Documentation Report

FFCWS Polygenic Scores – Release 1
Genetic Data Freezes 1-3

Please cite in references:

"Ware EB, Fisher J, Schneper L, Notterman D, Mitchell C. FFCWS Polygenic Scores – Release 1. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan; and Princeton, NJ: Department of Molecular Biology, Princeton University; 2021."

Please cite grants **R01 HD36916, R01 HD073352, R01 HD076592**, which provide support for the collection, assay, and analysis of the genetic to create the polygenic scores.

Documentation report and polygenic scores prepared by

Erin B. Ware, University of Michigan
Jonah Fisher, University of Michigan
Colter Mitchell, University of Michigan

Genetic data prepared by

Lisa Schneper, Princeton University
Iulia Kotenko, Princeton University
Dan Notterman, Princeton University

Survey Research Center
University of Michigan
Ann Arbor, Michigan
August 2021

Contents

I. Introduction

This guide describes the construction of polygenic scores (PGSs) for a variety of phenotypes for FFCWS respondents who provided salivary DNA at age 9 (collected and genotyped 2007-2010). The mothers of these children are also being genotyped, but the current data only contain the children. These scores serve as an attempt to harmonize research across studies. 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 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. Data collection

For the Ages 9 and 15 follow-ups, FFCWS obtained genetic samples from the participants. In addition to the core mother and father surveys, the Age 9 follow-up included child surveys, teacher surveys, home assessments, interviewer observations, primary caregiver surveys, as well as DNA samples from both the children and their mothers. For more information on Core data collection, see the Year 9 User Guide. The Age 15 follow-up included primary caregiver surveys, in home assessments (for a subset of families), child surveys, and DNA samples. For more information, see the Year 15 User Guide.

C. Saliva Collection Procedures

Respondents were instructed to rinse out their mouth 30 minutes prior to providing saliva. Respondents were then instructed to spit into the saliva container until they filled it up to a line indicating 2 ml volume. Saliva was collected using the Oragene DNA Self-Collection Kit. DNA was extracted using the Oragene® Laboratory Protocol Manual Purification of DNA.

II. File Layout

The file contains observations for all 4,898 births in the FFCWS study. PGS data are available for 3,074 focal children. The remaining cases lack PGS data for one of two reasons: (1) the participant did not provide a saliva sample (coded as -9 No DNA sample/Not in wave), or (2) although a saliva sample was provided a PGS was not constructed because the ancestry group was too small (see section III.D) or the sample failed quality control (coded as -3 missing due to technical issues).

A. Variable Naming Convention

The PGS data file follows the standard FFCWS variable naming convention. All variables other than *idnum* begin with “g” because they are from our genetic and epigenetic data. Because most of the DNA comes from the age 9 (wave 5) child data collection, the variable is labeled gk5. The next letters indicate if the variable is a polygenic score (pg; see section IIA), a principal component (pc; see section III.D), or an indicator of it is part of the more or less admixed sample (admix; see section III.D). For polygenic scores (i.e. gk5pg___) the next three letters will indicate which polygenic score. For PC measures (i.e. gk5pc_) the number indicates what principal component it is (see section III.D). The last letter for all variables (a, e, or h) indicates which ancestry the group is included in

that variable (see section III C and D): a- predominantly African analytic group, e- predominantly European analytic group, h- predominantly Hispanic analytic group.

B. Missing Values

Three missing data codes are used in these data and follow general FFCWS codes:

-9 Not in Wave (i.e. did not provide saliva at Age 9)

-7 Not Applicable—because PGS and PC data are generated separately by analytic ancestry groups, participants with PGS data who are not in that analytic group are coded as "Not applicable" for that specific measure. As described in more detail below, different variables were created for each analytic group because comparing mean differences in the PGS by analytic ancestry group is scientifically invalid.

-3 Missing due to technical issues (i.e. PGS data were not generated despite have a DNA sample either due to having too small an ancestry group or sample failed genomic data quality control).

III. FFCWS Genomic Data

Specimen processing was conducted at the Notterman laboratory at Princeton University from 2015-2019 (R01 HD36916, R01 HD 073352, R01HD 076592). Genotype data on FFCWS participants was obtained using the Illumina PsychChip_v1-1. Individuals with missing call rates >2%, SNPs with missing call rates >2%, and chromosomal anomalies were removed. 3,074 individuals and 273,800 SNPs passed filters and QC.

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 among SNPs (i.e. clumping vs. pruning). Using a population-representative study [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.[1] The method we choose for PGS construction is referred to as "P+T," or pruning and thresholding.

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, we have chosen to provide scores that include all available SNPs in the PGS that overlap between the GWAS meta-analysis and the FFCWS 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. Polygenic scores are additive in nature. PGSs are calculated using the following formula:

---

[1] For additional information on this analysis, see:
Ware EB, Schmitz LL, Faul JD, Gard AM, Smith JA, Mitchell CM, Weir DR, Kardia SLR. (2017) *Method of Construction Affects Polygenic Score Prediction of Common Human Traits*. BiorXiv. 2017 doi: https://doi.org/10.1101/106062

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

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 genotype given the allele frequency. Scores were similar when not employing the missing data imputation default. PGSs were calculated using PRSIce-2 polygenic score calculation program.[2]

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. SNP weights were downloaded from consortium webpages, requested from consortium authors, obtained from dbGap, or taken from published supplemental material. All base SNP files from GWAS meta-analyses were converted to NCBI build 37 annotation for compatibility with FFCWS SNP data.

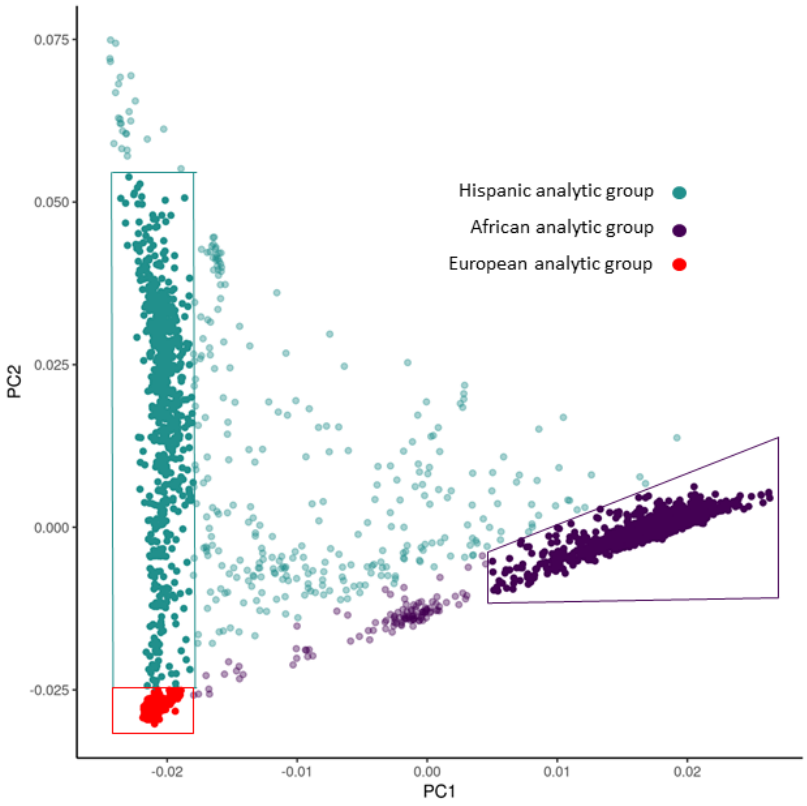
C. Notes about the use of PGSs

We release polygenic scores for a European analytic group, a more admixed African analytic group and a less admixed African analytic group, and a more admixed Hispanic analytic group and a less admixed Hispanic analytic group, separately. **However, it should be noted that the majority of 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).** We are currently evaluating other methods of constructing trans-ethnic polygenic scores that better account for population stratification and admixture.

D. Analytic groups and genetic principal components

Global genetic 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 < 2% 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, we excluded the HLA, 8p23, and 17q21.31 regions from the initial pool. We identified analytic genetic groups in FFCWS through PC analysis on genome-wide SNPs calculated across all participants using the aforementioned filtering criteria. We are currently evaluating methods to create analytic groups for admixed populations. It is possible that future releases of polygenic scores for the FFCWS will include different analytic groups.

[2] Choi, S.W., Mak, T.S. & O'Reilly, P.F. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc (2020). https://doi.org/10.1038/s41596-020-0353-1



Principal component plot of global principal component 1 (PC1) and principal component 2 (PC2). Individuals are colored by European analytic group (boxed and red), African analytic group (boxed less admixed dark purple; more admixed dark purple plus light purple), and Hispanic analytic group (boxed less admixed dark green; more admixed dark green plus light green). This plot has been rotated to show the European ancestry in the bottom left corner of the plot.

European analytic sample (n=475): The European analytic sample includes all participants who had PC loadings of PC1 > 0.018 and PC2 > -0.0075 from the global PC analysis of *all* unrelated study subjects. This area is consistent with 1000G EUR super population clusters. It is important to note that this group may contain individuals who self-identify as a race/ethnicity other than "White". Also, although there is an admixed indicator variable **gk5admixe**, there are no participants within the European analytic group classified as coming from a more admixed population.

African analytic sample – less admixed (n=1528; gk5admixa=0): The less admixed African analytic sample includes all participants who had PC loadings of P1 < -0.005 and PC2 > 0.007+0.75(PC1) from the global PC analysis of *all* unrelated study subjects. We have flagged this group in the data for those wishing to do analyses or sensitivity analyses on a more genetically clustered group. This area is consistent with 1000G AFR super population clusters. It is important to note that this group may contain individuals who self-identify as a race/ethnicity other than "Non-Hispanic Black". A more **Admixed African analytic sample (n=1640; gk5admixa=0 or 1)** has been defined in the data for those wishing to do analyses with a more admixed sample of African ancestry.

Hispanic analytic sample – less admixed (n=640; gk5admixh=0): The less admixed Hispanic analytic sample includes all participants who had PC loadings of PC1 > 0.018 and -0.055 < PC2 < 0.025 from the global PC analysis of *all* unrelated study subjects. We have flagged this group in the data for those wishing to do analyses or sensitivity analyses on a more genetically clustered group. Because Hispanic genetic ancestry is incredibly diverse, this area is consistent with only some of the 1000G AMR super population clusters (specifically CLM-Medellin, Colombia, PEL-Lima, Peru). It is important to note that this group may contain individuals who self-identify as a race/ethnicity other than "Hispanic". A more

Admixed Hispanic analytic sample (n=959; gk5admixh=0 or 1) has been defined in the data for those wishing to do analyses or sensitivity analyses on a more admixed sample of Hispanic ancestry.

Local "within-analytic-group" genetic principal components: Once analytic samples were identified (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959), 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, within-analytic group. We refer to these as "within-analytic-group PCs".

Within-analytic-group PCs 1-20 (**gk5pc1e-gk5pc20e, gk5pc1a-gk5pc20a, gk5pc1h-gk5pc20h**) are included for each analytic group. The PCs control for any genetic aspects of common ancestry that may spuriously correlate with the polygenic score and the outcome of interest (Price et al., 2006). **We highly recommend that users perform analyses separately by ancestral group. We recommend the following number of PCs for each analytic group, but at the very least recommend adjusting for PCs 1-5**:

Analytic group	Sample size	Number of PCs recommended in analyses
European	475	10
More admixed African*	1640	5
Less admixed African	1528	5
More admixed Hispanic**	959	10
Less admixed Hispanic	640	5

*The less admixed African analytic group is included in the more admixed African analytic group;
**The less admixed Hispanic analytic group is included in the more admixed Hispanic analytic group

References

Choi, S.W., Mak, T.S. & O'Reilly, P.F. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc (2020). https://doi.org/10.1038/s41596-020-0353-1Euesden 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, Kardia 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, Kardia SLR. (2017). Method of Construction Affects Polygenic Score Prediction of Common Human Traits. BiorXiv. 2017 doi: https://doi.org/10.1101/106062

IV. Polygenic score GWAS descriptions and distributions

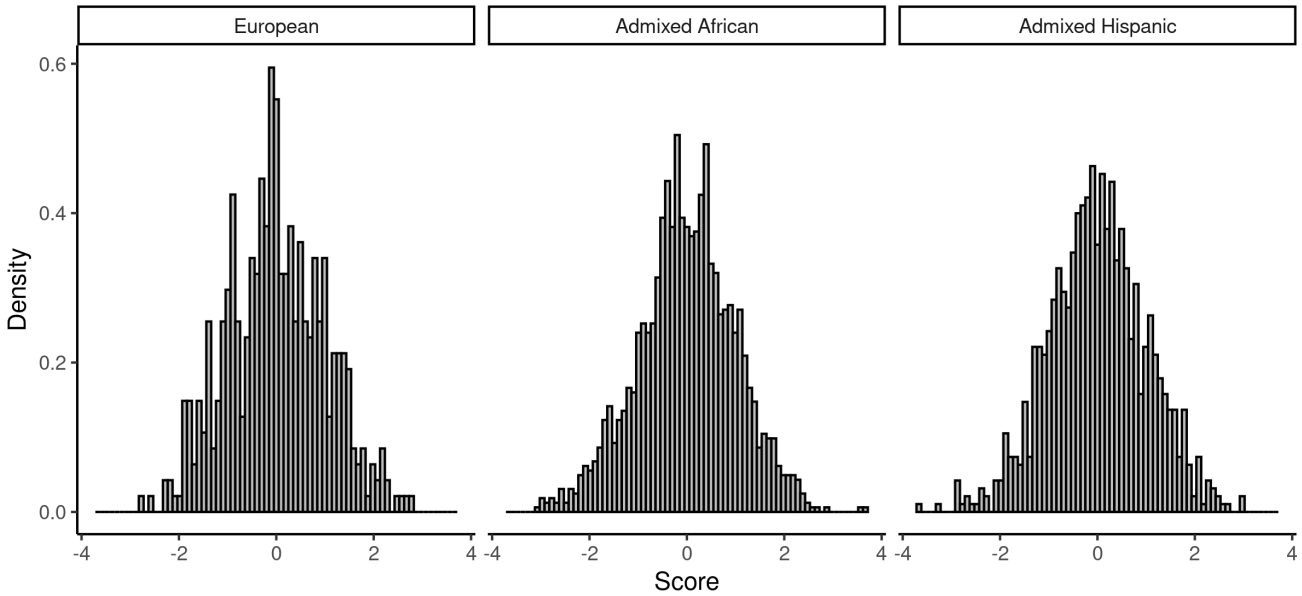
A. Body Mass Index (BMI)

BMI: gk5pgbmie, gk5pgbmia, gk5pgbmih

PGSs for body mass index (BMI) were created using results from a 2018 study conducted by the Genetic Investigation of Anthropometric Traits (GIANT) consortium. The GWAS meta-analysis files are publicly available on the Broad Institute data download page: (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files#2018_GIANT_and_UK_BioBank_Meta-analysis). The meta-analysis included 681,275 participants from a total of 15 cohorts of European ancestry. The 15 cohorts include the UK Biobank (UKB) and 14 cohorts from the previous GIANT GWAS of BMI (Locke et al., 2015; *Nature*). Authors performed a fixed effect inverse-variance weighted meta-analysis of the UKB results with GWAS summary statistics from Locke et al. (2015). 2,334,002 SNPs imputed from NCBI Build 37 HapMap phase 2 data were included in the meta-analysis. The GWAS of BMI in UKB was conducted in 456,426 participants of European Ancestry, using 16,653,239 SNPs imputed to the Haplotype Reference Consortium imputation reference panel. Associations were adjusted for 10 principal components to reduce confounding by population stratification, as well as for age, sex, recruitment center, and genotyping batch. The study identified 941 genome-wide significant SNPs ($P < 10^{-8}$) (**Figure 1** and **Table 1**).

The GIANT BMI2 PGSs contains 202,501 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ancestry to a standard normal curve (mean=0, standard deviation = 1).

Please note that the GIANT-BMI2 summary statistics are from a GWAS on individuals of European ancestry (see Section C. "Notes about the use of PGSs" for more information on the use of PGSs in other ancestry groups).



Distribution of BMI PGS, by analytic group standardized within analytic group
(European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)

References

Yengo L, Sidorenko J, Kemper KE, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641–3649. doi:10.1093/hmg/ddy271

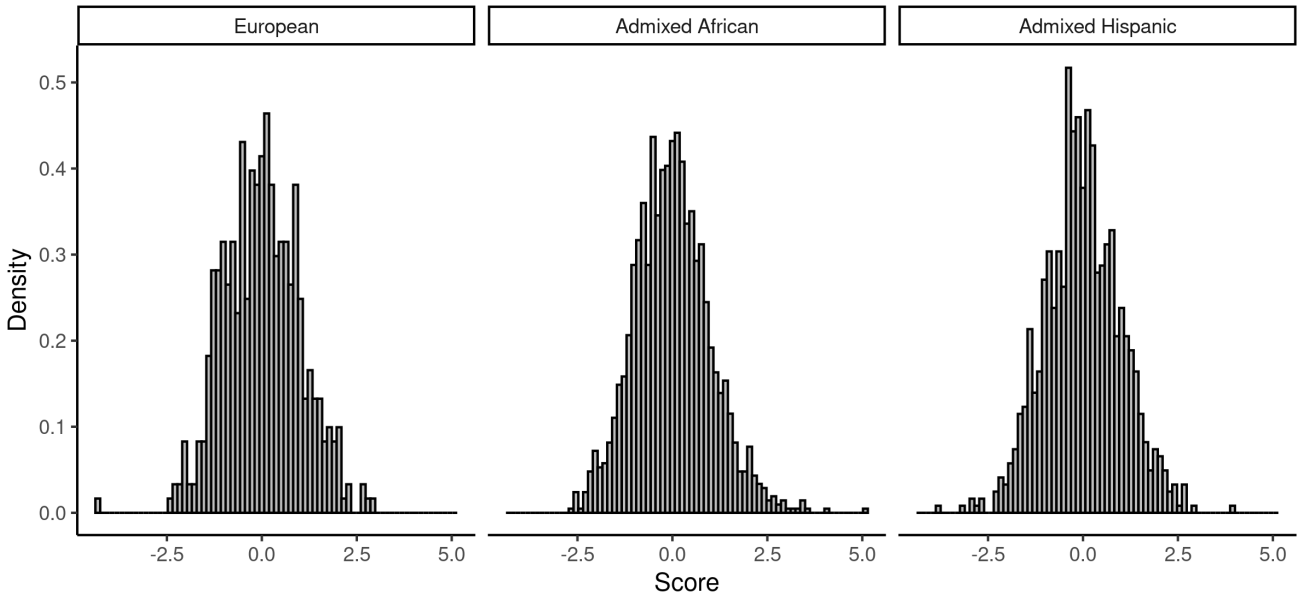
B. Height

Height: gk5pghgte, gk5pghgta, gk5pghgth

PGSs for Height were created using results from a 2018 study conducted by the Genetic Investigation of Anthropometric Traits (GIANT) consortium. The GWAS meta-analysis files are publicly available on the Broad Institute data download page: (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files#2018_GIANT_and_UK_BioBank_Meta-analysis). The GIANT meta-analysis included 693,529 participants from a total of 6 cohorts of European ancestry. The 6 cohorts include the UK Biobank (UKB) and 5 cohorts from the previous GIANT GWAS of Height (Wood et al., 2014; *Nature Genetics*). Authors performed a fixed effect inverse-variance weighted meta-analysis of the UKB results with GWAS summary statistics from Wood et al. (2014). SNPs were imputed from NCBI Build 37 HapMap phase 2 data were included in the meta-analysis. The GWAS of Height in UKB was conducted in 456,426 participants of European Ancestry, using 16,653,239 SNPs imputed to the Haplotype Reference Consortium imputation reference panel. Associations were adjusted for 10 principal components to reduce confounding by population stratification, as well as for age, sex, recruitment center, and genotyping batch. The study identified 3,290 genome-wide significant SNPs ($P < 10^{-8}$) (**Figure 1** and **Table 1**).

The GIANT Height PGS contains 202,619 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ancestry to a standard normal curve (mean=0, standard deviation = 1).

Please note that the GIANT Height2 summary statistics are from a GWAS on individuals of European ancestry (see Section C. "Notes about the use of PGSs" for more information on the use of PGSs in other ancestry groups).



Distribution of Height PGS, by analytic group standardized within analytic group
(European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)

References

Yengo L, Sidorenko J, Kemper KE, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641–3649. doi:10.1093/hmg/ddy271

C. Waist Circumference and Waist-to-Hip Ratio

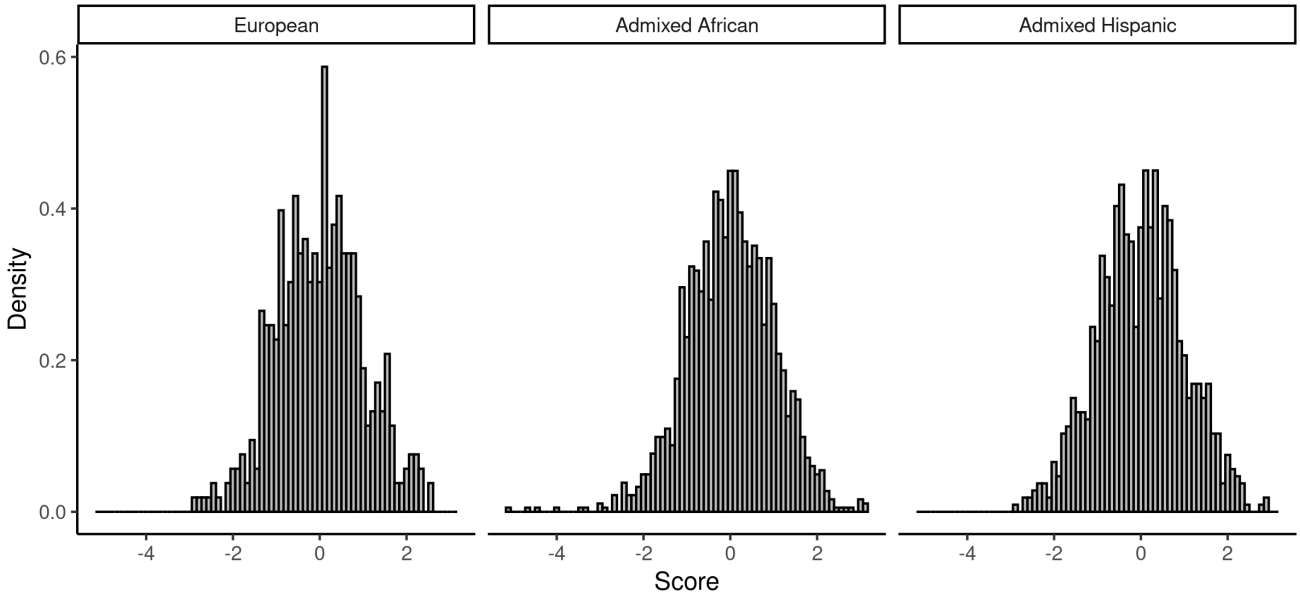
Waist Circumference: gk5pgwcre, gk5pgwcra, gk5pgwcrh
Waist-to-Hips Ratio: gk5pgwhre, gk5pgwhra, gk5pgwhrh

PGSs for waist circumference (WC) and waist-to-hip ratio (WHR) were created using results from a 2015 study conducted by the **G**enetic **I**nvestigation of **ANT**hropometric 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 (WC: GIANT 2015 WC COMBINED EUR.txt.gz, WHR: GIANT 2015 WHR COMBINED EUR.txt.gz).

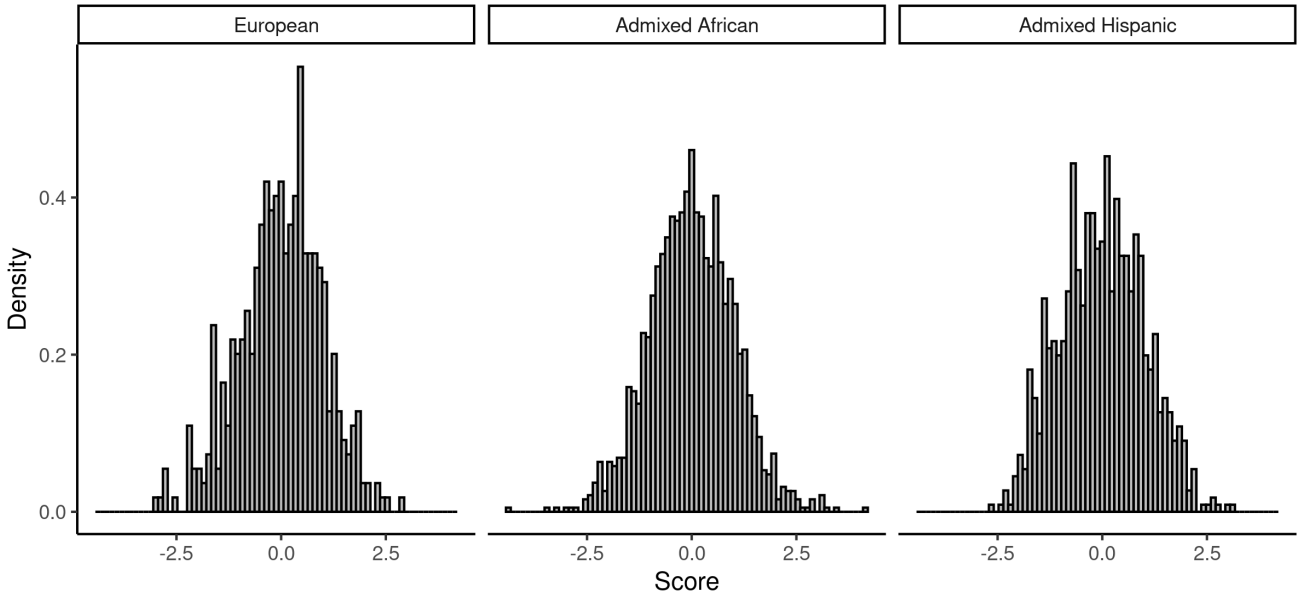
GWAS meta-analysis was performed on a sample of 142,762 individuals from 57 studies across 2,507,022 SNPs, and separately in a Metabochip (MC) meta-analysis on a sample of 67,326 individuals from 44 studies across 124,196 SNPs. A joint GWAS and MC meta-analysis was then conducted on 210,088 individuals 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, age^2, study-specific covariates if necessary, and BMI.

The GIANT WC PGSs contains 214,391 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The GIANT WHR PGSs contains 314,206 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ancestry to a standard normal curve (mean=0, standard deviation = 1).

These weights are from the joint analysis of GWAS and MC meta-analysis conducted on 210,088 individuals. **Please note that the GIANT results are from a GWAS on individuals of European ancestry (see Section C. "Notes about the use of PGSs" for more information on the use of PGSs in other ancestry groups).**



Distribution of waist circumference PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)



Distribution of waist to hip ratio PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)

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.

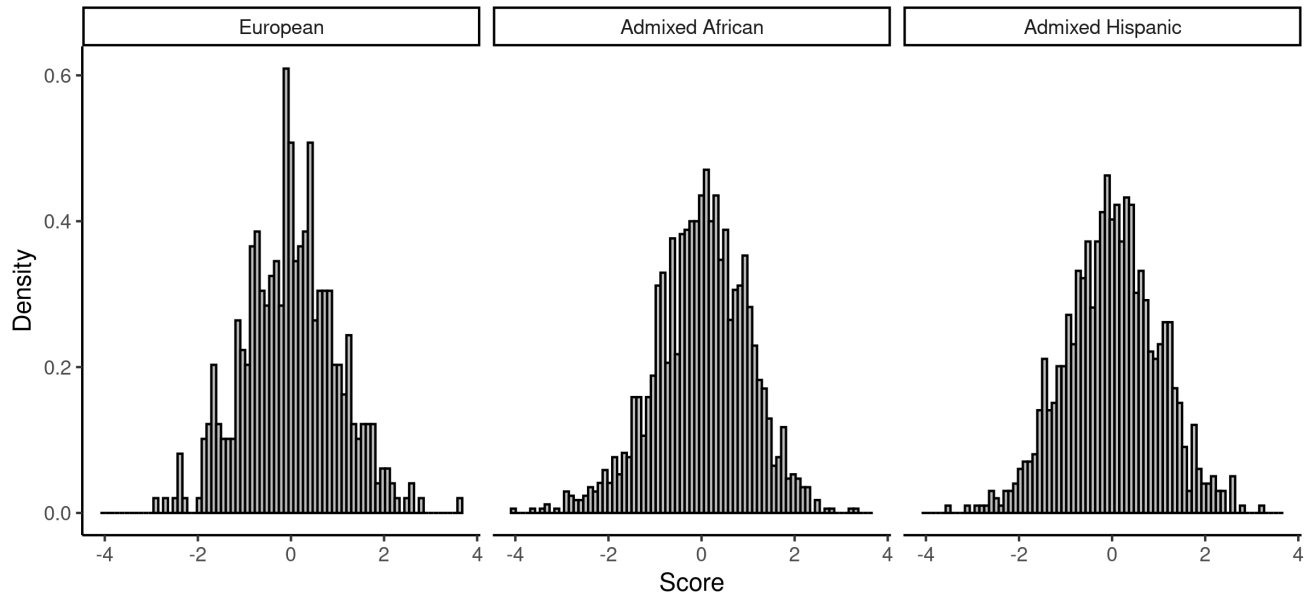
D. Myocardial Infarction

Myocardial Infarction: gk5pgmyie, gk5pgmyia, gk5pgmyih

The PGSs for myocardial infarction (MI) were created using 2015 results from a subgroup analysis of coronary artery disease (CAD) conducted by the Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) consortium. The GWAS meta-analysis files are publicly available and can be downloaded from www.cardiogramplusc4d.org (mi.add.030315.website.txt). The GWAS is a meta-analysis of 48 studies of mainly European, South Asian, and East Asian, descent imputed using the 1000 Genomes phase 1 v3 training set with 38 million variants. The study interrogated 9.4 million variants and involved 60,801 CAD cases and 123,504 controls. Case status was defined by an inclusive CAD diagnosis (for example, myocardial infarction, acute coronary syndrome, chronic stable angina or coronary stenosis of >50%). Thirty-seven previous loci and ten new loci achieved genome-wide significance (**Supplementary Table 2**). MI subgroup analysis was performed in cases with a reported history of MI (~70% of the total number of cases). No additional loci reached genome-wide significance in the MI analysis.

The CARDIoGRAM MI PGS contains 224,303 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ancestry to a standard normal curve (mean=0, standard deviation = 1). Weights are represented as log(OR).

Please note that the CARDIoGRAM results are from a GWAS on individuals of mostly European ancestry (see Section C. "Notes about the use of PGSs" for more information on the use of PGSs in other ancestry groups).



Distribution of myocardial infarction PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)

References

CARDIoGRAMplusC4D Consortium. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, *47*(10), 1121-1130.

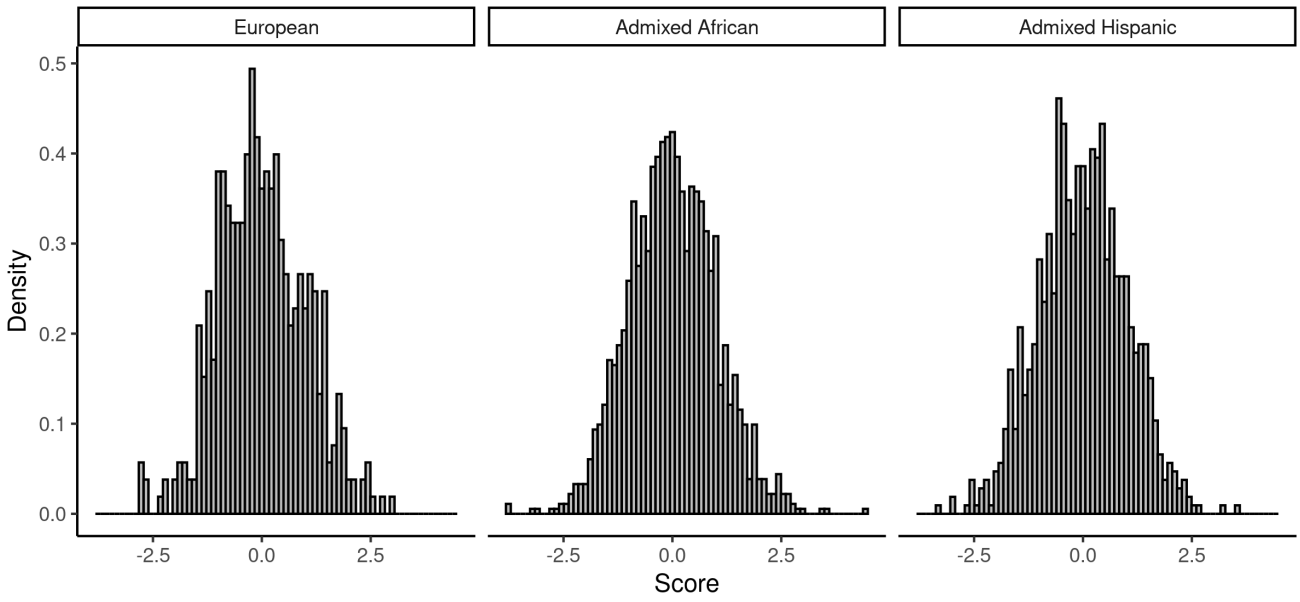
E. Age at Menarche

Menarche: gk5pgmnae, gk5pgmnaa, gk5pgmnah

PGSs for age at menarche were created using results from a 2014 study conducted by the Reproductive Genetics (ReproGen) consortium. The GWAS meta-analysis files are publicly available on the ReproGen data download page: http://www.reprogen.org/data_download.html (Menarche_Nature2014_GWASMetaResults_17122014.txt). The ReproGen meta-analysis included 182,416 women of European descent from 57 studies imputed to HapMap Phase 2 CEU build 35 or 36 with a total of 2,441,815 autosomal SNPs. Birth year was the only covariate included to allow for the secular trends in menarche timing. The study reported 3,915 genome-wide significant SNPs (**Figure 1**). Of these, the authors identified 123 independent signals for age at menarche, which they assessed further in an independent sample of 8,689 women from the EPIC-InterAct study.

The ReproGen age at menarche PGS contains 210,826 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ancestry to a standard normal curve (mean=0, standard deviation = 1).

Please note that the ReproGen results are from a GWAS on individuals of European ancestry (see Section C. "Notes about the use of PGSs" for more information on the use of PGSs in other ancestry groups).



Distribution of age at menarche PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)

References:

Perry, J. R., Day, F., Elks, C. E., Sulem, P., Thompson, D. J., Ferreira, T., ... & Albrecht, E. (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, *514*(7520), 92-97.

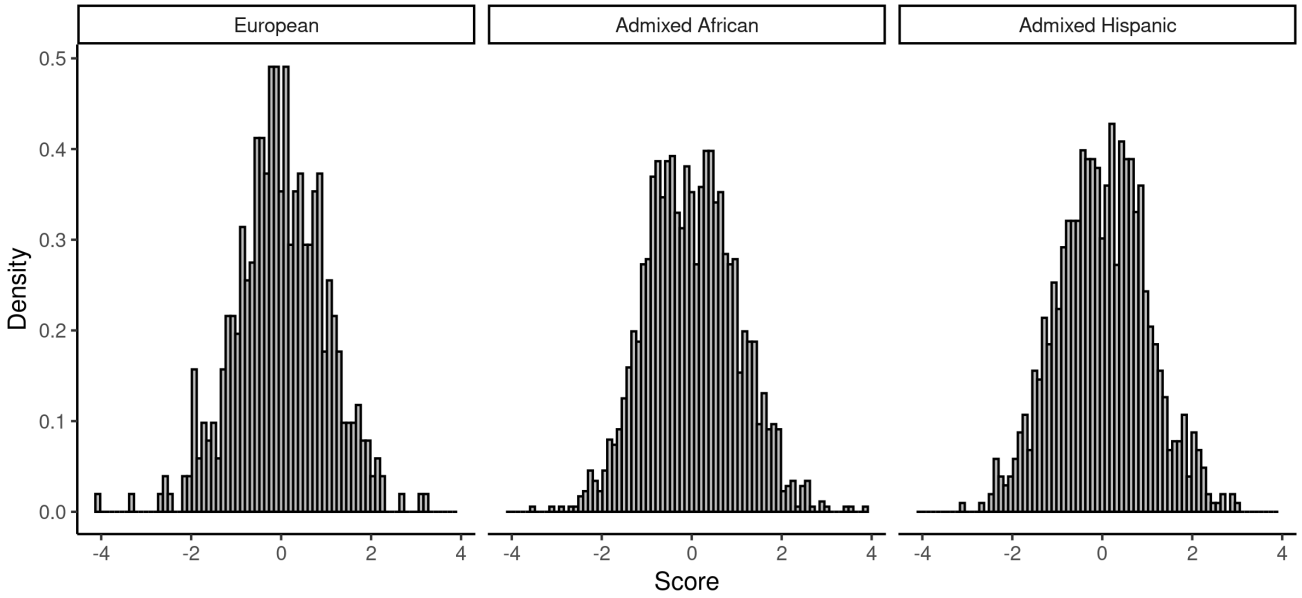
F. Lipid traits (High-density Lipoprotein (HDL), Low-density Lipoprotein (LDL), Total cholesterol (TC), Triglycerides)

HDL: gk5pghdle, gk5pghdla, gk5pghdlh
LDL: gk5pgldle, gk5pgldla, gk5pgldlh
Total Cholesterol: gk5pgtche, gk5pgtcha, gk5pgtchh
Triglycerides: gk5pgtrge, gk5pgtrga, gk5pgtrgh

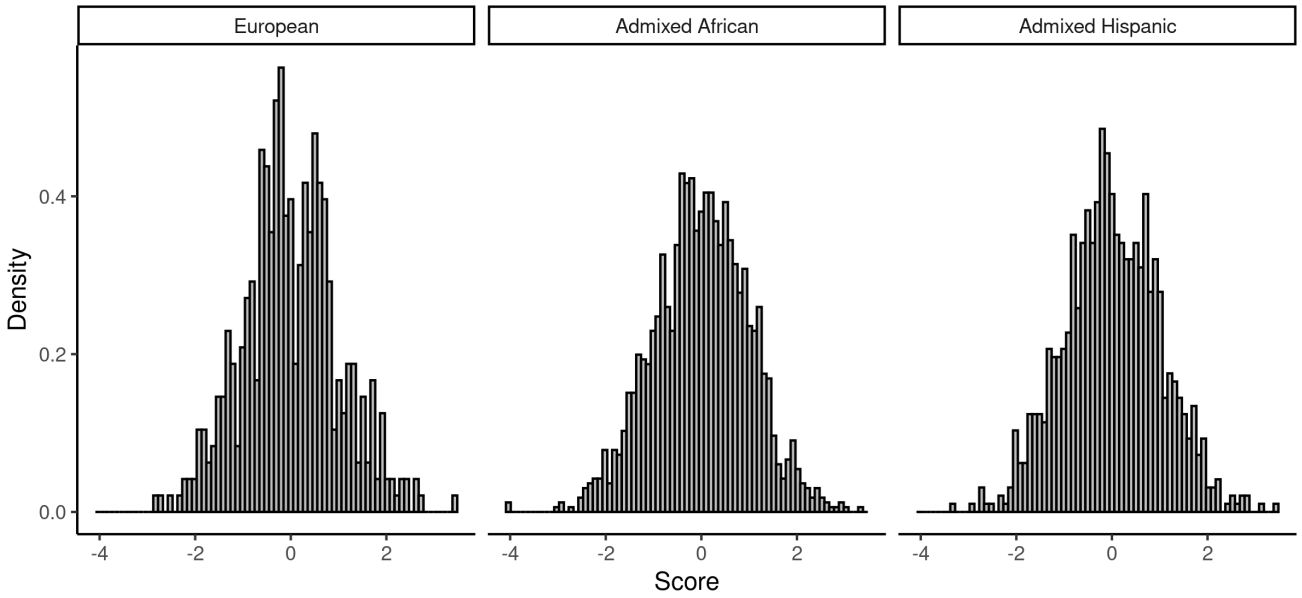
The HDL, LDL, and TC PGS were created using results from a 2013 study by the Global Lipid Genetics Consortium (Willer et al. 2013). Authors conducted separate GWAS for European (n=188,578) and non-European (n=7,898) ancestries followed by a meta-analysis of 7,168 individuals in a single ancestry group. Only European samples were used for discovery of novel genome-wide significant loci; non-European samples were meta-analyzed and examined only for fine-mapping analyses. Results are available for download directly from the Center for Statistical Genetics website (http://csg.sph.umich.edu/willer/public/lipids2013/) and results from the joint analysis of metabochip and GWAS data were used to create the PGSs. Results files were slightly modified on 11/26/2013. Sites with N<50,000 were removed from the joint meta-analysis results, sites with N<20,000 were removed from the Metabochip-only results and an rsid column was added to each dataset. Data was sourced by collecting summary statistics from 23 studies of European ancestry genotyped with GWAS arrays and 46 studies genotyped with Metabochip arrays, of which 37 studies consisted primarily of individuals of European ancestry. Nine studies using Metabochip arrays were of non-European ancestry: two studies were South Asian, two studies were East Asian, and five studies were African. Blood lipid levels were typically measured after > 8 hours of fasting and individuals known to be on lipid-lowering medication were excluded when possible. Hapmap release 22 CEU reference was used. In cases where Metabochip and GWAS array data were available for the same individuals, Metabochip data was used to ensure key variants were directly genotyped, rather than imputed. None of the GWAS meta-analyses included HRS. The study identified 157 loci associated with lipid levels at P <5×01-8, including 62 loci not previously associated with lipid levels in humans. Adjustments for population structure using principal component analysis or mixed model approaches were carried out in 24 studies (35% of individuals).

The GLGC HDL PGS contains 206,461 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The GLGC LDL PGS contains 206,253 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The GLGC TC PGS contains 206,449 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The GLGC TG PGS contains 206,267 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ancestry to a standard normal curve (mean=0, standard deviation = 1).

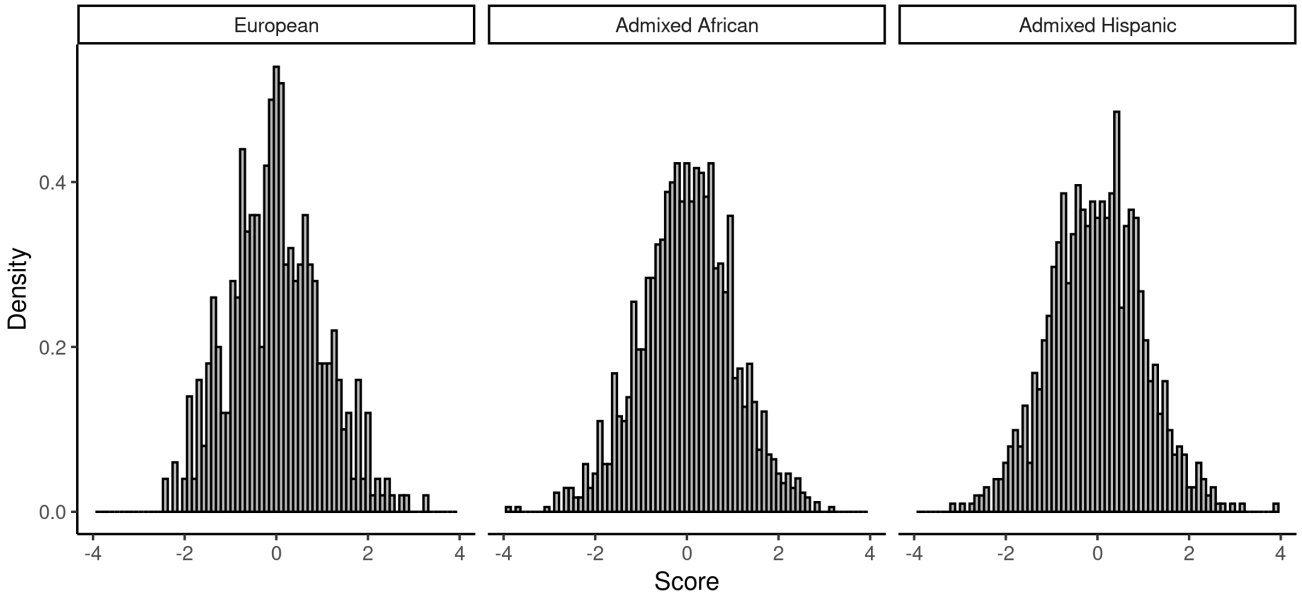
Please note that the GLGC-lipid results contain PGSs from European ancestry backgrounds (see Section C. "Notes about the use of PGSs" for more information on the use of PGSs in other ancestry groups).



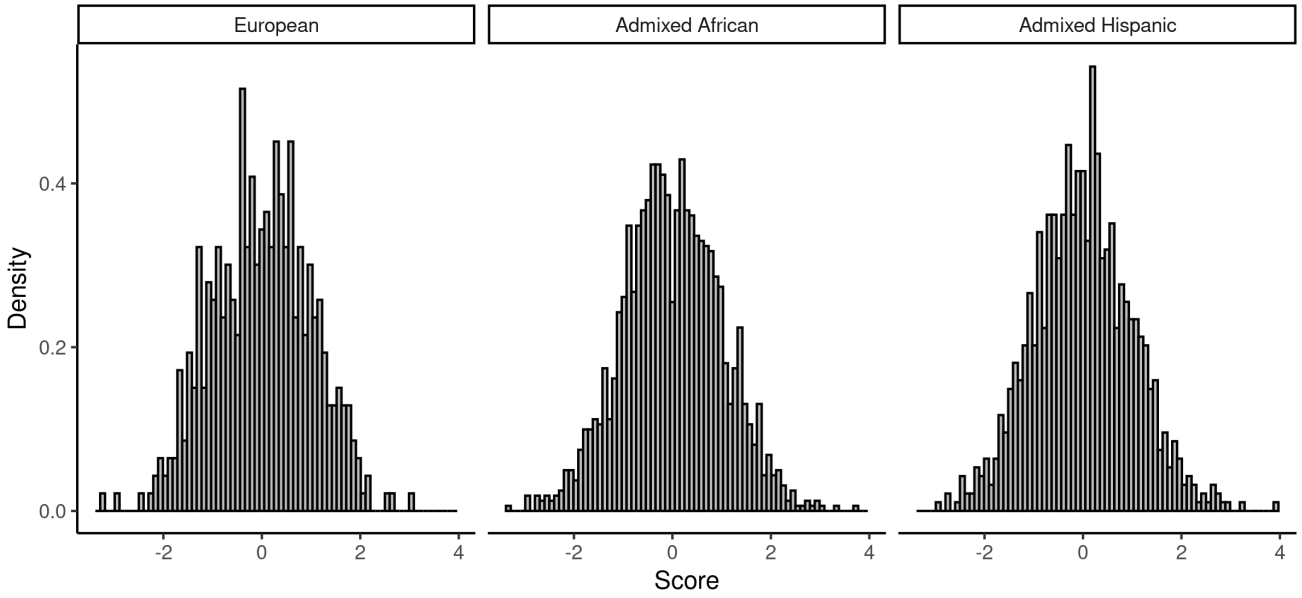
Distribution of HDL PGS, by analytic group standardized within analytic group
(European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)



Distribution of LDL PGS, by analytic group standardized within analytic group
(European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)



Distribution of total cholesterol PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)



Distribution of triglycerides PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)

References

Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., … & Global Lipids Genetic Consortium. (2013) Discovery and Refinement of Loci Associated with Lipid Levels. Nat Genet. 45(11), 1274-1283. doi:10.1038/ng.2797.

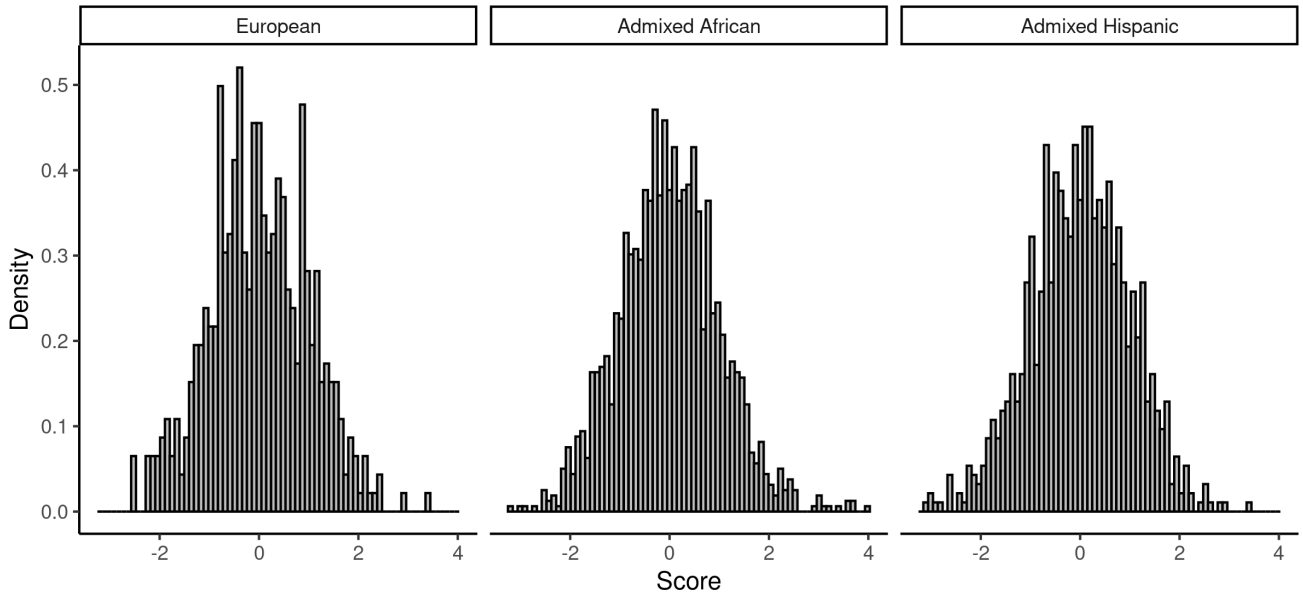
G. Blood pressure (Diastolic blood pressure (DBP) and Systolic Blood Pressure (SBP))

DBP: gk5pgdbpe, gk5pgdbpa, gk5pgdbph
SBP: gk5pgsbpe, gk5pgsbpa, gk5pgsbph

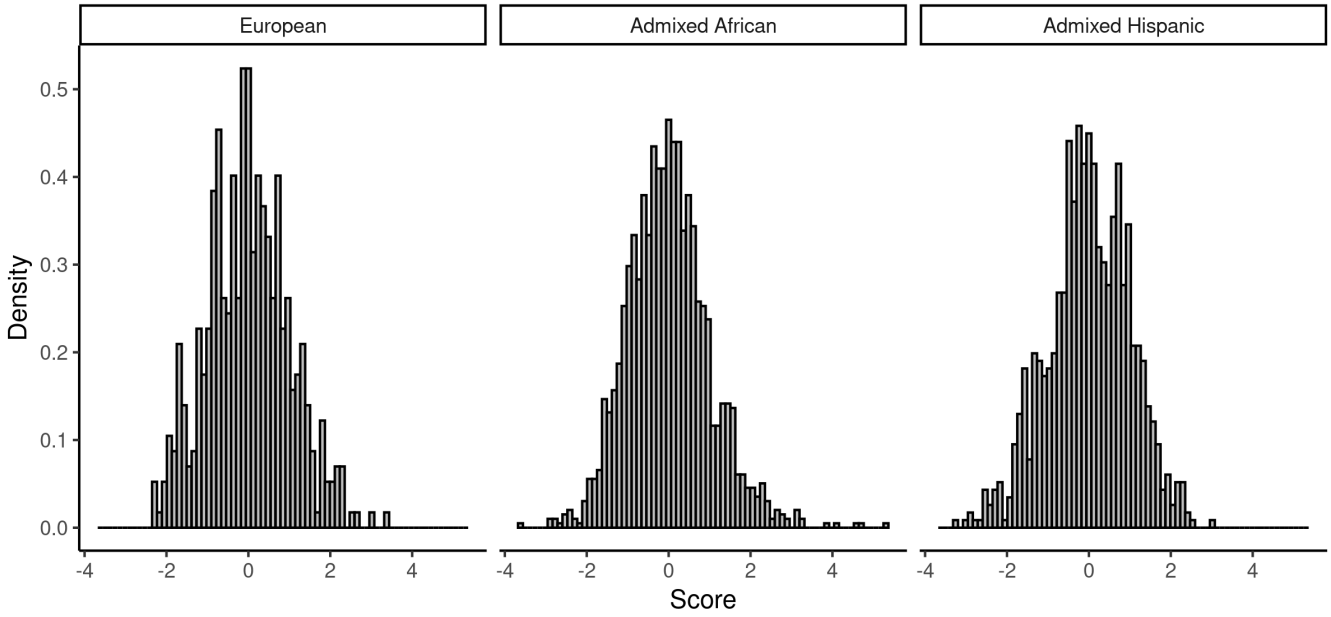
The DBP and SBP PGS were created using results from a 2018 study by the International Consortium of Blood Pressure-Genome Wide Association Studies (ICBP; Evangelou et al 2018). Discovery analyses were performed in people of European ancestry drawn from the UK Biobank and the International Consortium of Blood Pressure – Genome Wide Association Studies (Total N = 757,601). The discovery analysis included fixed-effects inverse variance weighted meta-analysis of ~7.1 Million SNPs with minor allele frequency greater than or equal to 1%. The ICBP analysis included previously reported GWAS data from 54 studies (N=150,134) plus new data from 23 additional studies (N=148,890). Full methods on these studies can be found in (**Supplementary Table 1b,** and **Supplementary Tables 20a-c**). The UK Biobank analysis included the following covariates: sex, age, age2, BMI and a binary indicator variable for UKB vs UK BiLEVE to account for the different genotyping chips. Blood pressure was assessed from the average of two automated (N=418,755) or two manual (N=25,888) BP measurements. For individuals with one manual and one automated BP measurement (N=13,521), BP was calculated as the mean of these two values. When only one BP measurement (N=413) was available, they used this single value. BP was adjusted for medication use by adding 15 and 10 mmHg to SBP and DBP, respectively, for individuals reported to be taking BP-lowering medication (N=94,289). Additional replication samples from the US Million Veterans Program (N=220,520) and the Estonian Genome Centre, University of Tartu (N=28,742) Biobanks. The UKB+ICBP summary data can be downloaded from the GWAS catalog (https://www.ebi.ac.uk/gwas/publications/30224653). After removing 274 loci (from 357 previously reported SNPs that were associated with blood pressure), the study reports 535 novel loci associated with blood pressure traits (including diastolic and systolic blood pressure, and pulse pressure).

The GLGC HDL PGS contains 206,461 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The GLGC LDL PGS contains 206,253 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The GLGC TC PGS contains 206,449 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The GLGC TG PGS contains 206,267 SNPs that overlapped between the FFCWS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ancestry to a standard normal curve (mean=0, standard deviation = 1).

Please note that the GLGC-lipid results contain PGSs from European ancestry backgrounds (see Section C. "Notes about the use of PGSs" for more information on the use of PGSs in other ancestry groups).



Distribution of diastolic blood pressure PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)



Distribution of systolic blood pressure PGS, by analytic group standardized within analytic group (European, n=475; Admixed African, n=1640; Admixed Hispanic, n=959)

References

Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, …, Million Veteran Program. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet. 2018 Oct;50(10):1412-1425. doi: 10.1038/s41588-018-0205-x. Epub 2018 Sep 17. Erratum in: Nat Genet. 2018 Dec;50(12):1755. PMID: 30224653; PMCID: PMC6284793.