consortium. The GWAS meta-analysis files are publicly available on their data download page: https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files. The GIANT height meta-analysis included 253,288 individuals from 79 studies imputed to HapMap II with a total of 2,550,858 SNPs. Replication was performed in a sample of 80,067 individuals. Height was measured as sex standardized height and participating studies adjusted for age and genetic principal components in their GWASs. There were 697 genome-wide significant SNPs reported (**Supplementary Table 1**). The European ancestry PGSs contain 814,243 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 791,321 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	4657.28	4839.26	182	4762.7	4761.8	0.193093
African ancestry	2279	4491.23	4647.92	156.7	4544.6	4547.6	0.499855



European Ancestry: Distribution of PGS_Height_GIANT14



African Ancestry: Distribution of PGS_Height_GIANT14

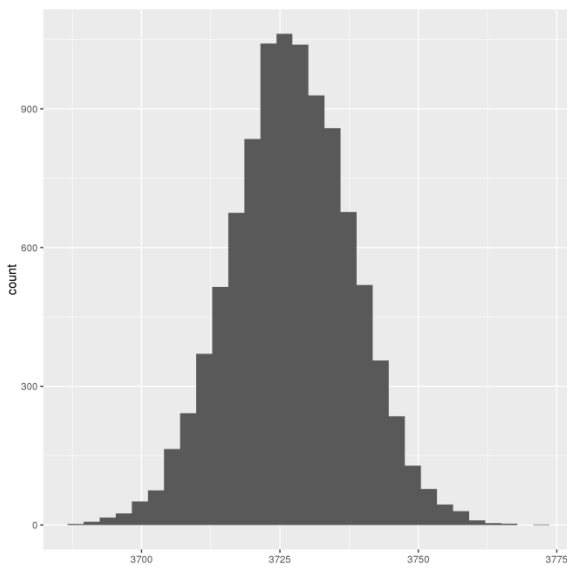
References

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., ... & Amin, N. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11), 1173-1186.

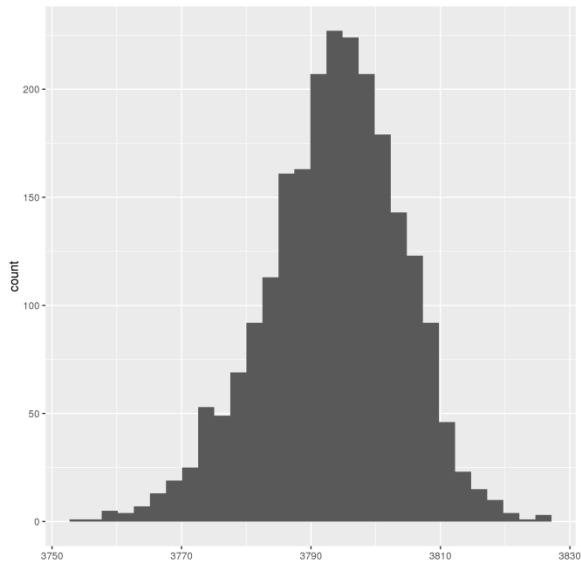
F. Body Mass Index (BMI)

PGSs for BMI were created using results from a 2015 GWAS conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium. The GWAS meta-analysis files are publicly available on their data download page: (https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files) GWAS meta-analysis was performed on a sample of 234,069 individuals from 80 studies across 2,550,021 SNPs, and separately in a Metabochip (MC) meta-analysis on a sample of 88,137 individuals from 34 studies across 156,997 SNPs. A joint GWAS and MC meta-analysis was then conducted on 332,154 individuals across 2,554,623 SNPs. Adjustment covariates within each contributing cohort GWAS included age, age², sex, and genetic principal components. A total of 97 SNPs were reported as genome-wide significant (**Table 1 and 2 and Extended Data Table 2**). The European ancestry PGSs contain 774,728 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 779,542 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	3686.91	3771.06	84.15	3727.1	3727.1	0.109833
African ancestry	2279	3754.41	3826.41	72	3794.1	3793.5	0.221635



European ancestry: Distribution of PGS_BMI_GIANT15



African ancestry: Distribution of PGS_BMI_GIANT15

References

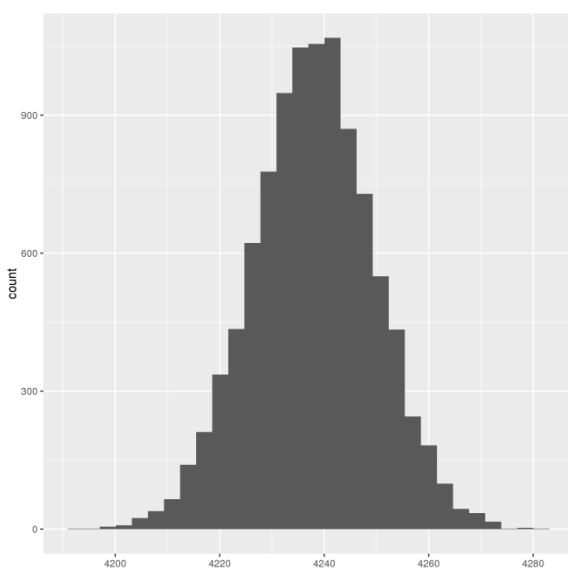
Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., ... & Croteau-chonka, D. C. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538), 197-206.

G. Waist Circumference and Waist-to-Hip Ratio

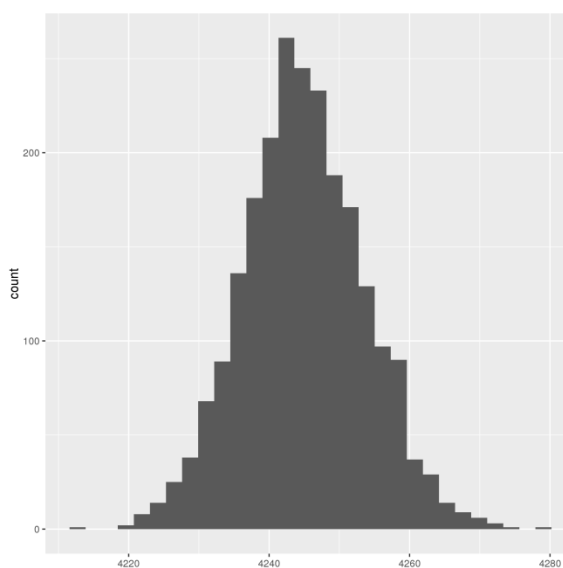
PGSs for waist circumference (WC) and waist-to-hip ratio (WHR) were created using results from a 2015 study conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium. The GWAS meta-analysis files are publicly available on their data download page:

https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files GWAS meta-analysis was performed on a sample of 142,762 individuals from 57 studies across 2,507,022 SNPs, and separately in a Metabochip (MC) meta-analysis on a sample of 67,326 individuals from 44 studies across 124,196 SNPs. A joint GWAS and MC meta-analysis was then conducted on 210,088 individuals across 93,057 SNPs. The GWAS identified 49 loci associated with WHR and an additional 19 loci associated with WC at the genome-wide significance level (**Table 1**). Association analyses adjusted for age, age², study-specific covariates if necessary, and BMI. The European ancestry PGSs for WC contain 778,121 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 776,945 SNPs. Similarly, the European ancestry PGSs for WHR contain 777,158 SNPs and African ancestry PGSs for WHR contain 774,709 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

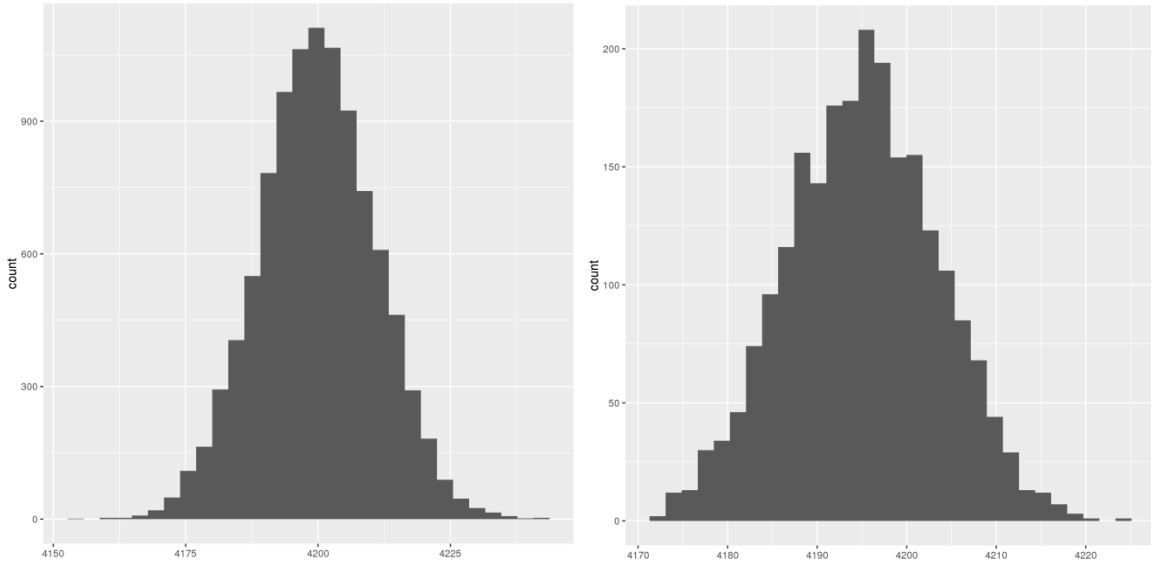
Waist circumference	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	4192.10	4281.30	89.20	4238.00	4237.89	0.12
African ancestry	2279	4212.86	4279.22	66.36	4244.65	4244.94	0.18
Waist:Hip ratio	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	4154.69	4242.69	88.00	4199.70	4199.84	0.11
African ancestry	2279	4171.91	4223.95	52.04	4195.22	4195.05	0.18



European ancestry: Distribution of PGS_WC_GIANT15



African ancestry: Distribution of PGS_WC_GIANT15



European ancestry: Distribution of PGS_WHR_GIANT15 African ancestry: Distribution of PGS_WHR_GIANT15

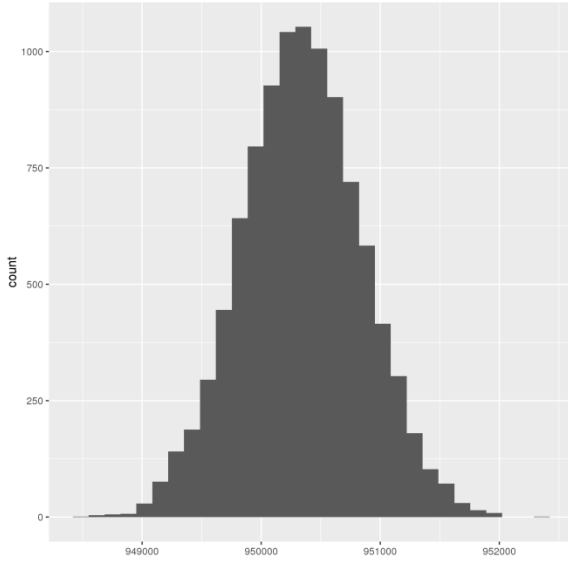
References

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., ... & Workalemahu, T. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538), 187.

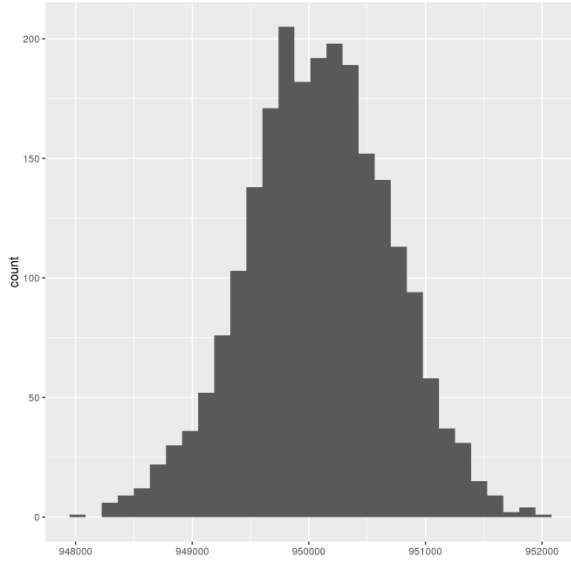
H. Blood Pressure: Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), Pulse Pressure (PP), and Mean Arterial Pressure (MAP)

The PGSs for DBP, SBP, PP, and MAP were created using results from two separate 2011 GWASs conducted by the Internal Consortium for Blood Pressure (ICBP). The GWAS meta-analysis files are publicly available on the dbGaP website: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000585.v1.p1. For DBP and SBP, primary analyses evaluated associations between 2.5 million SNPs in 69,395 individuals of European ancestry from 29 studies. Replication analysis was conducted in a follow-up sample of 133,661 individuals. Twenty-nine independent SNPs were reported as having genome-wide significant associations with DBP, SBP, or both in the discovery and replication samples (**Table 1**). For PP and MAP, primary analysis evaluated 74,064 individuals and replication analysis evaluated 48,607 individuals of European descent from 35 studies. Eight SNPs reached genome-wide significance with PP, MAP, or both (**Table 1**). For all four phenotypes, age, age², BMI, and sex were included as covariates in association analyses. In the presence of anti-hypertensive medication, the BP was corrected by adding a constant (15mmHg for SBP and 10mmHg for DBP). The European ancestry DBP PGSs contain 808,751 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis and African ancestry PGSs for DBP contain 808,545 SNPs; SBP PGSs contain 803,755 SNPs for European ancestry individuals and 803,700 SNPs for African ancestry individuals; PP PGSs contain 790,979 SNPs for European ancestry individuals and 790,664 SNPs for African ancestry individuals; MAP PGSs contain 791,334 SNPs for European ancestry individuals and 790,921 SNPs for African ancestry individuals. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

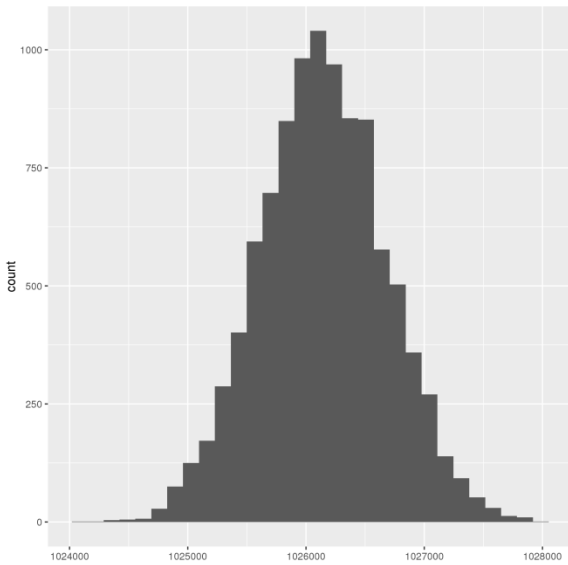
	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
Diastolic Blood Pressure							
European ancestry	9991	948456.00	952329.00	3873.00	950337.00	950340.78	5.04
African ancestry	2279	948046.00	952042.00	3996.00	950081.00	950079.85	12.92
Systolic Blood Pressure							
European ancestry	9991	1024070.00	1027970.00	3900.00	1026140.00	1026142.68	5.31
African ancestry	2279	1023600.00	1028910.00	5310.00	1026530.00	1026515.87	16.34
Pulse Pressure							
European ancestry	9991	897243.00	901075.00	3832.00	899208.00	899213.19	5.01
African ancestry	2279	896436.00	900801.00	4365.00	898909.00	898905.21	12.16
Mean Arterial Pressure							
European ancestry	9991	895966.00	899811.00	3845.00	897898.00	897904.16	5.00
African ancestry	2279	895293.00	899247.00	3954.00	897212.00	897214.15	11.88



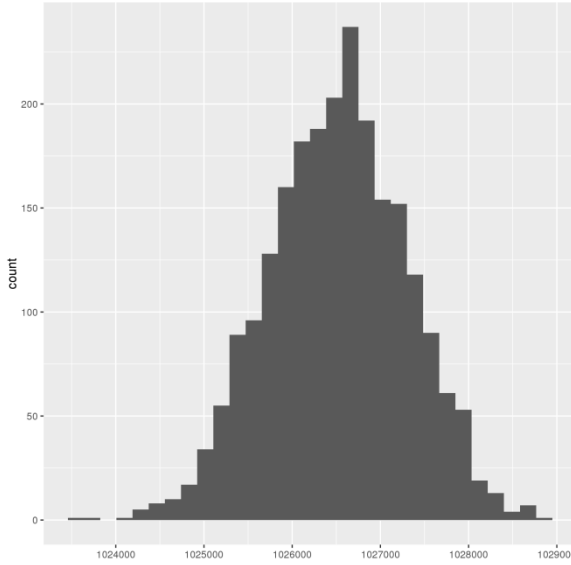
European ancestry: Distribution of PGS_DBP_ICBP11



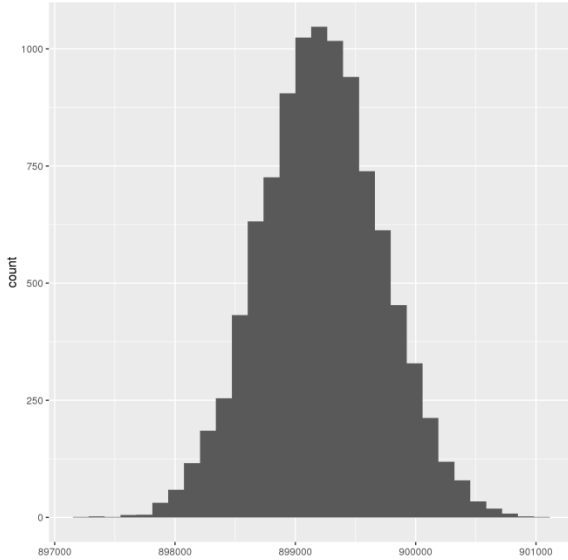
African ancestry: Distribution of PGS_DBP_ICBP11



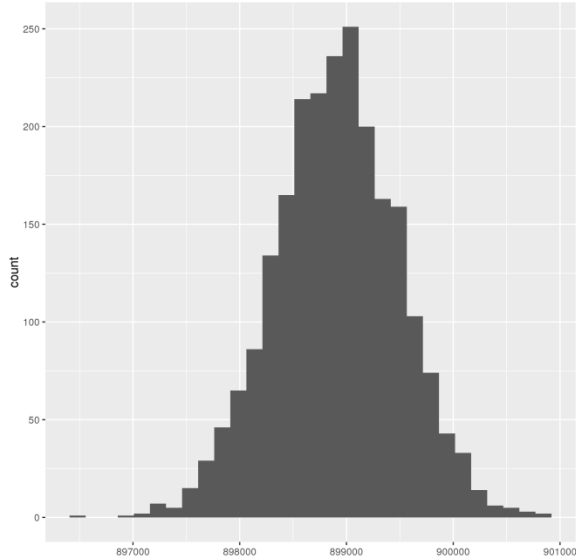
European ancestry: Distribution of PGS_SBP_ICBP11



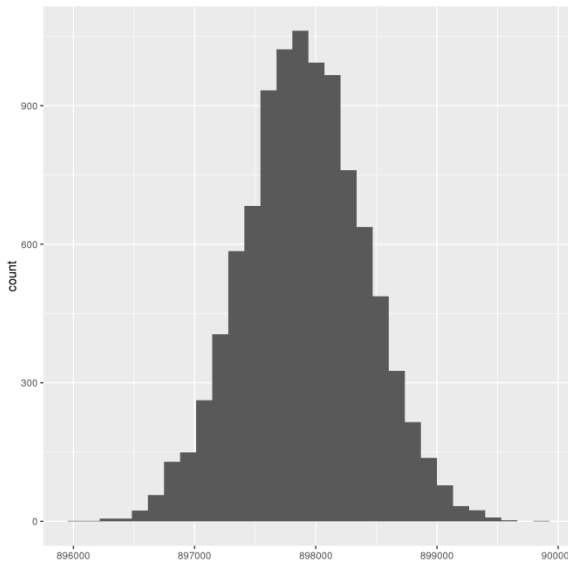
African ancestry: Distribution of PGS_SBP_ICBP11



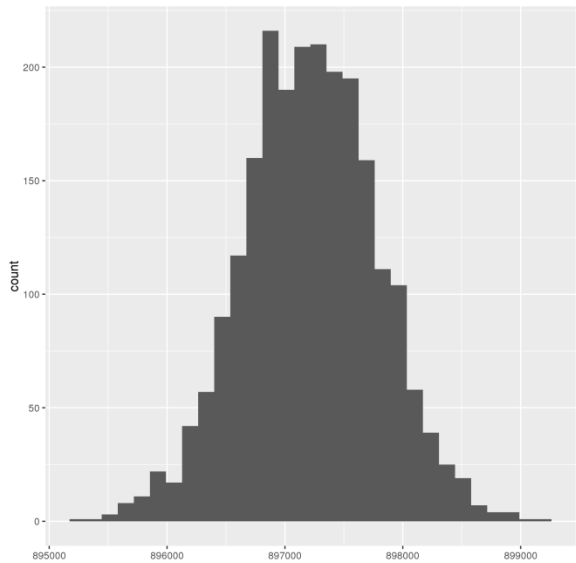
European ancestry: Distribution of PGS_PP_ICBP11



African ancestry: Distribution of PGS_PP_ICBP11



European ancestry: Distribution of PGS_MAP_ICBP11



African ancestry: Distribution of PGS_MAP_ICBP11

References

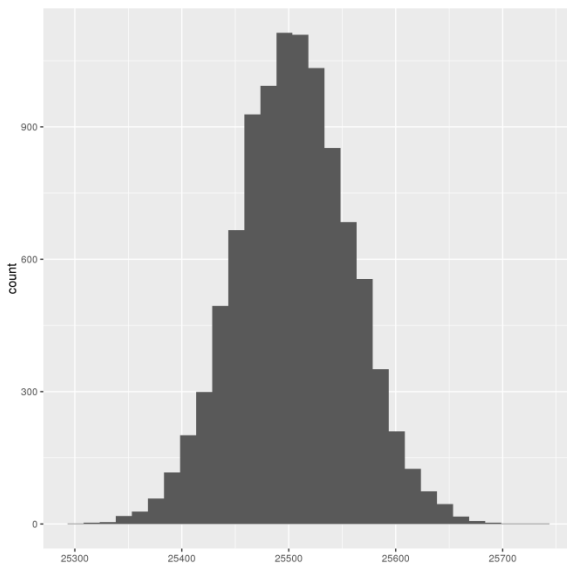
Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., ... & Okamura, T. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, *478*(7367), 103-109.

Wain, L. V., Verwoert, G. C., O'Reilly, P. F., Shi, G., Johnson, T., Johnson, A. D., ... & Ehret, G. B. (2011). Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nature Genetics*, *43*(10), 1005-1011.

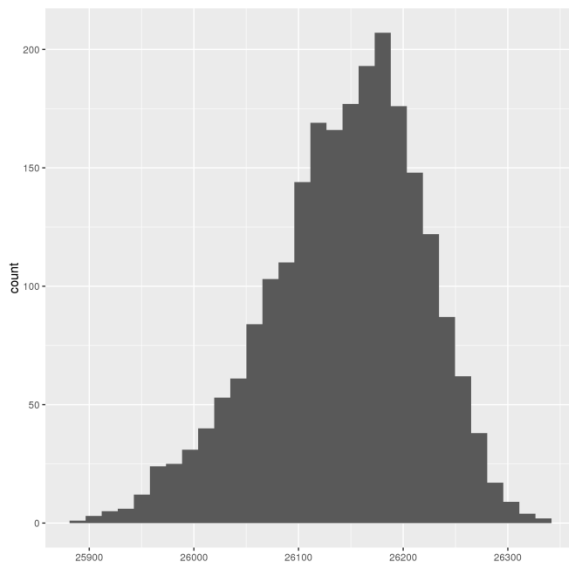
I. Alzheimer's Disease

The PGSs for Alzheimer's disease (AD) were created using results from a 2013 GWAS conducted by the International Genomics of Alzheimer's Project (IGAP): http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php. A GWAS meta-analysis of AD was conducted across 20 independent studies using data from four international consortia: Alzheimer's Disease Genetic Consortium (ADGC), the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, the European Alzheimer's Disease Initiative (EADI), and the Genetic and Environmental Risk in Alzheimer's Disease (GERAD) Consortium. The stage 1 meta-analysis included 54,162 participants ($N_{\text{cases}}=17,008$ and $N_{\text{controls}}=37,154$) of European descent with a total of 7,055,881 SNPs imputed to 1000 Genomes (2010 release). The stage 2 replication sample included 19,884 participants of European ancestry ($N_{\text{cases}}=8,572$ and $N_{\text{controls}}=11,312$) with a total of 11,632 genotyped SNPs. In addition to the *APOE* locus (encoding apolipoprotein E), the two-stage combined discovery and replication GWAS revealed 19 SNPs that reached genome-wide significant associations with AD (**Table 2**). Adjustment covariates within each contributing cohort included age, sex, and genetic principal components. The European ancestry PGSs contain 1,163,744 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 1,177,015 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	25293.4	25729.1	435.7	25504	25505	0.536299
African ancestry	2279	25891.6	26336.9	445.3	26152	26143	1.551465



European Ancestry: Distribution of PGS_AD_IGAP13



African Ancestry: Distribution of PGS_AD_IGAP13

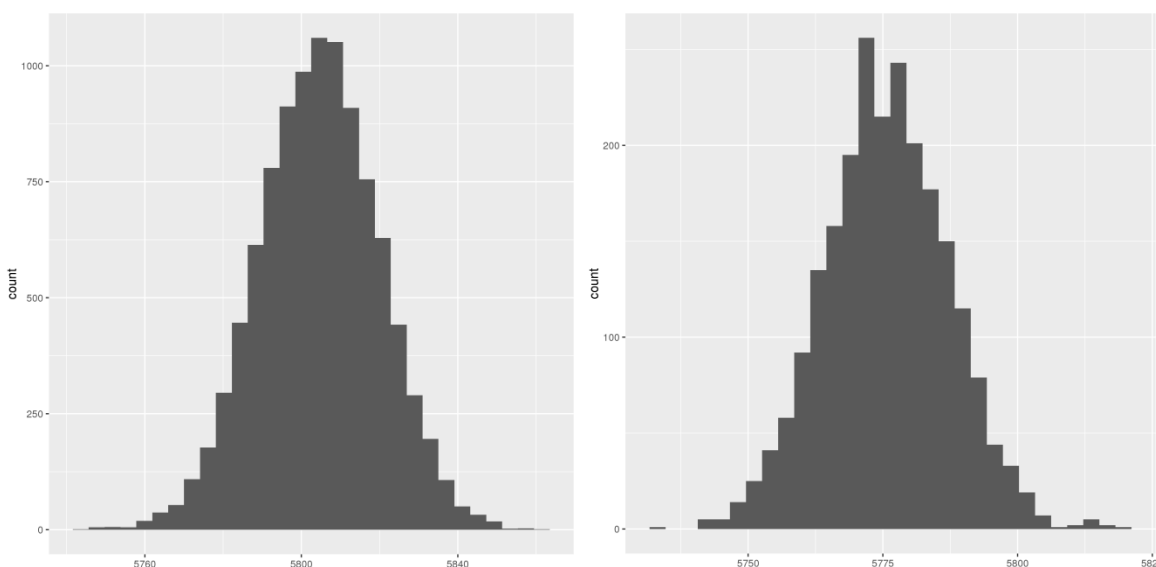
References

Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet.* 2013;45(12):1452-1458.

J. General Cognition

The PGSs for general cognition were created using results from a 2015 GWAS conducted across 31 cohorts by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. A total of 53,949 participants undertook multiple, diverse cognitive tests from which a general cognitive function phenotype was created within each cohort by principal component analysis. Thirteen genome-wide significant SNPs in three separate regions previously associated with neuropsychiatric phenotypes were reported (**Supplementary Table S3**). The original GWAS included the HRS. To compute the PGSs for HRS respondents, weights were provided by the CHARGE consortium from a meta-analysis that excluded the HRS. Adjustments for age, sex, and population stratification were included in study-specific GWAS association analyses. Cohort-specific covariates—for example, site or familial relationships—were also fitted as required. A total of 2,478,500 SNPs were included in the meta-analysis. The European ancestry PGSs contain 771,570 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 766,869 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	5744.18	5862.06	117.9	5804.7	5804.5	0.153097
African ancestry	2279	5732.25	5818.71	86.46	5775.6	5775.7	0.241321



European ancestry: Dist. of PGS_GenCog_CHARGE15 African ancestry: Dist. of PGS_GenCog_CHARGE15

References

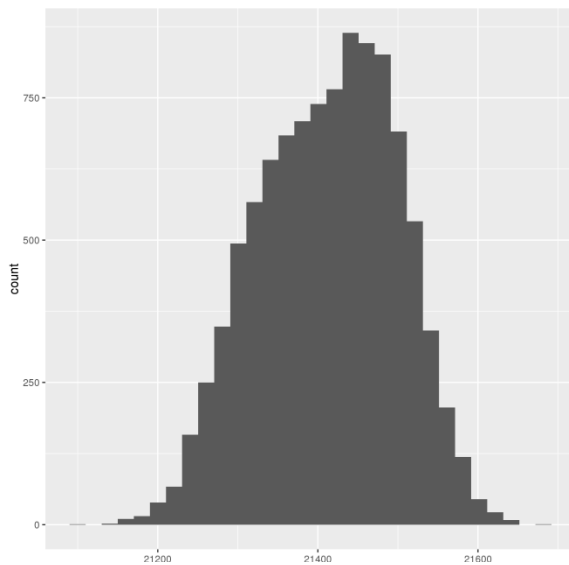
Davies, G., Armstrong, N., Bis, J. C., Bressler, J., Chouraki, V., Giddaluru, S., ... & Van Der Lee, S. J. (2015). Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N= 53,949). *Molecular Psychiatry*, 20(2), 183.

K. Schizophrenia

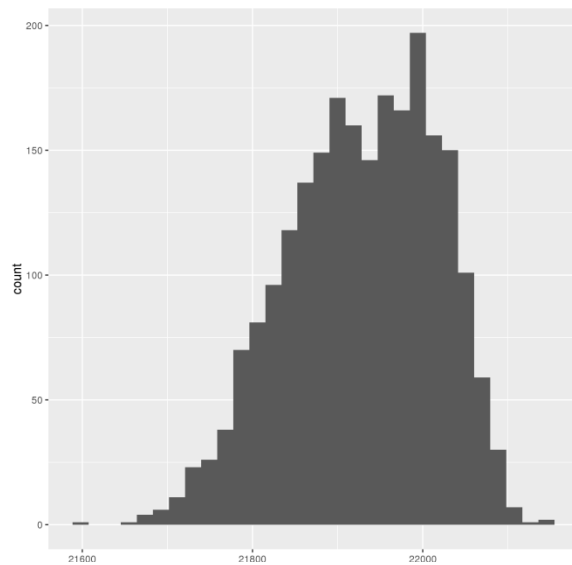
The PGSs for schizophrenia were created using results from a 2014 GWAS conducted by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC). The GWAS meta-analysis files are publicly available on their data download page: <https://www.med.unc.edu/pgc/results-and-downloads>. The schizophrenia GWAS combined meta-analysis included 36,989 cases and 113,075 controls (N=152,805) and identified 128 loci that meet genome-wide significance (**Supplementary Table 2**). The replication sample consisted of 1,513 cases and 66,236 controls. After quality control, around 9.5 million SNPs were analyzed. To enable acquisition of large samples, some of the participating groups ascertained cases via clinician diagnosis rather than a research-based assessment. Genetic principal components and study identifiers were included as covariates.

The European ancestry PGSs contain 1,263,949 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 1,279,781 SNPs.

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	21098.80	21681.00	582.20	21418.00	21412.33	0.86
African ancestry	2279	21599.70	22147.20	547.50	21934.70	21928.33	1.84



European Ancestry: Distribution of PGS_SCZ_PGC09



African Ancestry: Distribution of PGS_SCZ_PGC09

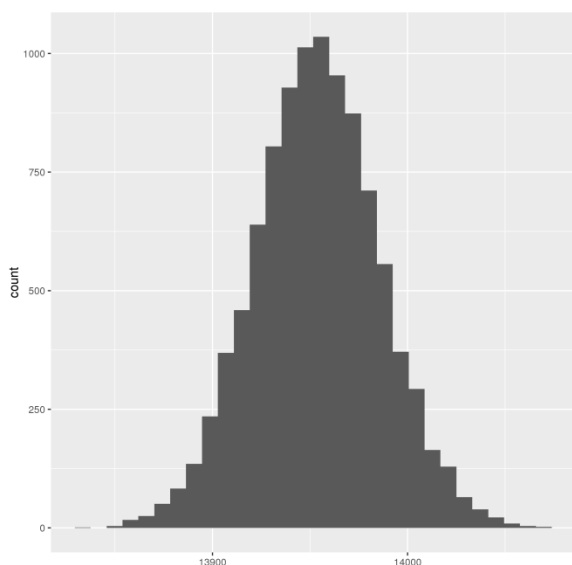
References

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K. H., Holmans, P. A., ... & Pers, T. H. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421-427.

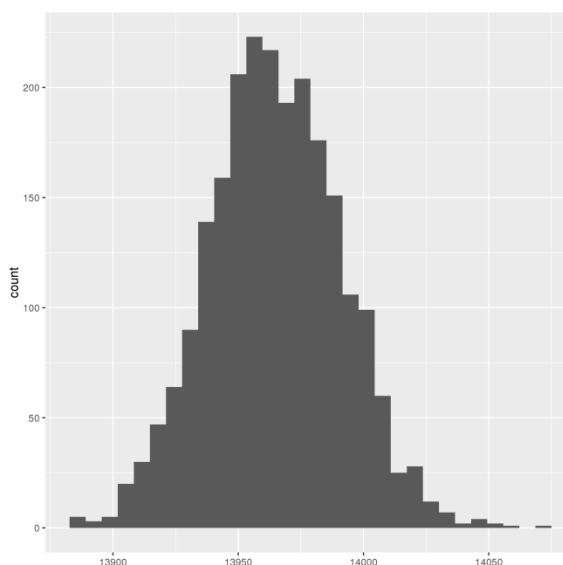
L. Smoking Initiation

The PGSs for smoking initiation were created using results from a 2010 GWAS conducted by the Tobacco and Genetics Consortium (TAG). The GWAS meta-analysis files are publicly available on the Psychiatric Genomics Consortium (PGC) website: <https://www.med.unc.edu/pgc/results-and-downloads>. A total of 74,053 participants were included in the discovery phase of the analysis, and 143,023 were included in a follow up meta-analysis of the 15 most significant regions. Eight SNPs exceeded genome-wide significance. HapMap-2 CEU was used as the imputation panel resulting in a common set of ~2.5 million SNPs across studies. Individuals who were recorded as having ever been regular smokers were defined as those who reported having smoked at least 100 cigarettes during their lifetime, and never regular smokers were defined as those who reported having smoked between 0 and 99 cigarettes during their lifetime. Study-specific GWASs controlled for imputed allele dosage for a SNP plus whether a subject was classified as a case in the primary study. If the primary study was case-control in design and the phenotype being studied was known to be associated with smoking, the GWAS adjusted for case status to reduce potential confounding. Analyses were run and meta-analyzed separately for males and females. The European ancestry PGSs contain 779,538 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 777,095 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	13834.90	14071.30	236.40	13953.50	13953.61	0.32
African ancestry	2279	13883.50	14069.50	186.00	13964.00	13964.71	0.55



European ancestry: Dist. of PGS_EvrSmk_TAG10



African ancestry: Dist. of PGS_EvrSmk_TAG10

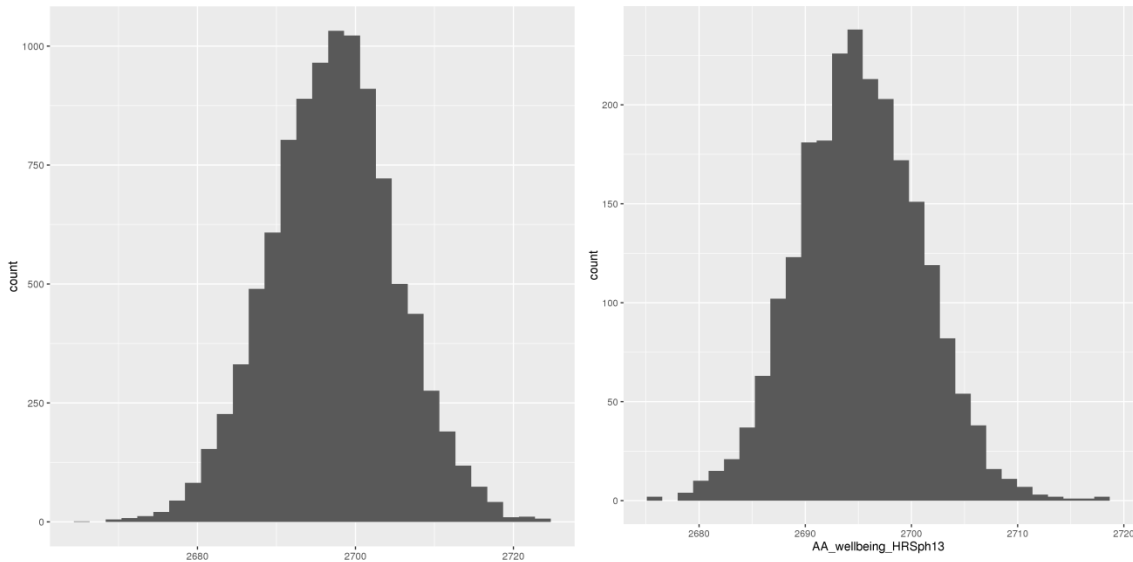
References

Tobacco and Genetics Consortium. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, 42(5), 441-447.

M. Subjective Wellbeing

The PGSs for subjective wellbeing were created using results from a 2016 GWAS conducted by the Social Science Genetic Association Consortium (SSGAC). The subjective wellbeing analyses included 298,420 European ancestry individuals in the discovery sample. Genome-wide significant SNPs were identified in 3 loci (**Table 1**). A quasi-replication analysis tested whether these three SNPs were associated with depressive symptoms and neuroticism. The phenotype measure was life satisfaction, positive affect, or in some cohorts a measure combining both. Approximately 9.3 million SNPs were included in the analyses, with cohorts utilizing SNPs imputed to the 1000 genomes reference panel (1000G) or the HapMap 2 Panel. Adjustments for age, age², sex, and population stratification (first four PCs from the genotypic data) were included in study-specific GWAS association analyses. Cohorts were also asked to include any study-specific covariates such as study site or batch effects. The original subjective wellbeing GWAS included the HRS. To compute PGSs for HRS respondents, the SSGAC provided SNP weights with the HRS and 23andMe results (due to data use agreements) removed (combined discovery + replication sample size without the HRS: N=288,478). The European ancestry PGSs contain 1,433,080 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 1,432,906 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	2666.21	2724.59	58.38	2697.28	2697.14	0.08
African ancestry	2279	2676.08	2718.23	42.15	2695.01	2695.15	0.12



European ancestry: Dist. of PGS_wellbeing_SSGAC16 African ancestry: Dist. of PGS_wellbeing_SSGAC16

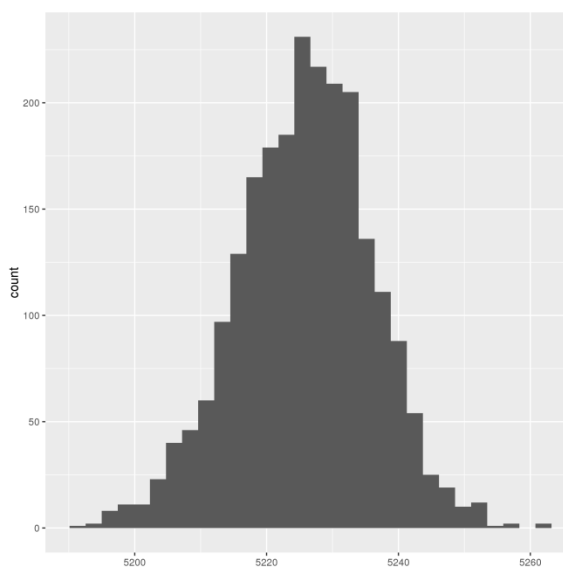
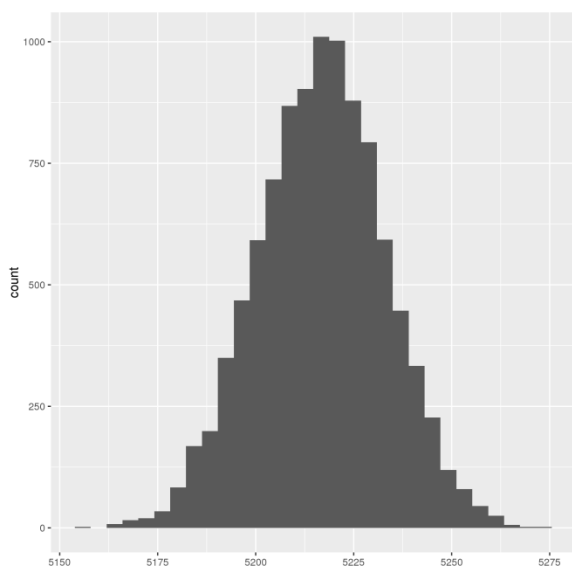
References

Okbay A, Baselmans BM, De Neve JE, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet.* 2016;48(6):624-633.

N. Neuroticism

The PGSs for neuroticism were created using results from a 2016 auxiliary GWAS conducted by the Social Science Genetic Association Consortium (SSGAC) as part of their subjective wellbeing GWAS (see above). The GWAS meta-analysis files are publicly available on the SSGAC website: <https://www.thessgac.org/data>. The entire meta-analysis included 170,911 individuals. Meta-analysis was performed on publicly available results from the Genetics of Personality Consortium (GPC) (N=63,661) with results from UK Biobank data (N=107,245). The meta-analysis yielded 11 lead SNPs, 2 of which tag inversion polymorphisms (**Table 1**). A quasi-replication analysis tested whether these SNPs were associated with subjective wellbeing. A replication analysis was also performed using data from 23andMe (N=368,890). In UKB, the phenotype measure was the respondent's score on a 12-item version of the Eysenck Personality Inventory Neuroticism scale. The GPC harmonized different neuroticism batteries. In the UKB, analyses controlled for the first 15 PCs, indicator variables for genotyping array, sex, indicator variables for age ranges, and sex-by-age interactions. Model adjustments for the 29 cohorts contributing to the GPC meta-analysis varied (see de Moor et al., p. 644, 2015). The European ancestry PGSs contain 1,152,920 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 1,148,174 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	5156.65	5274.29	117.64	5216.96	5216.71	0.16
African ancestry	2279	5191.19	5261.87	70.68	5226.18	5225.79	0.21



European ancestry: Dist. of PGS_neuroticism_SSGAC16

African ancestry: Dist. of PGS_neuroticism_SSGAC16

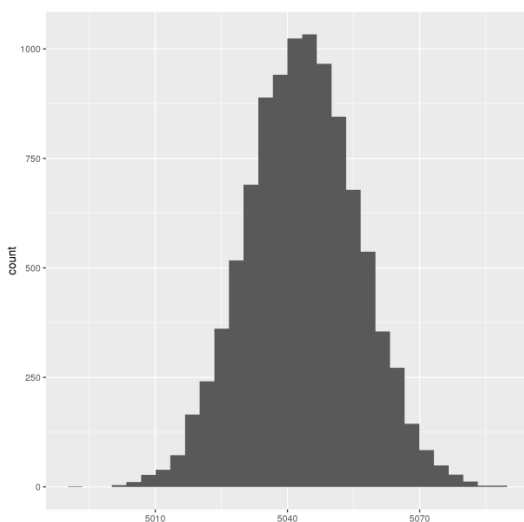
References

- Okbay, A., Baselmans, B. M., De Neve, J. E., Turley, P., Nivard, M. G., Fontana, M. A., ... & Gratten, J. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6), 624-633.
- De Moor, M. H., Van Den Berg, S. M., Verweij, K. J., Krueger, R. F., Luciano, M., Vasquez, A. A., ... & Gordon, S. D. (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry*, 72(7), 642-650.

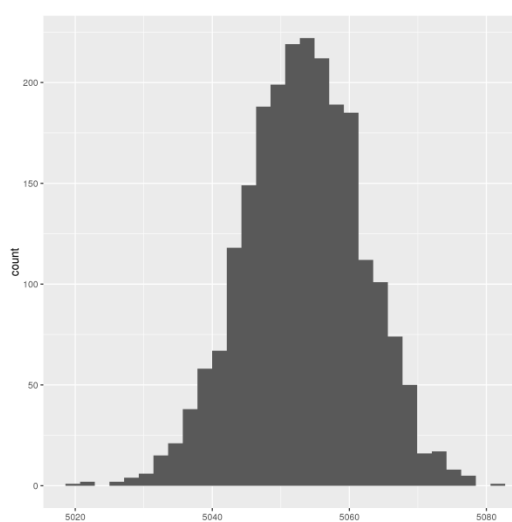
O. Depressive Symptoms

The PGSs for depressive symptoms were created using results from a 2016 auxiliary GWAS conducted by the Social Science Genetic Association Consortium (SSGAC) as part of their subjective wellbeing GWAS (see above). The GWAS meta-analysis files are publicly available on the SSGAC website: <https://www.thessgac.org/data>. The GWAS included 180,866 individuals and meta-analyzed publicly available results from a study performed by the Psychiatric Genomics Consortium (PGC) ($N_{\text{cases}} = 9,240$, $N_{\text{controls}} = 9,519$) with results from analyses of UK_Biobank (UKB) data ($N = 105,739$), and the Resource for Genetic Epidemiology Research on Aging (GERA) Cohort ($N_{\text{cases}} = 7,231$, $N_{\text{controls}} = 49,316$). The meta-analysis identified two genome-wide significant SNPs (**Table 1**). A quasi-replication analysis tested whether these SNPs were associated with subjective wellbeing. A replication analysis was also performed using data from 23andMe ($N=368,890$). In UKB, a continuous phenotype measure was used that combined responses to two questions, which ask about the frequency in the past two weeks with which the respondent experienced feelings of unenthusiasm/disinterest and depression/hopelessness. The PGC and GERA cohorts utilized case-control data on major depressive disorder. In the UKB, analyses controlled for the first 15 PCs, indicator variables for genotyping array, sex, indicator variables for age ranges, and sex-by-age interactions. In GERA, analyses controlled for the first four PCs of the genotypic data, sex, and 14 indicator variables for age ranges. The PGC included controls for five PCs, sex, age, and cohort fixed effects (for details see Ripke et al., 2013). The European ancestry PGSs contain 1,147,841 SNPs that overlapped between the HRS genetic database and the GWAS meta-analysis; African ancestry PGSs contain 1,144,655 SNPs. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation = 1). Raw PGS details are presented below:

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (Mean)
European ancestry	9991	4991.73	5088.11	96.38	5043.39	5043.42	0.13
African ancestry	2279	5019.32	5081.37	62.05	5053.23	5053.21	0.18



European ancestry: Dist. of PGS_depsymp_SSGAC16



African ancestry: Dist. of PGS_depsymp_SSGAC16

References

- Okbay, A., Baselmans, B. M., De Neve, J. E., Turley, P., Nivard, M. G., Fontana, M. A., ... & Gratten, J. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6), 624-633.
- Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., ... & Heath, A.

C. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, 18(4), 497.

III. Setup

By downloading this freely provided data set, you agree to use its contents only for research and statistical purposes, making no effort to identify the respondents. You also agree inform HRS of any papers, publications, or presentations based on this data set. Please send a copy of such publications in PDF format via e-mail to hquestions@umich.edu with "Attn: Papers and Publications" in the subject line. If you wish, you may include a bibliographical reference. As an alternative, you may transmit publications in paper format by postal mail:

Health and Retirement Study
 Attn: Papers and Publications
 The Institute for Social Research, Room 3410
 P.O. Box 1248
 Ann Arbor, Michigan 48106-1248

A. Distribution Set

The *HRS Polygenic Scores* data set is packaged for distribution in a ZIP file, **PGENSCORE.zip**. The .zip file contains two data files, PGENSCOREA_R (African Ancestry) and PGENSCOREE_R (European Ancestry). Both files are keyed on HRS Household Identifier (HHID) and Person Number (PN). To use these data, extract the data files, the program statement files that matches your analysis environment, the data description (this file), and the codebook files.

If you have problems when downloading this data product or in extracting its contents, please contact the [HRS Help Desk](#). See Table 1 (below) for a description of the .zip file contents as well as a suggested subdirectory structure.

Directory ³	File Type	Polygenic Scores	File Name
c:\polys\docs\	Codebook (ASCII text)		PGENSCORE.txt
	Data Description (this document)		PGENSCORESDD.pdf
c:\polys\data\	Data files (ASCII text)	AFRICAN ANCESTRY	PGENSCOREA_R.da (n=2279)
		EUROPEAN ANCESTRY	PGENSCOREE_R.da (n=9991)
c:\polys\sas\	SAS program statements	AFRICAN ANCESTRY	PGENSCOREA_R.sas
		EUROPEAN ANCESTRY	PGENSCOREE_R.sas
c:\polys\spss\	SPSS program statements	AFRICAN ANCESTRY	PGENSCOREA_R.sps
		EUROPEAN ANCESTRY	PGENSCOREE_R.sps
c:\polys\stata\	Stata dictionary/"do" files	AFRICAN ANCESTRY	PGENSCOREA_R.dct/.do
		EUROPEAN ANCESTRY	PGENSCOREE_R.dct/.do

³ When using HRS data products, you should feel free to create the directory structure that is most suitable for your needs. By using the suggested directory structure (or a Unix equivalent), you will not have to change the path name references in the data descriptor files. If you want to use a different structure, just change the directory references in the program files.

B. Program Statements

Each data file comes with associated SPSS, SAS, or Stata program statements to read the data. Files containing SPSS statements are named with an .sps extension, those with SAS statements with an .sas extension, and those with Stata statements with .do and .dct extensions.

1. Using the Files with SAS

To create a SAS system file for this data set, load the .sas program statement files into the SAS Program Editor and reference the appropriate .da data files. If the *.sas file is located in 'c:\polys\sas' and the data file is located in 'c:\polys\data', you can run the file as is. A SAS system file will be saved to directory 'c:\polys\sas'. If the files are not located in the specified directories, you will need to edit the *.sas file to reflect the proper path names prior to running the file.

2. Using the Files with SPSS

To create an SPSS system file for this data set, load the .sps program statement files into the SPSS syntax editor window, reference the appropriate .da data files, and select the *Run>All* option. If the *.sps file is located in 'c:\polys\spss' and the data file is located in 'c:\polys\data', you can run the file as is. An SPSS system file (*.sav) will be saved to directory 'c:\polys\spss'. If the files are not located in the specified directories, you will need to edit the *.sps file to reflect the proper path names prior to running the file.

3. Using the Files with Stata

To use Stata with this data set, three file types must be present for that data set: .dct, .do, and .da. Files with the suffix ".da" contain the raw data for Stata to read. Files with the suffix ".dct" are Stata dictionaries used by Stata to describe the data. Files with the suffix ".do" are short Stata programs ("do files") which you may use to read in the data. Load the .do file into Stata and then submit it. If the *.do and .dct files are located in 'c:\polys\stata' and the data file is located in 'c:\polys\data', you can run the .do file as is. If the files are not located in these directories, you must edit the *.do and *.dct files to reflect the proper path names before you run the files.

C. Non-Windows Environments

Non-Microsoft users should modify the default Windows file structure syntax to match that of their own operating system. The following examples should work for both Macintosh OS X and any Unix/Linux distribution. Open the SAS program file(s), SPSS syntax file(s) or the Stata do/dct files in an ASCII editor and make the changes indicated below.

1. SPSS in an OSX environment

In this example, we assume that the user has downloaded the region dataset and placed the files in a **Desktop** folder called **Polys** with the ASCII data file stored in subfolder **data** and the syntax file in subfolder **spss**. Then the commands in the syntax file that reads PGENSCOREA_R.da would be modified to look like this:

```
FILE HANDLE polys /name='Desktop/Polys/data/PGENSCOREA_R.da' LRECL=431.
DATA LIST FILE= polys/
HHID 1-6(A)
[rest of syntax file goes here]
.
execute.
SAVE /outfile 'Desktop/Polys/spss/PGENSCOREA_R.sav'. Execute.
```

2. STATA in an OS X Environment

In the following example we assume that:

- The username is “user1”
- The zip file containing tracker information has been downloaded to the user’s desktop from the HRS file download site
- The user has decompressed the zip file (use Stuffit for OS X) into a desktop folder named **Polys**
- The statistical package is stata

File `PGENSCOREA_R.do` should be modified as follows:

Change...

```
infile using c:\polys\stata\PGENSCOREA_R.dct
```

To...

```
infile using /Users/user1/Desktop/Polys/stata/PGENSCOREA_R.dct
```

Change...

```
save c:\polys\stata\PGENSCOREA_R.dta
```

To...

```
save /Users/user1/Desktop/Polys/stata/PGENSCOREA_R.dta
```

File `PGENSCOREA_R.dct` should be modified as follows

Change...

```
dictionary using c:\polys\data\PGENSCOREA_R.da {
```

To...

```
dictionary using /Users/user1/Desktop/Polys/data/PGENSCOREA_R.da {
```

IV. If You Need to Know More

This document is intended to serve as a brief overview to provide guidelines for using the *HRS Polygenic Scores* data product. If you have questions or concerns that are not adequately covered here or on our Web site, or if you have any comments, please contact us. We will do our best to provide answers.

A. HRS Internet Site

Health and Retirement Study public release data and additional information about the study are available on the Internet. To access public data or to find out more about restricted data products and procedures, visit the [HRS Web site](#).

B. Contact Information

If you need to contact us, you may do so by one of the methods listed below.

Internet: Help Desk at the HRS Web site (<http://hrsonline.isr.umich.edu>)

E-mail: hrsquestions@umich.edu

Postal Service:

Health and Retirement Study
The Institute for Social Research
426 Thompson Street, 3410 ISR
Ann Arbor, Michigan 48104