Panel Study of Income Dynamics, Child Development Supplement 2014: User Guide to Genomic Data

3 December 2021
(Early Release)

Narayan Sastry, Paula Fomby, and Yi-Miau Tsai

Panel Study of Income Dynamics
Survey Research Center
University of Michigan

Abstract

The 2014 Child Development Supplement (CDS) to the Panel Study of Income Dynamics (PSID) collected saliva samples in a nationally representative sample of children aged 5–17 years and their primary caregivers—in addition to collecting extensive and detailed data on children's health, development, and well-being. The CDS-2014 saliva samples were used to obtain genomic data, which are now being provided to the research community through three components. First is a set of public use files with polygenic scores for a number of social, psychological, and health outcomes calculated from genome-wide association study results. Second is a restricted data file that describes the genetic relatedness among all pairs of sample members that is available to researchers through a restricted data use contract. Third is the full set of genetic sequence data in the form of a Variant Call Format data file that is available to users who submit an application package that is approved by PSID. A separate restricted data application is required for data users who want to link measures from the full genetic sequence data to PSID data at the individual level. This User Guide to Genomic Data provides essential information to researchers planning or undertaking research using the CDS-2014 genomic data.

Suggested citation of the CDS-2014 User Guide:

> Sastry, N., P. Fomby, and Y-M. Tsai. “Panel Study of Income Dynamics, Child Development Supplement 2014: User Guide to Genomic Data,” Institute for Social Research, University of Michigan, 2021.

Suggested citation of the CDS-2014 data:

> Child Development Supplement to the Panel Study of Income Dynamics, public use dataset [restricted use data, if appropriate]. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (year data were downloaded).

Suggested acknowledgement of the CDS-2014 data:

> The collection of data used in this study was supported by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development under grant number R01 HD052646 (Narayan Sastry, Principal Investigator).

Preface

The 2014 Child Development Supplement (CDS) to the Panel Study of Income Dynamics (PSID) was funded by Grant R01 HD052646 (Narayan Sastry, Principal Investigator) from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD). Funding for CDS-2014 genetic sequencing was provided by NICHD and by the Russell Sage Foundation and the Ford Foundation through Grant 96-17-02 (Narayan Sastry, Principal Investigator). PSID is very grateful to these sponsors for their support.

Contents

Acknowledgements

The effort to collect, process, and disseminate genomic data from the 2014 wave of the Child Development Supplement (CDS) to the Panel Study of Income Dynamics (PSID) was based at the Survey Research Center in the Institute for Social Research at the University of Michigan. The effort was directed by Narayan Sastry in collaboration with Paula Fomby, the current Associate Director of CDS, and Yi-Miau Tsai, the Project Manager for CDS.

The management of the CDS-2014 saliva samples following data collection was performed by Maryam Buagelila of the Survey Research Operations (SRO) unit and by staff members Laura Mayo-Bond and Lesli Scott of the BioSpecimen Lab located within the Institute of Social Research at the University of Michigan.

Joseph Pickrell at Gencove, Inc., directed the genetic sequencing of the CDS-2014 saliva samples with assistance from Kaja Wasik.

The genomic data were processed by a team at the Survey Research Center comprised of Colter Mitchell, Erin Ware, and Jonah Fisher. PSID project team members Flannery Campbell and Mohammad Mushtaq undertook several data management tasks associated with the final processing and release of the data.

Special acknowledgement and thanks go to Dalton Conley from Princeton University who provided valuable guidance, advice, and fundraising help to the project.

Finally, we are grateful to Regina Bures at the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development for her support of the CDS-2014 project.

1. INTRODUCTION

The Panel Study of Income Dynamics (PSID) is a longitudinal survey of a nationally representative sample of U.S. families that began in 1968.[1] The PSID Child Development Supplement was launched in 1997 with the goal of improving the understanding of social, psychological, and economic aspects of childhood within an ongoing nationally representative, longitudinal study of U.S. families.

CDS has collected information on psychological and social wellbeing, health status and behavior, family environment, education, child care, time use, sibling relationships, caregiver social and psychological resources, non-coresident parents, future work and schooling expectations, and religiosity. CDS data also support research on the relationship between children's characteristics and contemporaneous family decisionmaking and behavior. Finally, data from CDS allow researchers to study the effects of childhood factors on subsequent social, demographic, economic, and health outcomes because CDS sample members are followed into young adulthood as part of the PSID Transition into Adulthood Supplement and over the rest of the life course as part of Core PSID.

CDS-2014 marked a relaunching of CDS and a switch from a cohort design to a design in which all PSID children aged 0–17 years would be eligible for participation in CDS every five years.[2] These new data will allow studies of health, development, and well-being in childhood; the relationship between children's characteristics and contemporaneous family decisionmaking and behavior; and the effects of childhood factors on subsequent social, demographic, economic, and health outcomes over the entire life course for these individuals as they are followed into the future as part of the ongoing Core PSID.

The CDS-2014 sample included all PSID families that completed a Core PSID interview in 2013 and had one or more resident children. CDS-2014 participants form a nationally representative sample of children descended from the original 1968 families and the 1997 new immigrant refresher sample. (The CDS-2014 sample does not cover children from families in which both parents are post-1997 immigrants to the U.S.) All eligible PSID children in each family were selected for CDS-2014.

CDS-2014 was primarily a telephone interview; however, a random 50 percent of households were selected to receive an in-home visit to collect information that could not be obtained reliably by telephone. The in-home visits facilitated the collection of other study components that were otherwise collected using a mail-out/mail-back protocol, including saliva samples for subsequent genotyping.

The purpose of this User Guide to Genomic Data in CDS-2014 is to provide information about the sample and response rates, the genetic sequencing data, and the data files that are available. In Chapter 2, we describe the CDS-2014 fieldwork outcomes related to the collection of saliva samples. In Chapter 3, we describe the three separate data products that are being released as part of the CDS-2014 genomic data. The appendix provides details on the construction of the polygenic scores.

---

[1] McGonagle, K., Schoeni, R., Sastry, N., & Freedman, V. (2012). The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Course Research. *Longitudinal and Life Course Studies,* 3, 268–284.

[2] Sastry, N., & P. Fomby. (2017). Panel Study of Income Dynamics, Child Development Supplement 2014: User Guide. Institute for Social Research, University of Michigan, 2017. https://psidonline.isr.umich.edu/cds/CDS2014_UserGuide.pdf

2. COLLECTION AND PROCESSING OF SALIVA SAMPLES

In this chapter, we provide an overview of the collection of saliva samples in CDS-2014.

A total of 5,815 children in the PSID sample were deemed eligible to participate in CDS-2014.[3] CDS-2014 was fielded over a 26-week period between late October 2014 and late April 2015. Data were collected successfully on 4,333 children from 2,525 families. Accounting for 180 children who were subsequently determined to be non-sample, the overall unconditional response rate was 77 percent. The response rate conditional on completing the coverscreen interview was 88 percent, which is directly comparable to—and similar in magnitude to—the response rate for the original CDS in 1997.

Children aged 5–17 years and their primary caregiver (typically one of the child's parents) were asked to contribute saliva samples using an Oragene OG-500 DNA Genotek saliva kit for subsequent genetic analysis. Participants were asked to not eat, drink, chew gum, or smoke for thirty minutes prior to administering the saliva collection. All eligible respondents were asked to provide signed informed consent before participating in the saliva collection. Respondents were offered a $10 incentive for each saliva sample that was provided. The response rate for collecting saliva samples was 43 percent. However, there was 39 percentage point differential in response rates between those that with a home visit (63 percent) compared to those following the mail-out/mail-back protocol (24 percent).

As part of CDS-2014, a total of 2,579 saliva samples were collected and received in Ann Arbor with a valid signed consent form. Upon receipt in Ann Arbor, the saliva samples were logged and verified. The samples were initially stored at room temperature. Between August and September 2016, the saliva samples were transferred to the ISR BioSpecimen Lab (BSL) for processing and storage. As samples arrived, the ISR-BSL coordinator created the individual sample identifier labels and precise storage locations for each biospecimen. Between August and October 2017, the ISR-BSL carried out heat-treatment in a Thermo Heratherm Incubator (Model IMH100) for a minimum period of 8 hours and a maximum of 52 hours per DNA Genotek guidelines. The samples were then transferred to new 2 ml cryogenic sterile pyrogen-free tubes that were appropriate for -80 degree Celsius freezer storage. Samples were transferred into two panels, one 0.6 ml and the other 2.6-3.0 ml. IDs for each sample were verified at each step of the processing. Samples were stored in a freezer at -80 degrees Celsius.

Genetic sequencing of the CDS-2014 saliva samples was conducted in 2018 by Gencove, Inc., using ultra low-coverage genome sequencing with average 0.7X coverage using NovaSeq 5000/6000.[4] [5] [6] The CDS-2014 saliva samples were transferred to Gencove, Inc., on 20 February 2018 and the remaining saliva samples were retuned on 21 June 2018.

---

[3] Sastry, N., & P. Fomby. 2017. "Panel Study of Income Dynamics, Child Development Supplement 2014: User Guide." Institute for Social Research, University of Michigan. https://psidonline.isr.umich.edu/cds/CDS2014_UserGuide.pdf

[4] Gilly, A., Southam, L., Suveges, D., Kuchenbaecker, K., Moore, R., Melloni, G.E., Hatzikotoulas, K., Farmaki, A.E., Ritchie, G., Schwartzentruber, J. & Danecek, P., 2019. Very low-depth whole-genome sequencing in complex trait association studies. Bioinformatics, 35(15), pp.2555-2561.

[5] Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P. & Sullivan, P.F., 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nature Genetics, 44(6), pp.631-635.

[6] Li, J.H., Mazur, C.A., Berisa, T. & Pickrell, J.K., 2021. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Research, 31(4), pp.529-537.

Of the 2,579 valid saliva samples that were received in Ann Arbor and sent to Gencove for genetic sequencing, a total of 66 samples are excluded from the final data release. Genomic data could not be obtained by Gencove for 37 samples. Of the remaining 29 samples that were returned from Gencove, 16 cases had problematic genomic data as determined by various quality control checks and 13 cases were determined to be invalid through quality control checks based on the sample composition. The latter quality control checks indicated, for example, instances in which the same person apparently provided saliva samples for several different family members. The final number of cases with genomic data is 2,513 individuals. Table 1 summarizes the disposition of valid saliva samples collected in CDS-2014.

Table 1. Disposition of Saliva Samples Collected in CDS-2014

Description	Count	Percent
Saliva samples collected	2,579	100.0%
No genomic data from lab	37	1.4%
Removed after quality control checks of genomic data	16	0.6%
Removed after genetic relatedness checks	13	0.5%
Final released sample	2,513	97.4%

3. GENOMIC DATA

In this chapter, we describe the genomic data available from CDS-2014. Three separate data files are available as part of the CDS-2014 Genomic Data Early Release in November 2021. We describe each of these files and the process to obtain them. Please note that these data are classified as early release—which means that they have undergone limited checking and testing. Data users are encouraged to provide comments and feedback on these data to the PSID project through the helpdesk: psidhelp@umich.edu. Suggestions are welcome for additional measures, such as new polygenic scores, that can be constructed and released from the CDS-2014 genomic data.

Users analyzing the Early Release CDS-2014 genomic data should be aware that no separate sample weight has been developed yet for use with these measures. We plan to develop such a weight for a subsequent data release.

3.1 Polygenic Scores and Genetic Ancestry

Polygenic scores (PGS) are genetic risk indices for observing a particular phenotype that are constructed from results of genome-wide association studies. PGS have been developed for a substantial number of phenotypes, such as educational attainment, tobacco use, obesity, and other health and developmental measures, that can serve as independent predictors (and control variables) or as moderators of social influences on behavior.[7] [8] Polygenic scores provide an attractive approach for including genetic information in standard models for social and behavioral outcomes, thereby reducing bias due to unobserved heterogeneity among purely "social" variables and increasing statistical efficiency.

As part of this Early Release of CDS-2014 genomic data, we have constructed and made available a set of 27 polygenic scores, which are summarized in Table 2, across 22 specific domains. These variables cover outcomes related to reproduction (age at menarche and at first birth), social outcomes (educational attainment and subjective well-being), personality measures (extraversion and neuroticism), several mental health measures (including anxiety, depressive symptoms, post-traumatic stress disorder, and autism), substance use (cannabis and alcohol dependence), anthropometry (waist circumference and waist-to-hip ratio), several physical health measures (measures of blood pressure and blood lipids), and chronic or acute health conditions (type-2 diabetes, blood urea nitrogen, and myocardial infarction). Appendix A provides detailed information about how the polygenic scores were constructed and identifies the genome-wide association study on which each score is based.

The files with PGS include two additional sets of variables. The first are two pairs of identification variables to allow the data to be merged with other data from CDS-2014 or from Core PSID or other PSID supplements. See Chapter 4 of this User Guide for detailed information on data merging. The second set of variables are three global genetic principal components and an indicator of the assigned ancestry group derived from the genetic principal components. The genetic principal components were constructed globally—i.e., for the entire sample together—rather than separately by race/ethnic group because of small sample sizes and relatedness among sample members of the same race and ethnicity.

---

[7] Belsky, D.W., & S. Israel. (2014). Integrating genetics and social science: Genetic risk scores. Biodemography and Social Biology, 60(2): 137-155.

[8] Harden, K.P., & Koellinger, P.D. (2020). Using genetics for social science. Nature Human Behaviour, 4(6), 567-576.

Note that the CDS-2014 polygenic scores are provided in three separate files based on ancestry, separating cases classified as being of European, African, or admixed ancestry. This classification of ancestry is based on reported race and ethnicity from the CDS-2014 or PSID data and the genetic ancestry assignment based on the three global genetic principal components. Data users should use caution before pooling data across ancestries and analyzing them together. We reinforce this warning by providing within-ancestry standardization of the polygenic scores to have a mean of zero and a standard deviation of one. In the future, improved measures of race and ethnicity will be used in conjunction with the genetic ancestry assignment to verify the assignment of individuals to specific ancestry files, which may result in reassignments.

The CDS-2014 polygenic scores are available to users as Early Release public use data and can be downloaded from "Early Release" section of the "Packaged Data" section of the CDS Online Data Center (http://cds-tas.org/) or the PSID Online Data Center (http://psidonline.isr.umich.edu/).

3.2 Genetic Relatedness

The second CDS-2014 genomic data release is a single file available as restricted data to approved users through PSID's virtual data enclave. This file includes the genetic relatedness coefficient for all pairs of individuals with genomic data.

The relatedness measure is based on a calculated coefficient of relatedness that summarizes the correlation in genomic data for all participant pairs. The genetic relatedness measure can be used to characterize the biological relatedness among pairs of observations, with, for example, a value of 1.0 for identical twins, 0.5 between a parent and child and between full siblings, and 0.25 between half-siblings. The relatedness measure was derived through the identity-by-descent function within the PLINK software Version 1.9 (see https://www.cog-genomics.org/plink/1.9/ibd).

The relatedness file includes an additional set of variables, which is two pairs of identification variables to allow the data to be merged with other data from CDS-2014 or from Core PSID (or other PSID supplements). These identification variables are available for both individuals for whom the corresponding relatedness coefficient is calculated.

The relatedness file includes a total of 3,156,328 observations, which corresponds to the lower diagonal elements of the symmetric relatedness matrix. The count is obtained as the squared number of cases minus the number of cases divided by two: $((2{,}513)^2 - 2{,}513) / 2 = 3{,}156{,}328$.

Users interested in obtaining the CDS-2014 genetic relatedness file can learn more about the application process here: https://simba.isr.umich.edu/restricted/RestrictedUse.aspx.

3.3 Variant Call Format Genomic Data

The third CDS-2014 genomic data release component is the full set of genomic data in Variant Call Format (VCF). As part of the early release of the CDS-2014 genomic data, the VCF data are being distributed by PSID to data users who complete an application process (described below). We ultimately plan to make these same data available to users through the database of Genotypes and Phenotypes (dbGaP) at the National Institutes of Health (see https://www.ncbi.nlm.nih.gov/gap/).

The CDS-2014 VCF data are available in compressed format that requires approximately 0.5 terabytes of disk space. Also included with the CDS-2014 VCF data is a file with a limited set of variables described in Table 3, including ID variables and a small set of phenotypic data.

The application procedure for the CDS-2014 VCF data will mirror the approach required when the data are released through dbGaP. There are data use limitations for the CDS-2014 VCF data—in particular, data users must provide documentation of local IRB approval and use of data is limited to not-for-profit organizations. Users will need to complete a Data User Certification Agreement indicating their agreement to the terms of access under which the CDS-2014 VCF data are being made available.

Data users who would like to link summary measures derived from the CDS-2014 VCF data to other measures available through the CDS or PSID data archives will need to obtain a restricted data use agreement through the application process described here: https://simba.isr.umich.edu/restricted/RestrictedUse.aspx.

3.4 CDS-2014 Race/Ethnicity Variable

A variable describing the race/ethnicity of each CDS-2014 participant with genomic data is provided in the VCF Genomic Data File and is also used to stratify the polygenic score data by participants' ancestry based on a procedure that uses both this race/ancestry variable and global genetic principal components (see Section 2.4 in Appendix A).

The race/ethnicity variable is constructed for CDS-2014 individuals using data from the PSID Online Data Center (psid.org). Race/ethnicity of CDS-2014 children and primary caregivers was constructed in separate steps and then combined into a single race variable, with the same five categories: non-Hispanic white, non-Hispanic black, Asian, Hispanic, and other.

A child's race/ethnicity classification is based on a parental report (including a report by an adoptive parent). If a child has more than one record of race/ethnicity reported, his/her race/ethnicity is assigned the first available report from (and in the sequence of) the biological mother, biological father, adoptive mother, or adoptive father. A child's race/ethnicity is coded as "other" when no records are available, or when two or more categories of race are mentioned by parents for non-Hispanic individuals. The distribution of race/ethnicity variable for CDS-2014 children is 43.6% non-Hispanic white, 39.2% non-Hispanic black, 0.6% Asian, 6.1% Hispanic, and 10.5% other.

Race variable for CDS-2014 primary caregiver is constructed from information collected in the 2013 wave of the PSID main interview. A PCG's race is either self-reported or reported by a spouse or partner depending on who is the 2013 PSID interview respondent. The variables used are L39 Spanish Descent – Head, L40 Race of Head (mention 1 through mention 4), K39 Spanish Descent – Wife, and K40 Race of Wife (mention 1 through mention 4). A PCG's race is coded as "other" when no records are available, or when two or more categories of race are mentioned for PCGs who are non-Hispanic. Using these steps, the distribution of PCG's race/ethnicity is 43.8% non-Hispanic white, 34.6% non-Hispanic black, 0.9% Asian, 7.5% Hispanic, and 13.0% other.

Table 2. CDS-2014 Polygenic Scores

Variable Name	Variable Domain and Description
	A. Reproduction
G14PGSA1	AGE AT MENARCHE - POLYGENIC SCORE
G14PGSA2	AGE AT FIRST BIRTH - POLYGENIC SCORE
	B. Social Outcomes
G14PGSB1	EDUCATIONAL ATTAINMENT - POLYGENIC SCORE
G14PGSB2	SUBJECTIVE WELLBEING - POLYGENIC SCORE
	C. Personality
G14PGSC1	EXTRAVERSION - POLYGENIC SCORE
G14PGSC2	NEUROTICISM - POLYGENIC SCORE
	D. Mental health
G14PGSD1	ANXIETY - POLYGENIC SCORE
G14PGSD2	ANTISOCIAL BEHAVIOR - POLYGENIC SCORE
G14PGSD3	DEPRESSIVE SYMPTOMS - POLYGENIC SCORE
G14PGSD4	PTSD - POLYGENIC SCORE
G14PGSD5	OBSESSIVE COMPULSIVE D - POLYGENIC SCORE
G14PGSD6	BIPOLAR DISORDER - POLYGENIC SCORE
G14PGSD7	CROSS-PSYCH DISORDER - POLYGENIC SCORE
G14PGSD8	AUTISM - POLYGENIC SCORE
	E. Substance Use
G14PGSE1	CANNABIS USE - POLYGENIC SCORE
G14PGSE2	ALCOHOL DEPENDENCE - POLYGENIC SCORE
	F. Anthropometry
G14PGSF1	WAIST CIRCUMFERENCE - POLYGENIC SCORE
G14PGSF2	WAIST-TO-HIP RATIO - POLYGENIC SCORE
	G. Physical Health Measures
G14PGSG1	SYSTOLIC BP - POLYGENIC SCORE
G14PGSG2	DIASTOLIC BP - POLYGENIC SCORE
G14PGSG3	TOTAL CHOLESTEROL - POLYGENIC SCORE
G14PGSG4	LDL CHOLESTEROL - POLYGENIC SCORE
G14PGSG5	HDL CHOLESTEROL - POLYGENIC SCORE
G14PGSG6	TRIGLYCERIDE - POLYGENIC SCORE
	H. Chronic or Acute Health Conditions
G14PGSH1	DIABETES TYPE 2 - POLYGENIC SCORE
G14PGSH2	BLOOD UREA NITROGEN - POLYGENIC SCORE
G14PGSH3	MYOCARDIAL INFARCTION - POLYGENIC SCORE

Table 3. Additional Variables Provided with the CDS-2014 VCF Data

Variable Name	Variable Description
IID	DE-IDENTIFIED SUBJECT ID
SEX	SEX
RACE	RACE/ETHNICITY – 5 CATEGORIES
AGEGRP	AGE AT COLLECTION [0-9,10-19,…]
ADHEIGHT	HEIGHT IN INCHES FOR ADULT
CHHEIGHT	HEIGHT Z-SCORE FOR CHILD

4. MERGING GENOMIC DATA WITH CDS AND PSID DATA

In this chapter, we describe how to merge data from the CDS-2014 Genomic Data Early Release with data from CDS-2014 and other PSID components.

The early release polygenic scores file and genetic relatedness file each contain information on 2,513 individuals who participated in CDS-2014 for whom valid genomic data are available, including 1,413 children and 1,100 primary caregivers. The polygenic scores files include one record for each individual (N=2,513, summing across the three ancestry group files). The genetic relatedness file includes one record for each pair of individuals (N=3,156,328). The third genomic data component, which is the full set of genomic data in Variant Call Format (VCF), includes a de-identified subject identification number that cannot be linked directly with data from CDS-2014 or any other PSID components. However, data users can submit a PSID restricted data application to obtain linked variables from PSID or to use variables derived from the VCF genomic data within the PSID virtual data enclave that is used for restricted data analysis.

The CDS polygenic scores files and genetic relatedness file each include two sets of unique identifiers for each individual. The first set of identifiers allows users to merge to records in CDS-2014. The second set allows users to merge to individual-level records from the PSID main interview. The first set of identifiers is the same on each of these two genomic data files. The second set of identifiers is different between the polygenic scores files and the genetic relatedness file.

4.1 Merging Genomic Data to Records in CDS-2014

Two variables are required in order to merge records from the genomic data files with other CDS-2014 components.

On the **polygenic scores files**, G14CDSHID uniquely identifies CDS-2014 households and G14INST uniquely identifies each individual's roster position within their CDS-2014 household. (Note that values of G14INST repeat across CDS-2014 households.) Together, G14CDSHID and G14INST uniquely identify each individual.

On the **genetic relatedness file**, G14CDSHID1 uniquely identifies the CDS-2014 household and G14INST_1 uniquely identifies roster position within a CDS-2014 household for the first individual in a pair. G14CDSHID2 and G14INST_2 identify CDS-2014 household and roster position for the second individual in the pair.

The CDS-2014 household roster file provides an anchor to merge records between the genomic data files and other CDS-2014 components. The CDS-2014 household roster file is available as part of the complete CDS-2014 ZIP file on the Packaged Data page at the PSID Data Center (psid.org) or as a standalone file obtained using the File option to select variables at either the PSID Data Center or the CDS-TAS Data Center (cds-tas.org). The equivalent household and individual identifiers on the household roster file are R14CDSHID and R14INST. To merge observations between the **polygenic scores files** and the household roster file, first rename the variables on one file to match variable names on the other (e.g., on the household roster file, rename R14CDSHID to G14CDSHID and rename R14INST to G14INST). Then merge the two files together on these commonly-named variables.

To merge observations between Person 1 of each pair on the genetic relatedness file and the household roster file, rename R14CDSHID and R14INST to G14CDSHID1 and G14INST_1. To merge observations for Person 2 of each pair, rename the identifiers on the household roster with the suffix *2* in place of *1*. Note that each combination of unique identifiers for each person appears on the genetic relatedness on multiple records (2,513 occurrences as person 1 and 2,512 occurrences as person 2) but appears on the household roster file only once.

For guidance on how to conduct further merges to content on other CDS-2014 data files, refer to the CDS-2014 User Guide.[9]

4.2 Merging Genomic Data to Records in Other PSID Components

Users of the CDS-2014 genomic data files may wish to attach information that was collected in other PSID interview components, such as the PSID main interview. The required identifiers and strategy to achieve this differ between the polygenic scores files and the genetic relatedness file.

Merging Records from Polygenic Scores Files

To merge records on the polygenic scores files to other PSID components requires the identifier variables assigned to individuals in the family units they occupied during the 2013 PSID main interview. These identifier variables are different from the identifiers assigned on the CDS-2014 household roster described above. G14ID uniquely identifies family units (households) that participated in the 2013 PSID main interview. G14SN uniquely identifies individuals within their 2013 family units by their rostered sequence numbers. (Note that values of G14SN repeat across family units that participated in the 2013 main interview.) Together, G14ID and G14SN uniquely identify each individual.

The equivalent family unit and individual identifiers from the 2013 PSID main interview are available in the PSID and CDS-TAS Data Centers under the names ER34201 and ER34202 respectively. To merge records between the genomic data file and the 2013 PSID main interview, first rename the variables on one file to match variable names on the other (e.g., on the 2013 PSID main interview file, rename ER34201 to G14ID and rename ER34202 to G14SN). Then merge the files together on these two commonly-named variables.

Note that some CDS-2014 primary caregivers were not present in a family unit at the time of the PSID 2013 main interview. As a result, their records from the genomic data files will not match to any individual-level record from that main interview wave, and the number of records that successfully merge will be less than 2,513.

Merging Records from the Genetic Relatedness File

In addition to the unique identifiers described above, all individuals ever observed in PSID have a unique pair of time-invariant identifiers. These are the “1968 Family Interview Number” (ID68) and “Person Number” (PN). More information about these identifiers is available in PSID main interview main interview user guide.[10]

---

[9] Sastry, N., & P. Fomby. (2017). Panel Study of Income Dynamics, Child Development Supplement 2014: User Guide. Institute for Social Research, University of Michigan, 2017. https://psidonline.isr.umich.edu/cds/CDS2014_UserGuide.pdf

[10] PSID Main Interview User Manual: Release 2013. Institute for Social Research, University of Michigan, 2013. https://psidonline.isr.umich.edu/data/Documentation/UserGuide2013.pdf

These variables appear on the genetic relatedness file as GID68_1 and GPN_1 for person 1 in each pair and as G1D68_2 and GPN_2 for person 2 in each pair. They will appear as ER30001 (1968 family interview number) and ER30002 (person number) on any data extract from the PSID Data Center or CDS-TAS Data Center that includes individual-level records. To merge records between the genetic relatedness file and any extract, first rename the variables on one file to match variable names on the other. For example, if you wish to merge on information from a PSID data extract for person 1 on the genetic relatedness file, rename ER30001 to GID68_1 and rename ER30002 to GPN_1 in the PSID data extract). Then merge the files together on these two commonly-named variables.

Again, bear in mind that the genetic relatedness file includes many records for each set of unique identifiers, while the PSID data extract will include (at most) one record per person. Also note that a small number of CDS-2014 primary caregivers were never observed in a PSID family unit during any PSID main interview wave and so will not have a record in the PSID data extract. As a result, the number of individuals for whom genetic relatedness records are successfully merged will be less than 2,513.

Table 4 summarizes the names and locations of variables required for merging genomic data files to individual-level records in other CDS-2014 and PSID components.

Table 4. Variables Required for Merging CDS-2014 Genomic Data

	Merging to CDS-2014		Merging to PSID data extracts			
			Merging to 2013 PSID Main Interview		Merging to Any PSID Wave	
	Household ID	Position number	2013 Family interview ID	2013 sequence number	1968 family interview ID*	Person number*
Polygenic scores	G14CDSHID	G14INST	G14YRID	G14SN		
Genetic relatedness	G14CDSHID1, G14CDSHID2	G14INST_1, G14INST_2			G14ID68_1, G14ID68_2	G14PN_1, G14PN_2
Corresponding identifiers for merging	H14CDSHID	H14INST	ER34201	ER34202	ER30001	ER30002
Location of corresponding identifiers	CDS-2014 household roster		2013 PSID main interview, individual-level data		Any PSID data extract with individual-level records	

* The 1968 family interview ID and person number are also available on the CDS-2014 household roster as R14ID68 and R14PN.

Appendix A

CDS-2014 Polygenic Scores – Early Release 1

Erin Ware
Jonah Fisher
Colter Mitchell

Last compiled on 2021-11-19

Contents

1 Introduction

1.1 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 polygenic score (PGS) aggregates thousands to millions of individual loci across the human genome and weights them by effect sizes derived from a genome wide association study (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.

2 Polygenic Score Construction

Although 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) the 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. 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 CDS 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: $PGS_i = \sum_{j=1}^{J} W_j G_{ij}$, where i is an individual (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-2 polygenic score calculation program.

2.1 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 CDS SNP data.

2.2 Notes about the Use of PGSs

We provide polygenic scores separately for a European analytic group, an African analytic group, and an admixed analytic group. 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).

2.3 Genetic Processing

Gencove takes a low-pass sequencing (in our case 0.7x) approach to "sequence" the genomes of individuals. Roughly 45,000 variants (+/-1000) are sequenced on each individual, and overlap between any two people is fairly low. Gencove uses the sequenced variants to impute to a 1000G reference panel, one person at a time, resulting in roughly 37 million variants common across all individuals. No typical sample-level quality metrics are available for the imputed data, though a posterior genotype probability (i.e., P(AA), P(AB), P(BB)) can be used to capture variants with high probabilities of a specific call (accurately imputed).

In an example data set of 2,513 participants, these 37 million variants were reduced to the overlap of 1000 genomes SNPs that are used in GWAS (9,358,837) using three different prior probability thresholds: 0.8 (highly liberal), 0.9 (typical threshold), 0.95 (conservative). From here, we can remove variants with higher levels of missing individuals, and individuals with high levels of missing variants comparable to a standard genotyping chip. These QC metrics can be indicators of either poor quality samples or difficult-to-accurately-impute variants. In SNP-chip analyses, we use thresholds around 1-2%. In this case, such a conservative threshold eliminated all individuals and upwards of 6-8 million variants (out of 9 million). At a prior probability threshold of 0.9 and with a relaxed threshold of 5% for both criteria, we retained 90.1% of individuals in the sample and 7.3 million variants (78.5% of the 9 million 1000G variants). Further analysis of heterozygosity (the proportion of the variants that are heterozygous "AB" as opposed to homozygous "AA" or "BB") revealed multiple individuals with low heterozygosity. Indeed, increasing the strictness of the prior probability threshold (from 0.8 to 0.95) shifted the distribution of heterozygosity values lower—indicating many heterozygous calls are not imputed with high probability, as expected. These remaining variants should not be used for imputation and thus, the maximum number of variants would remain somewhere between 1-8 million with lower-than-anticipated heterozygotes. This does not account for allele frequency of the variants, which may reduce the total number of variants further.

2.4 Genetic Ancestry Assignment

Global genetic principal component (PC) analysis was performed to identify population group outliers and to provide sample convectors as covariates in the statistical model used for association testing to adjust for possible population stratification. After adding the 1000G known ancestry dataset to our QCed CDS sample, we selected SNPs for PC analysis by linkage disequilibrium (LD) pruning from an initial pool consisting of all autosomal SNPs that were present in both the CDS and 1000G data sets. In addition, we excluded the HLA, 8p23, and 17q21.31 regions from the initial pool. We identified analytic genetic groups in the CDS sample through PC analysis on genome-wide SNPs calculated across all participants plus 1000G samples using the aforementioned filtering criteria. This is based on the assumption that the most variation in this combined dataset will be derived from the genetic ancestry, and not within-family genetic similarity from the CDS sample. This assumption was satisfied by a visual check of the first few principal components. Figure 1, Panel A shows the first two principal components painted by 1000G super-population (African, American/Mexican, East Asian, European, South Asian, and the CDS sample). Figure 1, Panel C shows the second and third principal components with the same color legend as Panel A. Using these plots as guides, ancestry was assigned to the CDS sample following the bounds of the 1000G known genetic ancestry clusters. Figure 1, Panels B and D remove the 1000G samples from the plots and color the CDS sample with the same super-population scheme as 1000G (substituting mixed ancestry/no classification). Once these groups were defined, we took the union of estimated genetic ancestry from the principal components and reported race and ethnicity to create our analytic groups (Figure 2).

Figure 1. Genetic Principal Components (1–3) by Ancestry for 1000G and CDS

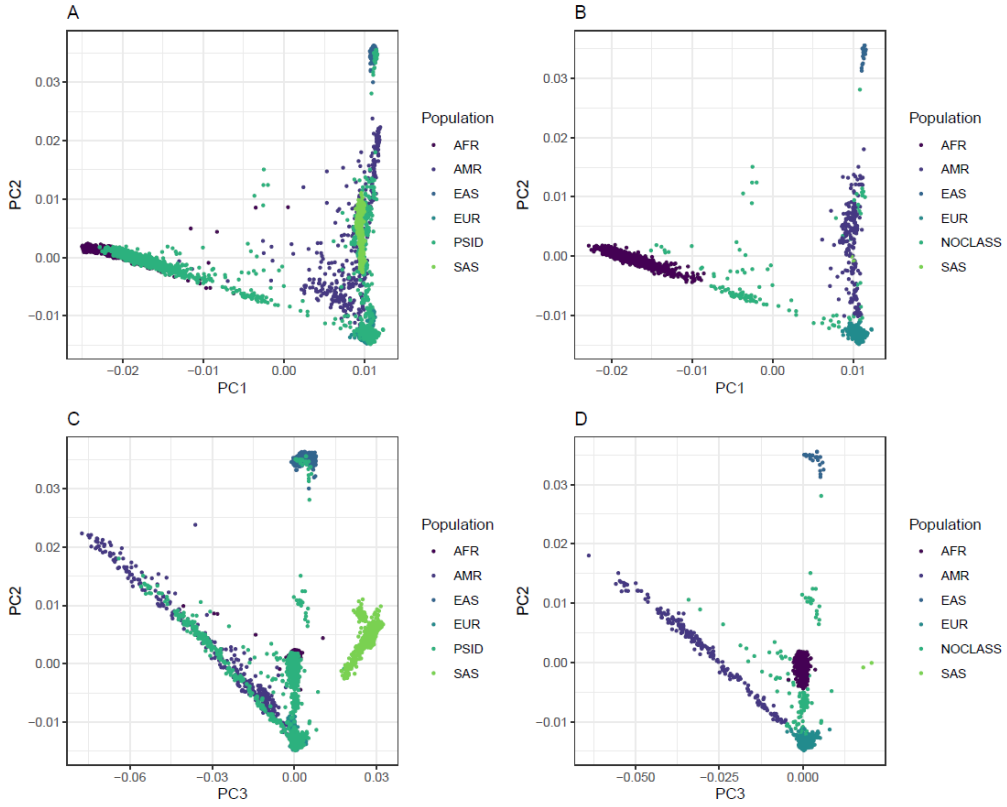
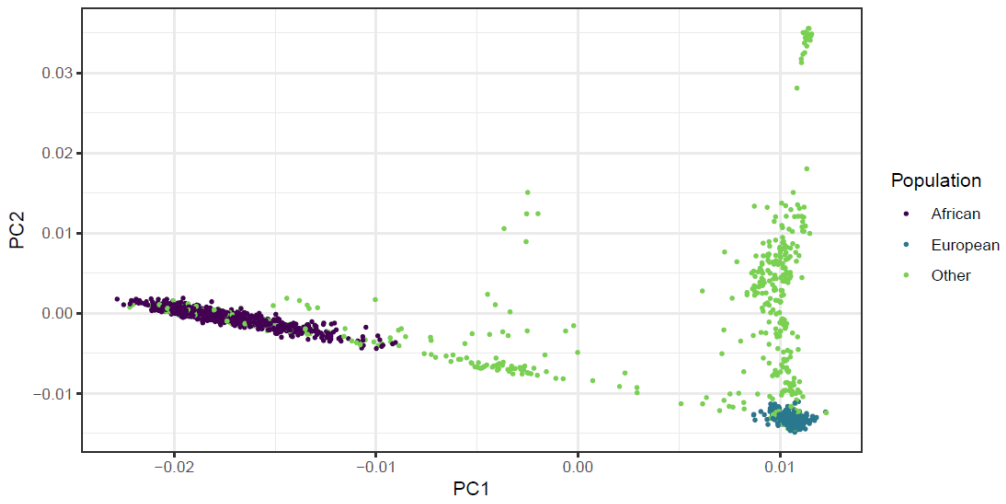


Figure 2. Assigned Ancestry Groups for CDS



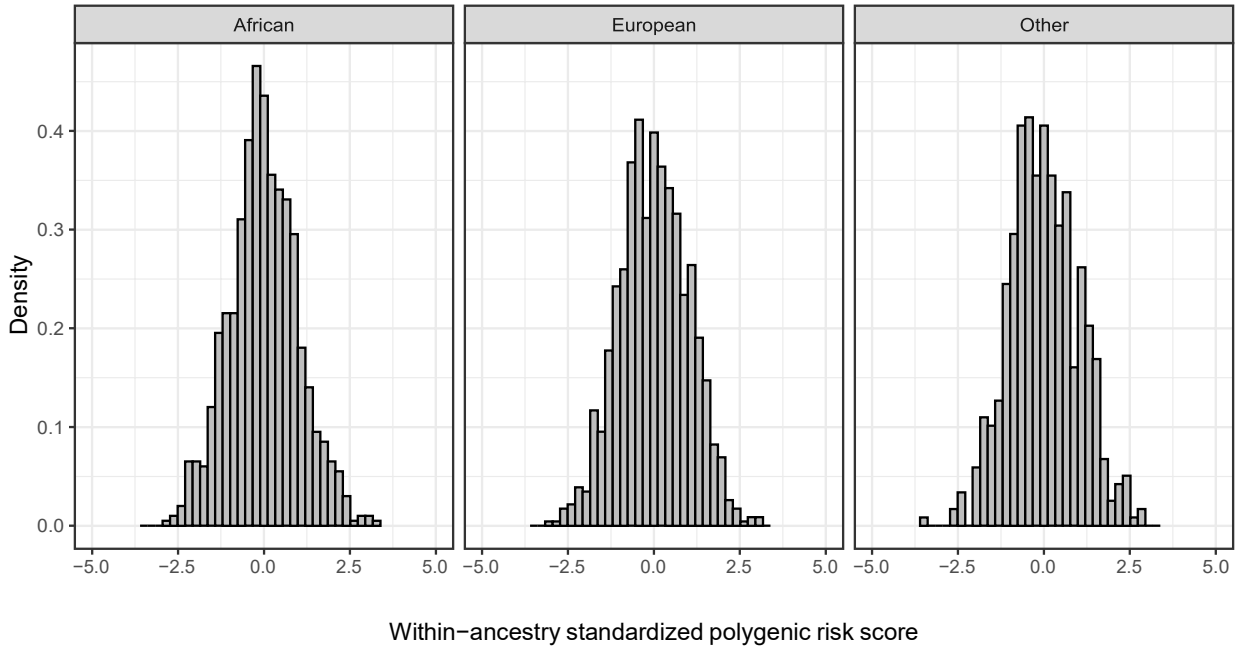
3 Polygenic Score GWAS Descriptions and Distributions

3.1 Age at Menarche

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 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 1,155,737 SNPs that overlapped between the CDS 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.



References

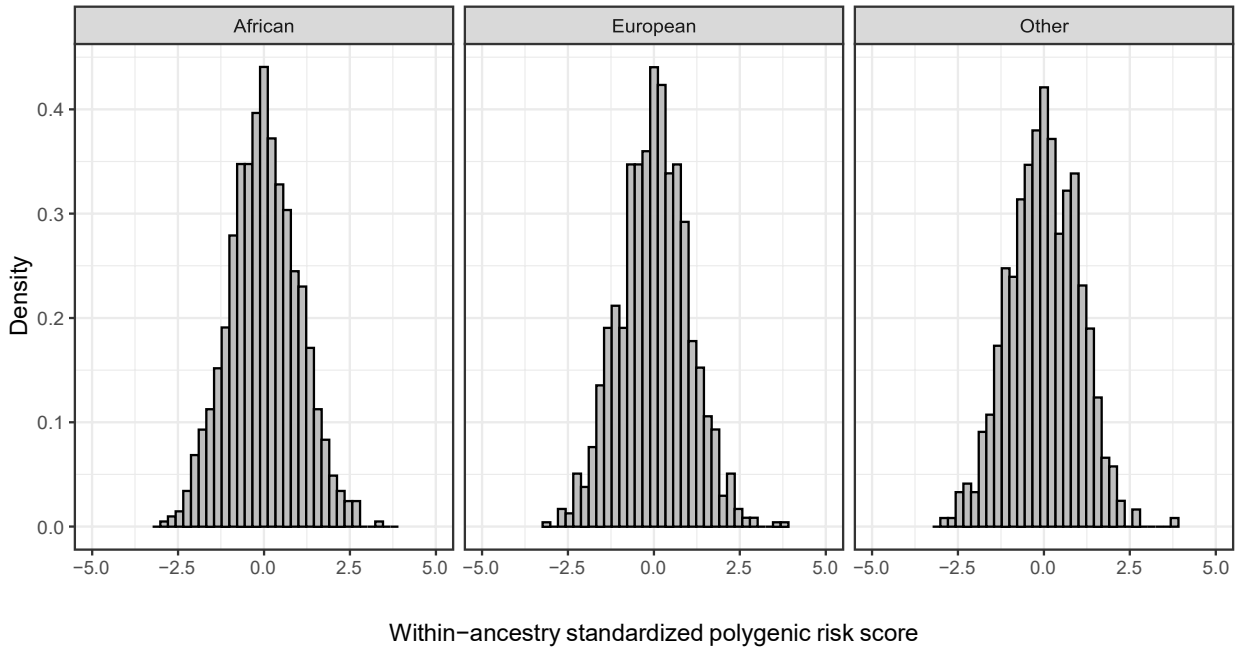
Perry, J. R., Day, F., Elks, C. E., Sulem, P., Thompson, D. J., Ferreira, T., … & Albrecht, E. (2014). Parent-of-origin- specific allelic associations among 106 genomic loci for age at menarche. Nature, 514(7520), 92-97.

3.2 Age at First Birth – Sociogenome Consortium 2016

PGSs for age at first birth were created using results from a 2016 study conducted by the Sociogenome Consortium. The GWAS meta-analysis files are publicly available on the Sociogenome data download page: http://www.sociogenome.com//data/ (AgeFirstBirth_Pooled.txt; AgeFirstBirth_Female.txt; AgeFirstBirth_Male.txt). The meta-analysis included 251,151 men and women from 62 cohorts of European ancestry. 2.4 million SNPs imputed from NCBI Build 37 HapMap phase 2 data passed quality control filters. Associations were adjusted for principal components to reduce confounding by population stratification, as well as for respondent birth year and its square and cube to control for nonlinear birth cohort effects. A single genomic control at the cohort level was applied and meta-analysis results were obtained using a sample-size-weighted fixed-effect method in METAL. AFB was only assessed for those who were parous. Meta-analysis results are reported for men and women combined and separately. The study identified ten genome-wide significant SNPs for combined results, nine of which were significantly associated in both sexes combined, and one of which was associated in women only (n=154,839) (Figure 1a and Table 1).

The PGS contains 1,165,034 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the age at first birth 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.



References

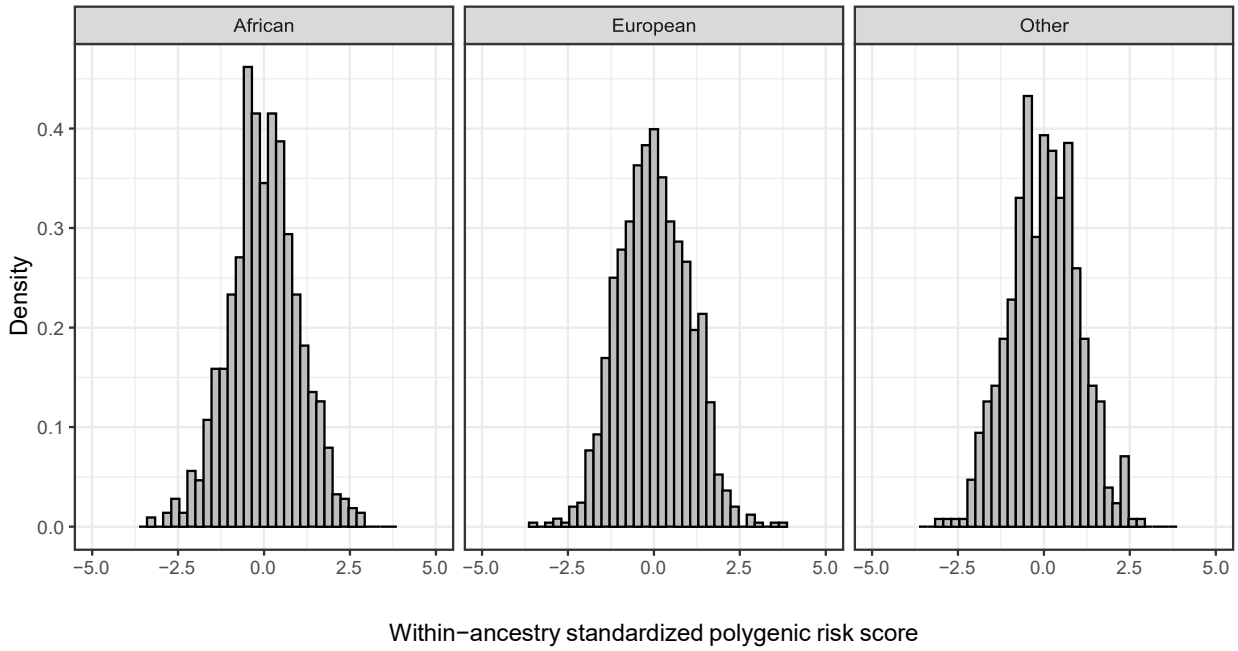
Barban, N., Jansen, R., De Vlaming, R., Vaez, A., Mandemakers, J. J., Tropf, F. C.,…& Tragante, V. (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. Nature genetics,48(12), 1462.

3.3 Educational Attainment – Social Science Genetic Association Consortium 2018

The educational attainment PGSs were created using results from a 2018 study by the Social Science Genetic Association Consortium (SSGAC). The meta-analysis included 405,073 individuals in the combined discovery and replication sample and 726,808 individuals that did not contribute to the analyses of the previous study and were used as replication in this study (total of 1,131,881 individuals). Genome-wide significant SNPs were identified in 1,271 loci (Supplementary Information table 2_1). Approximately 10.2 million SNPs were included in the analyses, with all cohorts utilizing SNPs imputed to the 1000 genomes reference panel (1000G). 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.

The PGS contains 1,381,562 individuals.

Please note that the SSGAC 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.



References

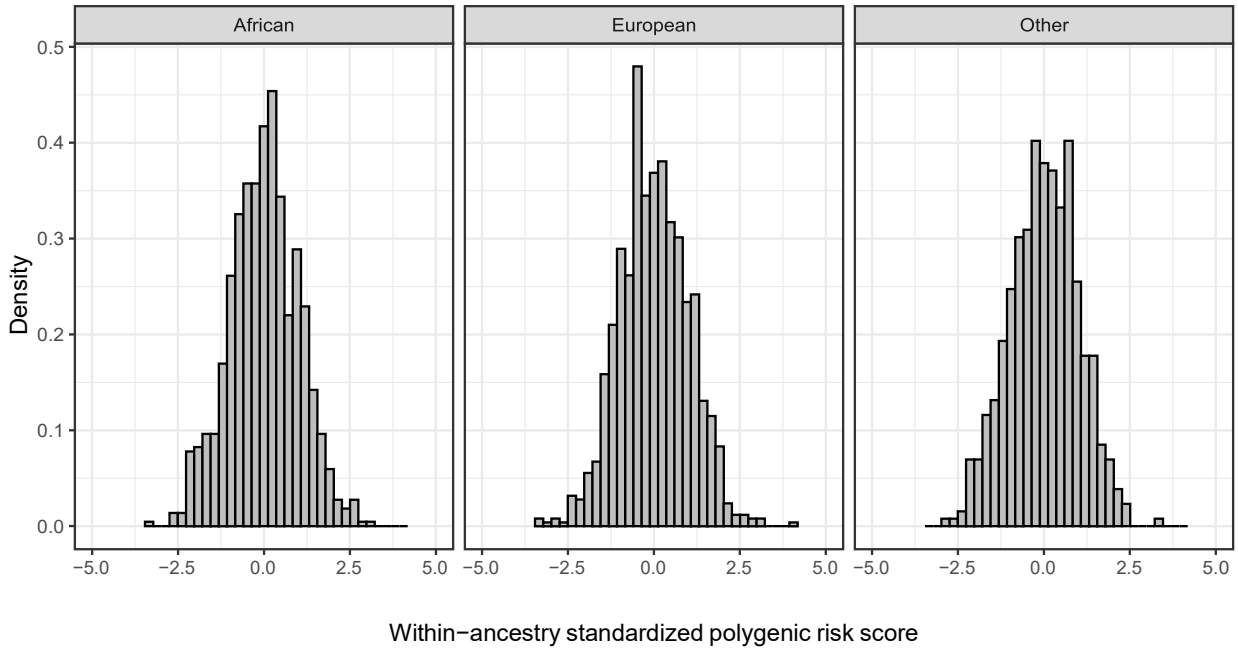
Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, & Cesarini D. (2016). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature genetics, doi: 10.1038/s41588-018-0147-3

3.4 Subjective Wellbeing – Social Science Genetic Association Consortium2016

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 European ancestry PGS contains 1,066,071 SNPs that overlapped between the CDS genetic database andthe GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normalcurve (mean=0, standard deviation=1).

Please note that the SSGAC 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.



References

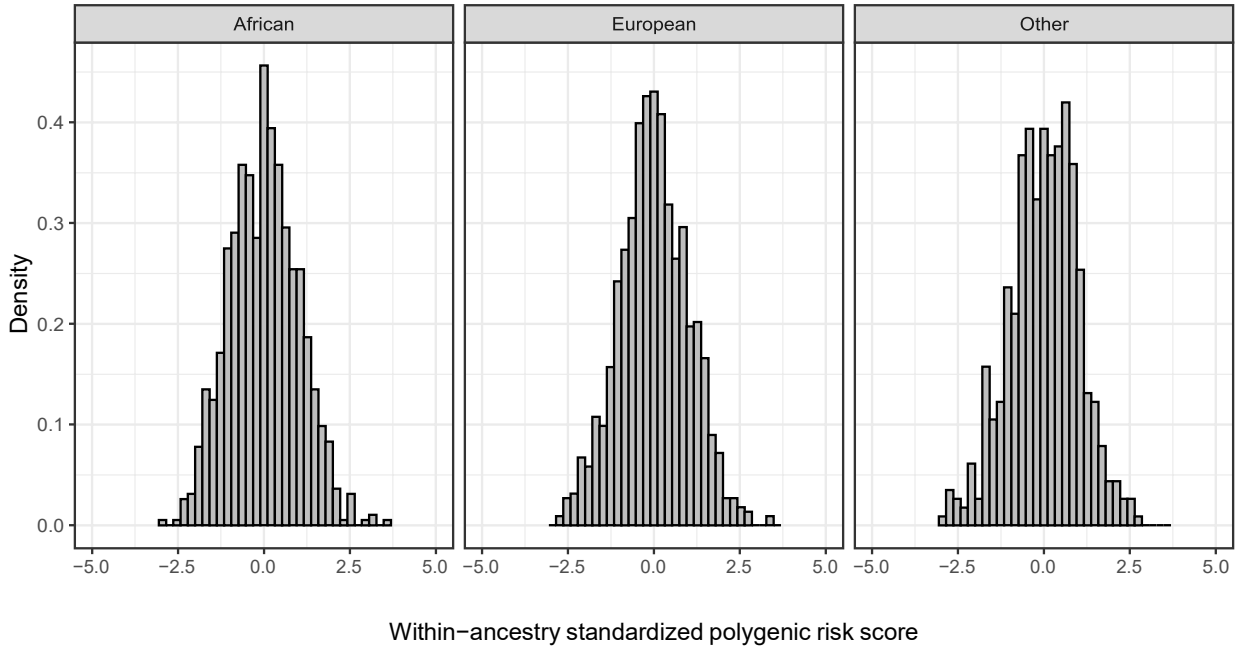
Okbay, A., Baselmans, B.M., De Neve, J.E., Turley, P., Nivard, M.G., Fontana, M.A., … & Gratten, J. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nature Genetics, 48(6), 624-633.

3.5 Extraversion – Genetics of Personality Consortium 2016

The extraversion PGSs were created using results from a 2016 study by the Genetics of Personality Consortium (GPC). The meta-analysis included 63,030 individuals of European ancestry from 29 cohorts in the discovery sample, and 9,783 individuals from one cohort in the replication sample. All datasets used a harmonized latent extraversion score as a continuous phenotype in study-specific GWAS. Summary statistics are freely available from the Netherlands Tweelingen register website (http://www.tweelingen-register.org/GPC/). Approximately 7.5 million SNPs were included in the meta-analyses, with all cohorts utilizing SNPs imputed to the 1000 genomes reference panel (1000G). No genome-wide significant SNPs were identified in the total sample; none of the 74 SNPs that were identified with a P-value < 1 x 10-5 were replicated in the replication sample. Study-specific GWASs accounted for sex and age; ancestry principal components were included as covariates at a study-specific determination.

The PGS contains 1,232,521 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the GPC 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.



References

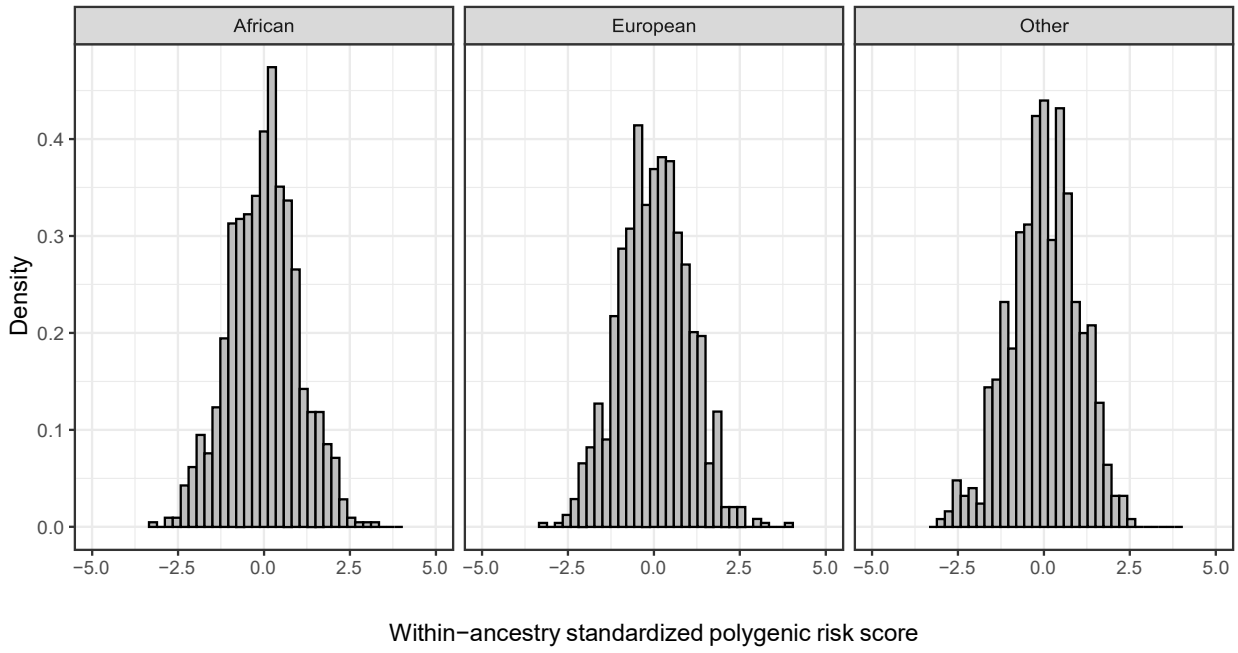
van den Berg, S. M., de Moor, M. H. M., Verweij, K. J. H., Krueger, R. F., Luciano, M., Arias Vasquez, A., & Boomsma, D. I. (2016). Meta-analysis of Genome-Wide Association Studies for Extraversion: Findings from the Genetics of Personality Consortium. Behavior Genetics, 46(2), 170-182. https://doi.org/10.1007/s10519-015-9735-5

3.6 Neuroticism – Social Science Genetic Association Consortium 2016

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 PGS contained 1,209,976 SNPs that overlapped between the CDS genetic database and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the SSGAC 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.



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.

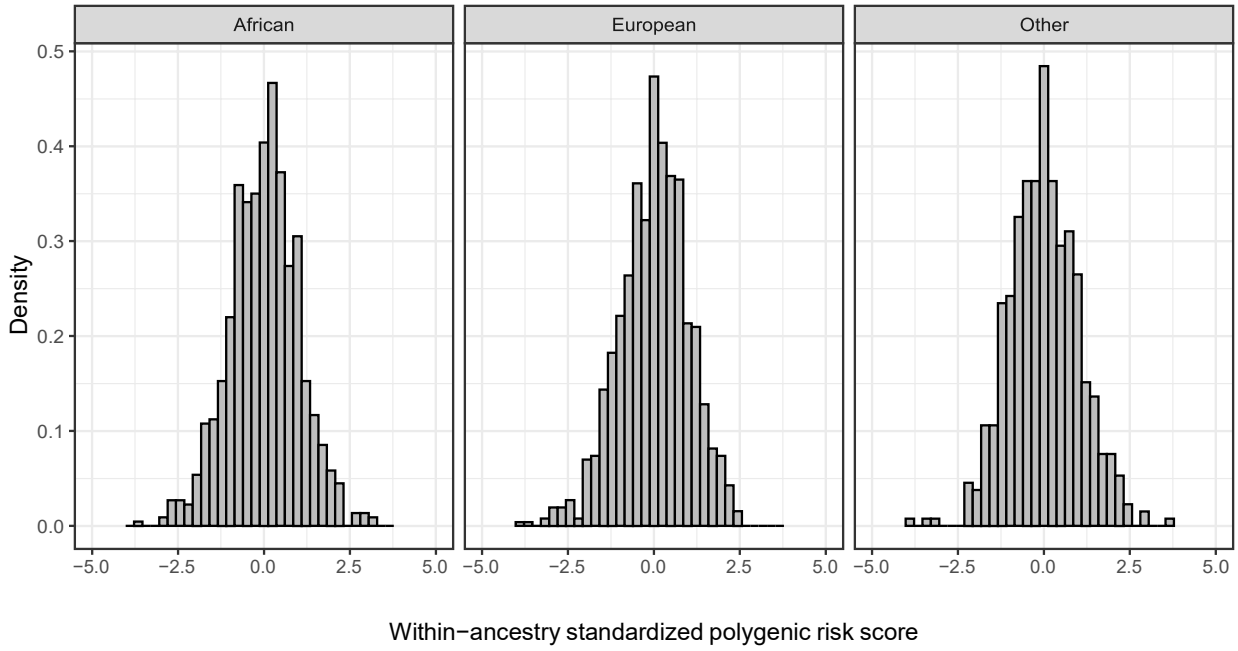
De Moor, M.H., Van Den Berg, S.M., Verweij, K.J., Krueger, R.F., Luciano, M., Vasquez, A.A., … & Gordon, S.D. (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. JAMA Psychiatry, 72(7), 642-650.

3.7 Anxiety – Anxiety NeuroGenetics Study 2016

The Anxiety PGSs were created using results from a 2016 study by the Anxiety NeuroGenetics Study (ANGST) Consortium (Otowa et al., 2016). Summary statistics are available from the PGC data download website (https://www.med.unc.edu/pgc/results-and-downloads). Data sourced 9 samples of European ancestry individuals, where each contributing cohort used standardized instruments to assess DSM-based anxiety disorder diagnoses (i.e., generalized anxiety disorder, panic disorder, social phobia, agoraphobia, and/or specific phobia). Parallel GWAS were conducted in each cohort (both in a case-control design and using a continuous factor score), followed by meta-analysis across all cohorts. The combined case-control meta-analysis included N=17,310 and the continuous factor score GWAS included N=18,186. All cohorts imputed SNPs to the 1000 Genomes Project references data (release v3, March 2012) and approximately 6.5 million SNPs were included in the combined meta-analysis. Sex and age at interview were included in the sample-specific GWAS and ancestry principal components were included on a sample-by-sample basis depending on their correlation with the outcome phenotypes' (Otowa et al., 2016). No genome-wide significant SNPs were identified in the GWAS meta-analyses.

The PGS contains 1,162,510 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis.

Please note that the anxiety 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.



References

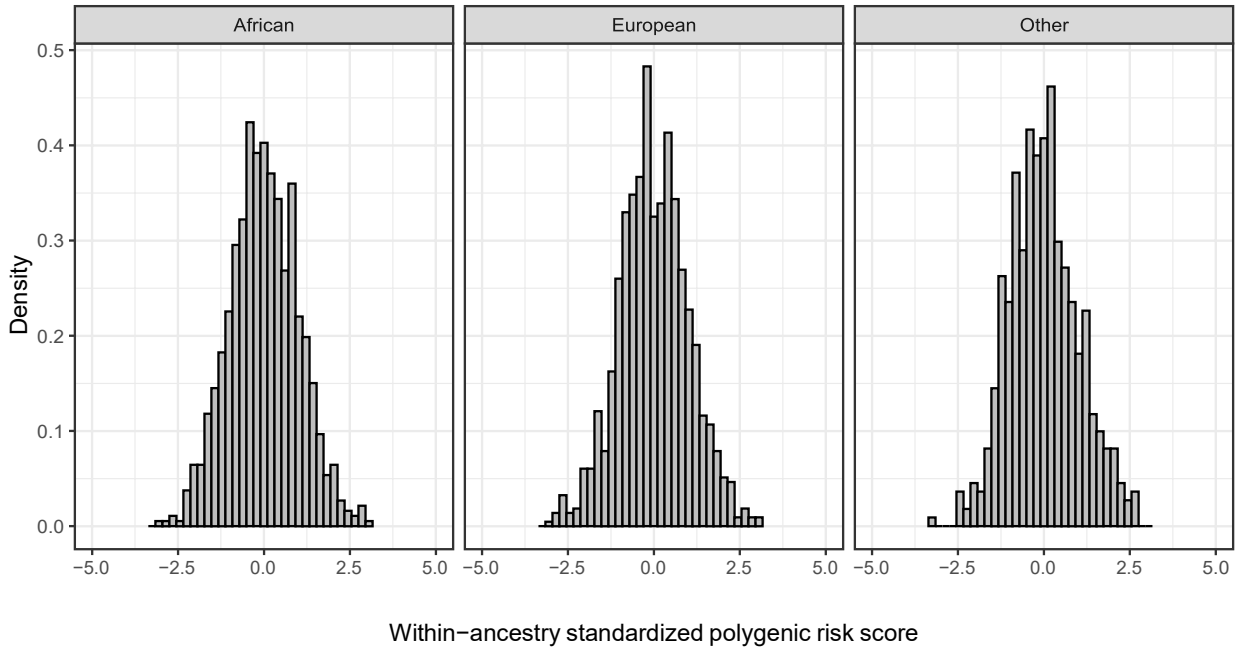
Otowa, T., Hek, K., Lee, M., Byrne, E.M., Mirza, S.S., Nivard, M. G., … & Fanous, A. (2016). Meta-analysis of genome-wide association studies of anxiety disorders. Molecular psychiatry, 21(10), 1391.

3.8 Antisocial Behavior – Broad Antisocial Behavior Consortium 2017

The antisocial behavior (AB) PGSs were created using results from a 2017 study by the Broad Antisocial Behavior Consortium. The meta-analysis included 16,400 individuals in the discovery sample and 9,381 individuals in the replication sample (both child and adult samples). All datasets used continuous phenotypes except for one study that used a case-control design (Tielbeek et al., 2017). Approximately 7.4 million SNPs were included in the meta-analyses, with all cohorts utilizing SNPs imputed to the 1000 genomes reference panel (1000G) (except two of the replication cohorts, which were not imputed). Sex-specific GWAS were also conducted. No genome-wide significant SNPs were identified in the total sample; three sex-discordant loci were identified just below the genome-wide significant p-value threshold, but were not replicated in the replication samples. Study-specific GWASs accounted for sex, age, the first four principal components, and study-specific covariates.

The PGS contains 386,853 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the antisocial behavior PGS 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.



References

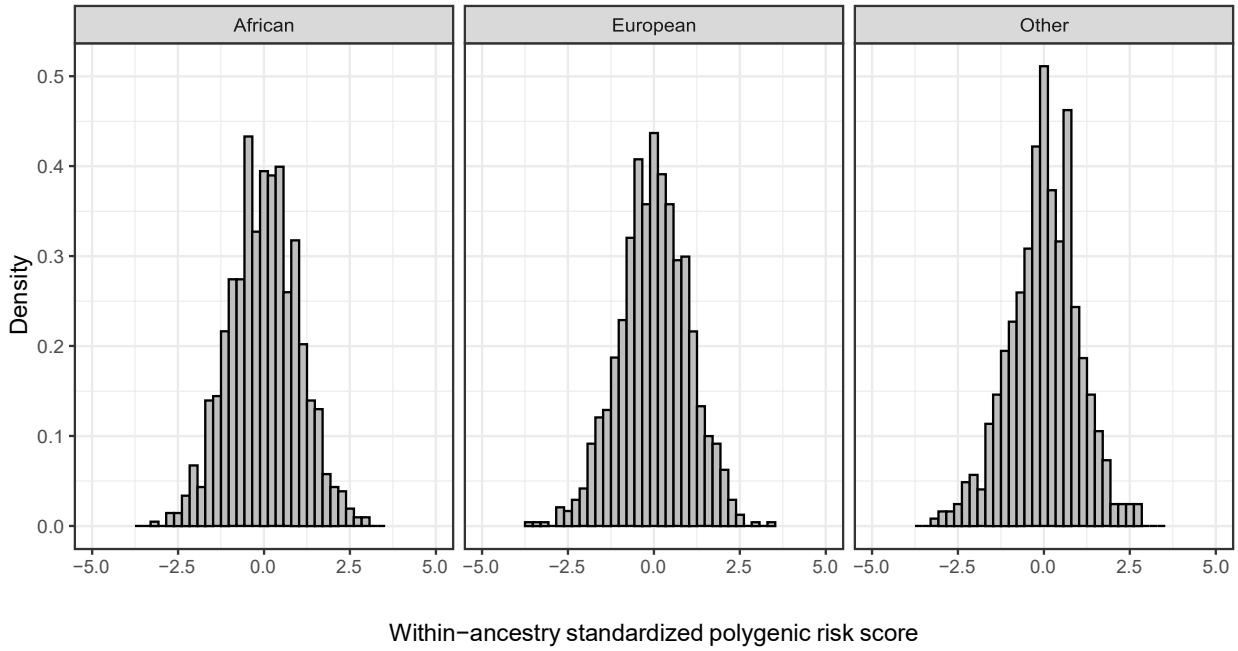
Tielbeek, J. J., Johansson, A., Polderman, T. J. C., Rautiainen, M.-R., Jansen, P., Taylor, M., & Posthuma, D. (2017). Genome-Wide Association Studies of a Broad Spectrum of Antisocial Behavior. JAMA Psychiatry, 74(12), 1242-1250. https://doi.org/10.1001/jamapsychiatry.2017.3069

3.9 Depressive Symptoms – Social Science Genetic Association Consortium 2016

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) (Ncases=9,240, Ncontrols=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 (Ncases=7,231, Ncontrols=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 PGS contains 1,209,987 SNPs that overlapped between the CDS genetic database and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the SSGAC 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.



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.

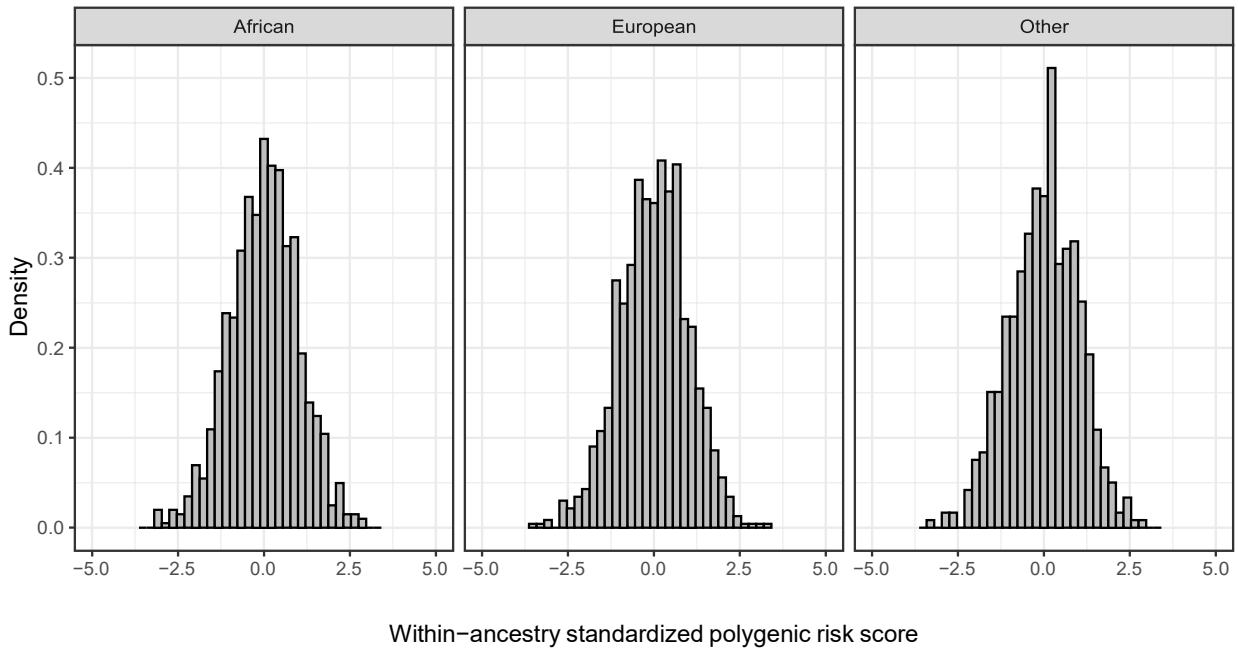
Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., … & Heath, A.C. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Molecular Psychiatry, 18(4), 497.

3.10 Post-Traumatic Stress Disorder – Psychiatric Genomics Consortium 2018

The post-traumatic stress disorder (PTSD) PGSs were created using results from a 2018 study by the Psychiatric Genomics Consortium (Duncan et al., 2018). Authors conducted separate GWAS for European ancestry datasets (Ncases=2,424, Ncontrols=7,113), African American datasets (Ncases=2,479, Ncontrols=6,744), Latino/Hispanic datasets (Ncases=98, Ncontrols=598), and South African datasets (Ncases=130, Ncontrols=254), followed by a meta-analyses across all four ancestry groups (Ncases=5,131, Ncontrols=15,092). Data sourced from 11 contributing studies; PTSD case status was measured using both self-report measures matched to DSM symptoms, clinician-administered questionnaires, and clinical interviews. Many of the controls were also exposed to trauma but did not meet PTSD criteria (Duncan et al., 2018). All cohorts imputed SNPs to the 1000 Genomes phase I reference panel. Approximately 21.2 million SNPs were included in the African American GWAS, 13.2 million SNPs in the European GWAS, and 25.5 million SNPs in the combined ancestry GWAS. No genome-wide significant SNPs were identified in either the transethnic or European meta-analyses. In the African American meta-analysis, one SNP exceeded the genome-wide significance threshold. All GWAS analyses controlled for the top 10 principal components.

The PGS contains 1,707,935 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the PGC-PTSD results contain PGSs from multiple ancestry backgrounds. See Section C, "Notes about the use of PGSs," for more information on the use of PGSs in other ancestry groups.



References

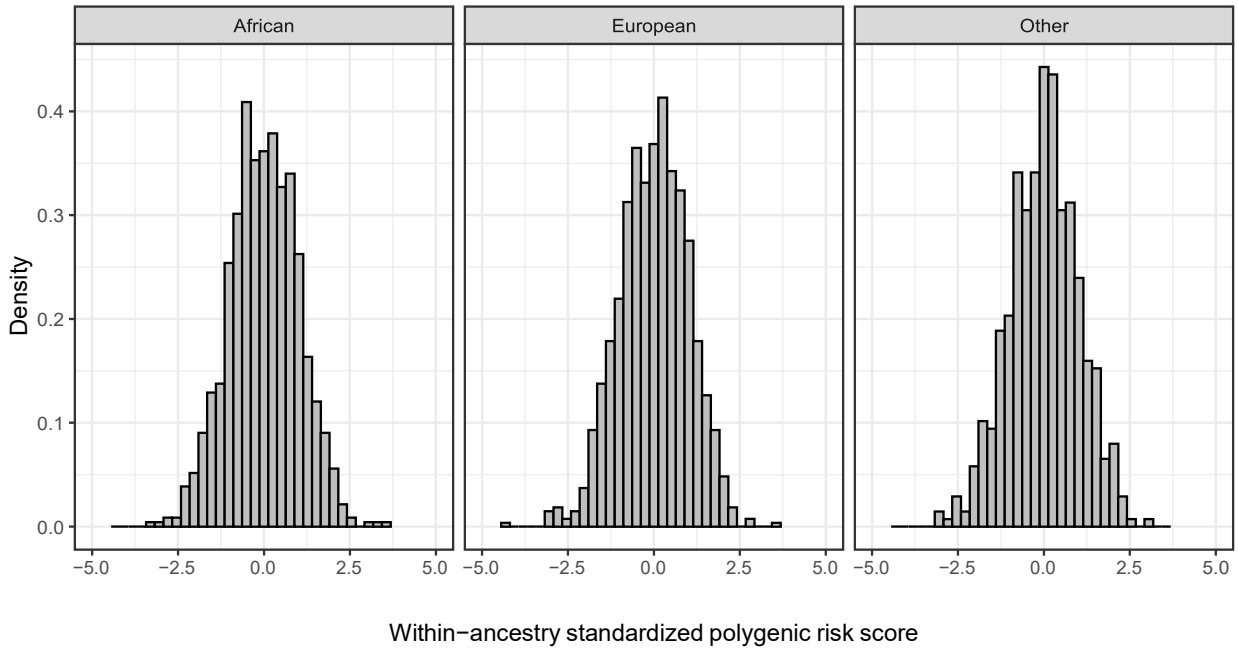
Duncan, L.E., Ratanatharathorn, A., Aiello, A.E., Almli, L.M., Amstadter, A.B., Ashley-Koch, A.E., … & Bradley, B. (2018). Largest GWAS of PTSD (N= 20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Molecular psychiatry, 23(3), 666.

3.11 Obsessive Compulsive Disorder – International Obsessive Compulsive Disorder Foundation-Genetics Collaborative 2017

The obsessive-compulsive disorder (OCD) PGSs were created using results from a 2017 study by International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS). The meta-analysis included 9,725 (2,688 OCD cases, 7037 controls)individuals. There was no replication sample for the GWAS analyses. No genome-wide significant SNPs were identified. Approximately 8.7 million SNPs were included in the analyses, with all cohorts using SNPs imputed to the 1000 genomes reference panel (1000G). Study-specific GWASs did not account for covariates; rather, separate association analyses were conducted for each case-control subpopulation (i.e., identified using principal components analyses).

The PGS contains 1,253,790 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity to a standard normal curve (mean=0, standard deviation=1).

Please note that the IOCDF-OCD 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 inother ancestry groups.



References

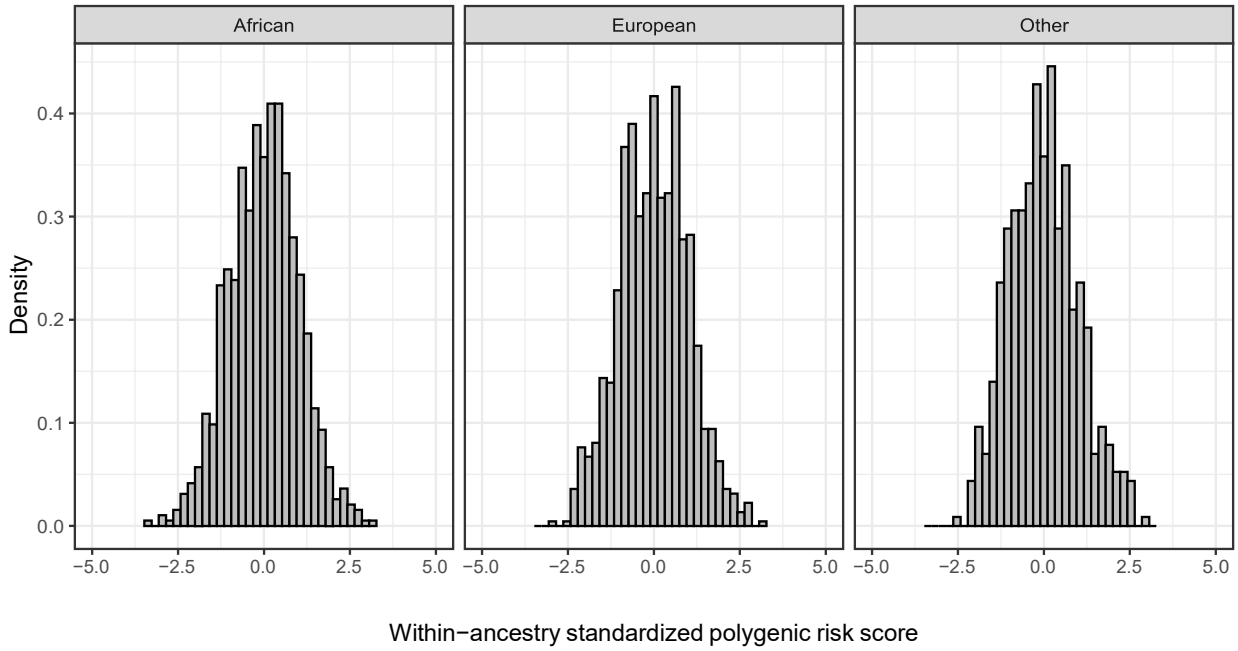
International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS). (2017). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. Molecular Psychiatry. https://doi.org/10.1038/mp.2017.154

3.12 Bipolar Disorder – Psychiatric Genomics Consortium 2011

The PGSs for bipolar disorder were created using results from a 2011 GWAS conducted by the Bipolar Disorder Working Group of the Psychiatric GWAS Consortium. The GWAS meta-analysis files are publicly available on the PGC website: http://www.med.unc.edu/pgc/results-and-downloads (pgc.bip.2012-04.zip). The discovery phase of the meta-analysis included 7,481 cases and 9,250 controls. Samples were drawn from 11 studies (see Skylar et al., Table 1 and supplemental materials, 2011). A follow up meta-analysis of the 34 most significant regions was conducted in a replication sample that included 4,496 independent cases and 42,422 independent controls. The combined GWAS meta-analysis yielded two genome-wide significance SNPs. BP case status was measured in all studies using standardized semi-structured interviews. Controls had a low probability of BP; some control selection criteria excluded individuals with a history of a mood disorder and other controls were unscreened. Meta-analyses were adjusted for the top five principal components and 10 dummy variables to account for differences between the 11 studies.

The PGS contains 1,146,972 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the PGC 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.



References

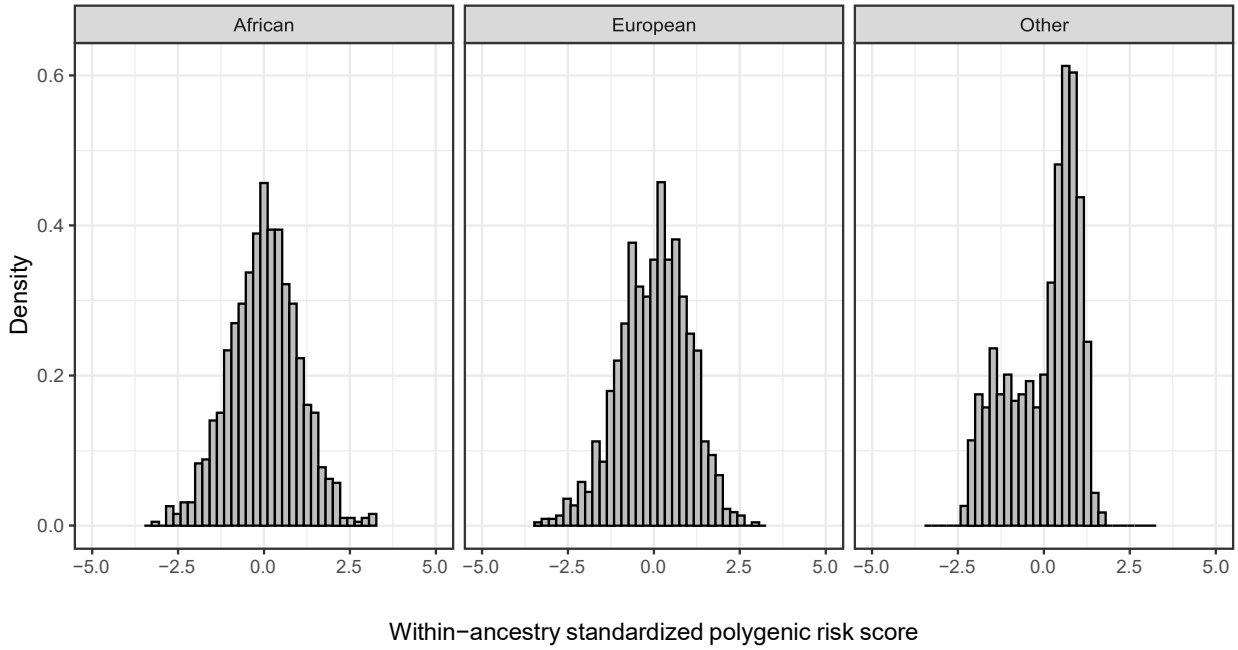
Psychiatric GWAS Consortium Bipolar Disorder Working Group. (2011). Large-scale genome-wide associationanalysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature Genetics, 43(10), 977-983.

3.13 Mental Health Cross Disorder – Psychiatric Genomics Consortium 2013

The PGSs for Mental Health Cross-Disorder were created using results from a 2013 GWAS conducted by the Cross Disorder working group of the Psychiatric Genomics Consortium. The GWAS meta-analysis files are pub- licly available on the PGC website: http://www.med.unc.edu/pgc/results-and-downloads (pgc.cross.full.2013- 03.zip). The discovery phase of the meta-analysis included 33,342 cases and 27,888 controls. Disorders that were counted as cases (DSM-III-R or DSM-IV criteria) included autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia. Four SNPs surpassed the genome-wide significance threshold. Analyses were adjusted for the top seven genetic principal components.

The PGS contains 597,573 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the PGC 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.



References

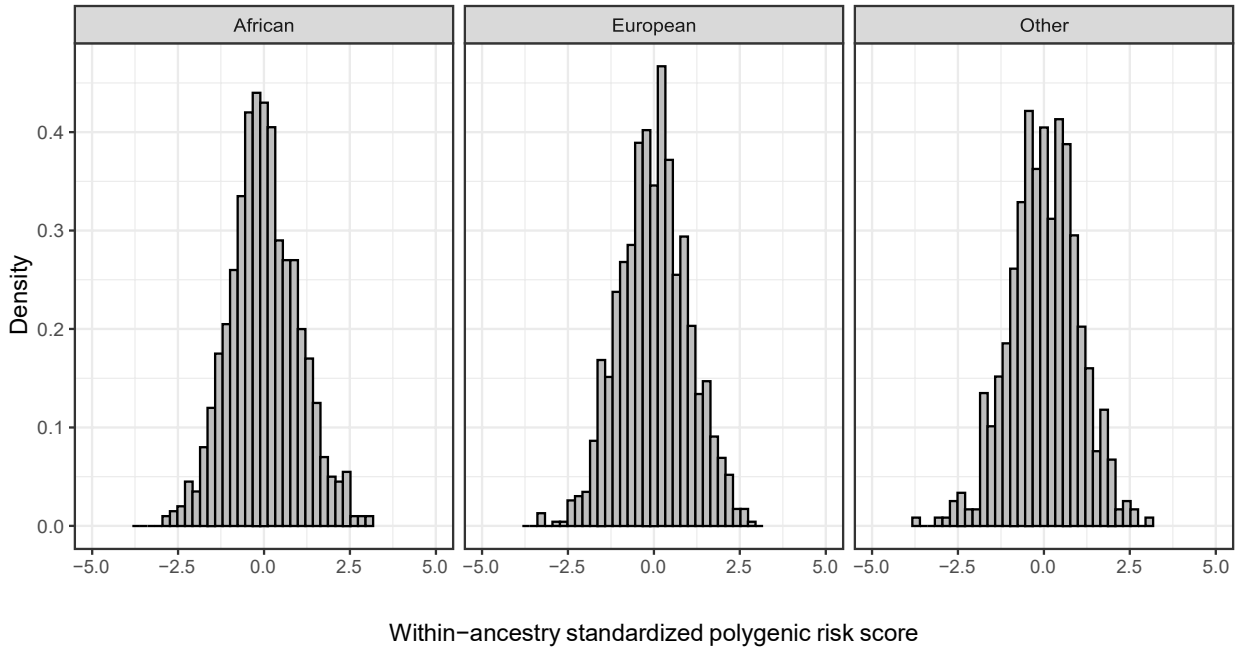
Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. The Lancet, 381(9875), 1371-1379.

3.14 Autism Spectrum Disorders – Psychiatric Genomics Consortium 2017

The PGSs for autism were created using results from a 2017 GWAS conducted by the Autism Spectrum Disorders Working Group of the Psychiatric Genomics Consortium. The GWAS meta-analysis files are publicly available on the PGC website: http://www.med.unc.edu/pgc/results-and-downloads. The phase I discovery sample included 7,387 Autism Spectrum Disorder (ASD) cases, and 8,567 controls. Samples were drawn from 14 independent cohorts. Covariates in the individual GWAS included genetic principal components. No SNPs exceeded genome-wide significance in the phase I discovery stage. Two independent samples were used for replication: the Danish iPSYCH Project (7,783 ASD cases, 11,359 controls) and a combined deCODE Collection and the Study to Explore Early Development (SEED) (1,369 ASD cases, 137,308 controls). Authors examined 180 LD-independent markers from the phase I discovery GWAS meta-analysis that were associated with ASD at $p < 5 \times 10\text{-}4$ in both replication samples: 6.1% and 5% of the markers were associated with ASD after multiple comparison correction in the iPSYCH and deCODE/SEED samples, respectively.

The PGS contains 1,132,964 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

Please note that the PGC 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.



References

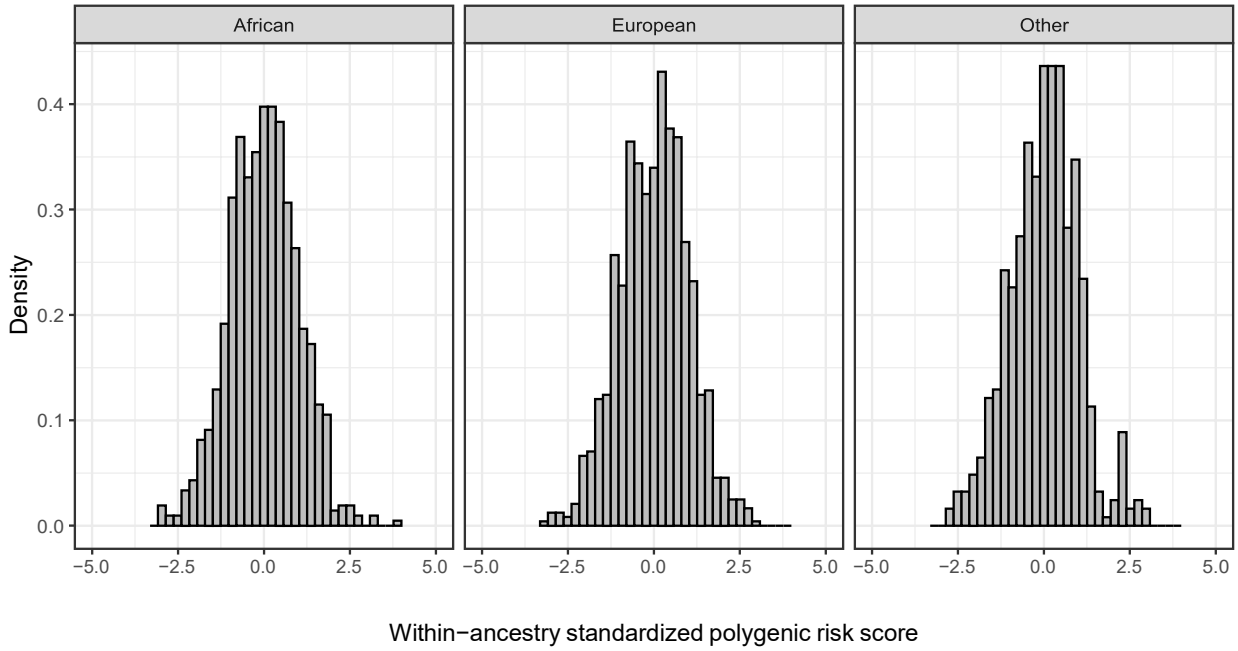
Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, Anney, R. J., Ripke, S., Anttila, V., Grove, J., Holmans, P., … & Neale, B. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. Molecular Autism, 8, 1-17.

3.15 Lifetime Cannabis Use – International Cannabis Consortium + UKBiobank 2019

PGSs for lifetime cannabis use (CANNA) were created using results from a 2018 study conducted by the International Cannabis Consortium. The GWAS meta-analysis files are publicly available on the ICC data download page: https://www.ru.nl/bsi/research/group-pages/substance-use-addiction-food-saf/vm-saf/genetics/international-cannabis-consortium-icc/. The meta-analysis included 184,765 participants of European ancestry, from the ICC, 23andMe, and UK Biobank cohorts. To compute the PGSs for CDS respondents, SNP weights were constructed from publicly available data and do not include results with 23andMe due to data use agreements. Thus, the PGSs for CDS respondents are based on a GWAS of 162,082 participants of European Ancestry from the ICC cohorts and the UK Biobank (see cannabis readme on ICC documentation page; website link above). For the CANNABIS phenotype, self-report data were available on whether the participant had ever used cannabis during their lifetime: yes (1) versus no (0); slightly different wording was used in each cohort (see Pasman et al., 2018). The GWAS in the ICC cohorts was based on 6,643,927 SNPs in 35,297 participants; covariates included age, sex, principal components to account for population stratification, birth cohort, and batch effects. The GWAS in the UKB was based on 10,827,718 SNPs in 126,785 participants; covariates included age, age2, sex, genotype array, and ancestry principal components to account for population stratification. This GWAS meta-analysis of the ICC cohorts and UKB (N=162,082) included 11,733,371 million SNPs imputed to Haplotype Reference Consortium imputation reference panel. The study identified 8 genome-wide significant SNPs ($P < 10\text{-}8$) (Figure 1 and Table 1).

The PGS contains 1,480,921 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity to a standard normal curve (mean=0, standard deviation=1).

Please note that the ICCUKB weights 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.



References

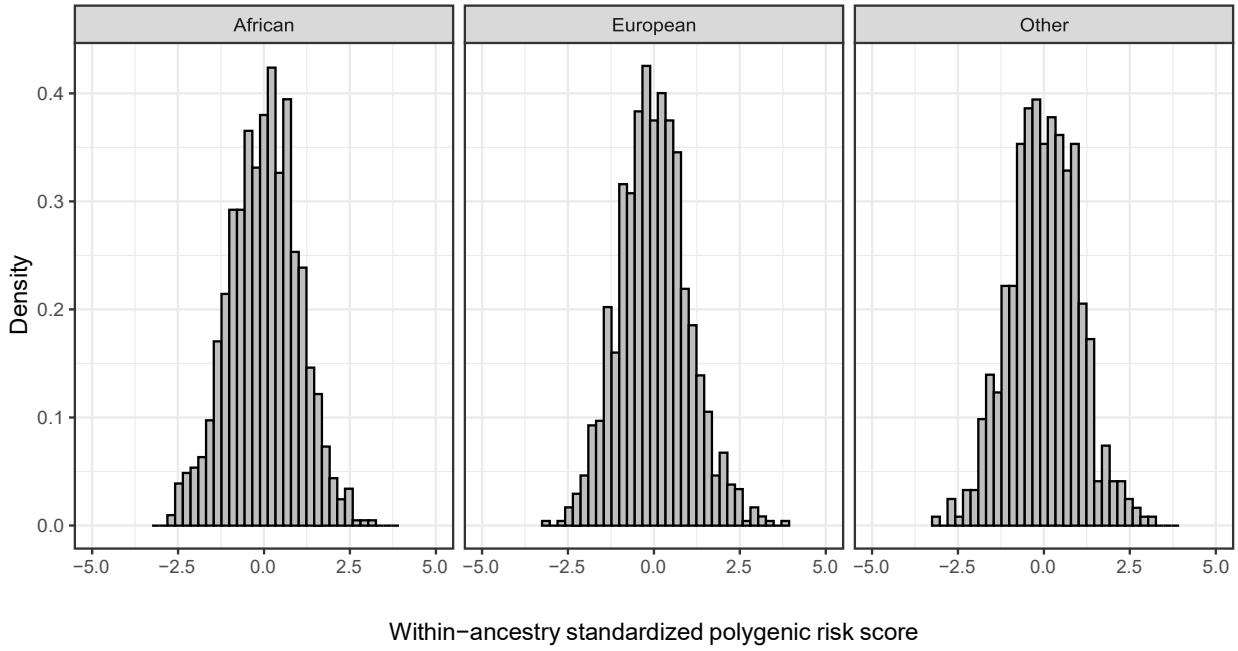
Pasman JA, Verweij KJH, Gerring Z, et al. GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia [published correction appears in Nat Neurosci. 2019 Jul;22(7):1196]. Nat Neurosci. 2018;21(9):1161-1170. doi:10.1038/s41593-018-0206-1

3.16 Alcohol Dependence – Psychiatric Genomics Consortium 2018

The alcohol dependence PGSs were created using results from a 2018 study by the Psychiatric Genomics Consortium Substance Use Disorder workgroup. Results from several GWAS were reported. The unrelated European ancestry GWAS meta-analysis included 11,569 cases and 34,999 controls from 27 cohorts; three SNPs reached genome-wide significance. Approximately ~9 million SNPs were included in the analyses, with all cohorts utilizing SNPs imputed to the 1000 genomes reference panel (1000G). Three replication cohorts were used: FINRISK (1,232 cases, 22,614 controls); Yale-Penn 2 (911 cases, 599 controls); and COGA African-American Family GWAS (880 cases, 1,814 controls). Study-specific GWASs controlled for sex and between five and ten principal components of the genotypic data in the European cohorts.

The PGS contains 1,522,593 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity to a standard normal curve (mean=0, standard deviation=1).

Please note that the PGC Alcohol Dependence weights 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.



References

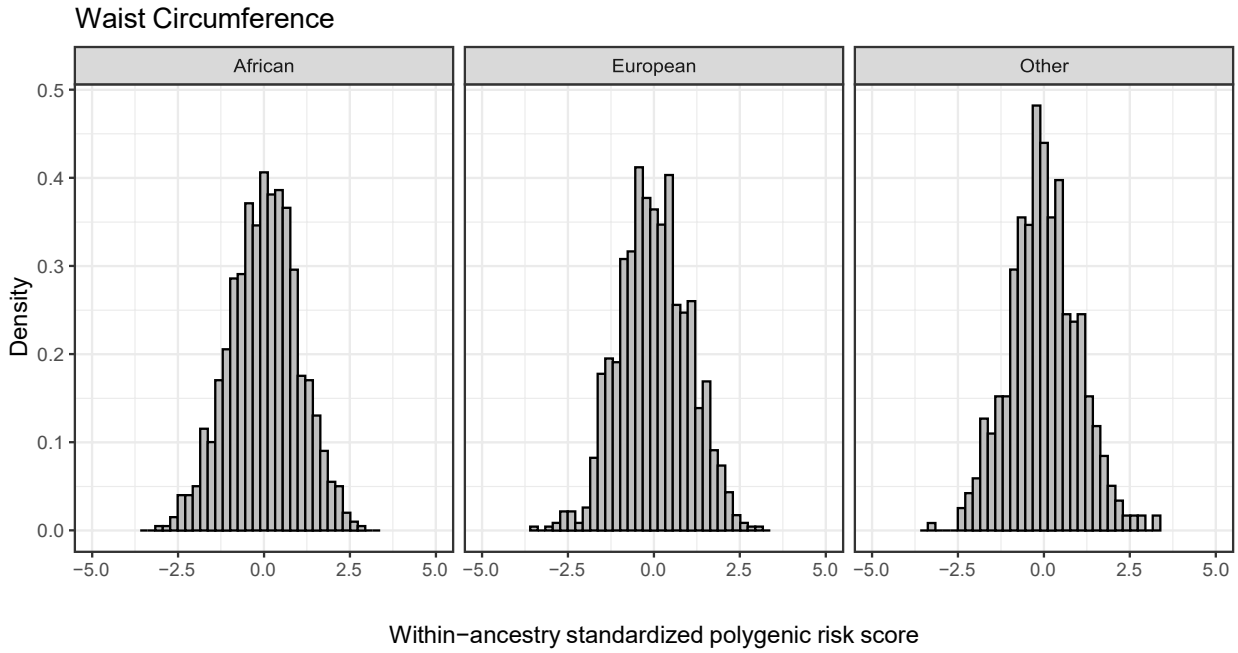
Walters RK, Polimanti R, Johnson EC, et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat Neurosci. 2018;21(12):1656-1669. doi:10.1038/s41593-018-0275-1

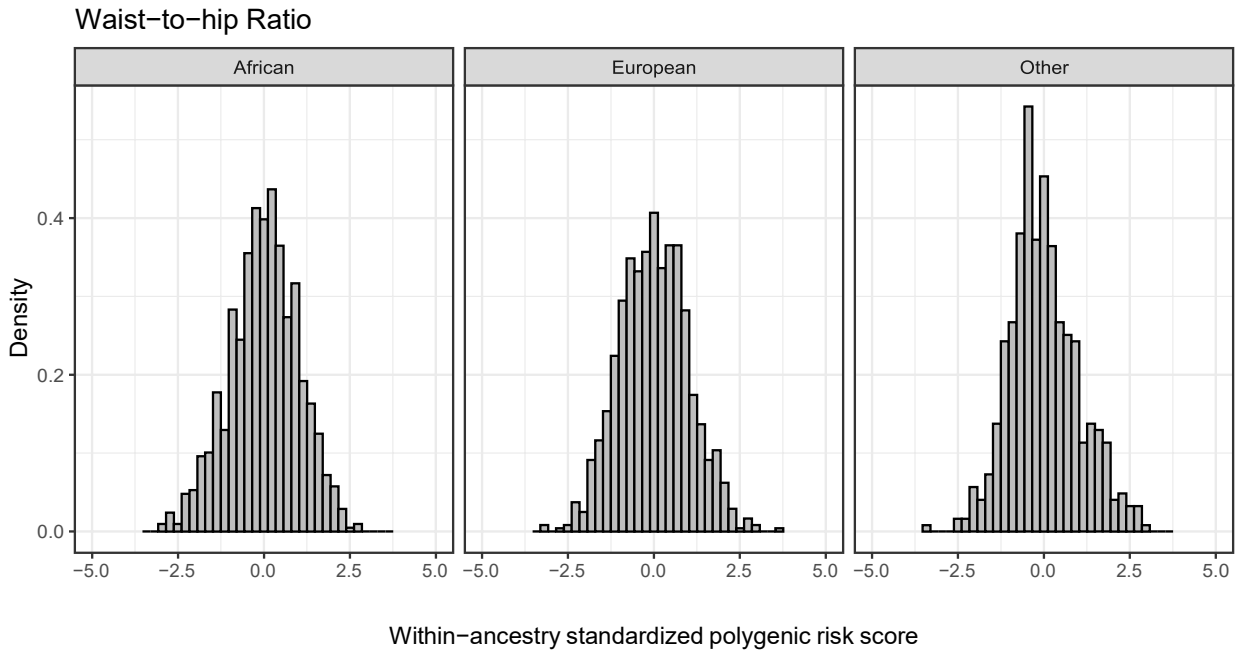
3.17 Waist Circumference and Waist-to-Hip Ratio

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 (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, age2, study-specific covariates if necessary, and BMI.

The GIANT WC PGSs contains 1,197,556 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The GIANT WHR PGSs contains 1,186,252 that overlapped between the CDS 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.





References

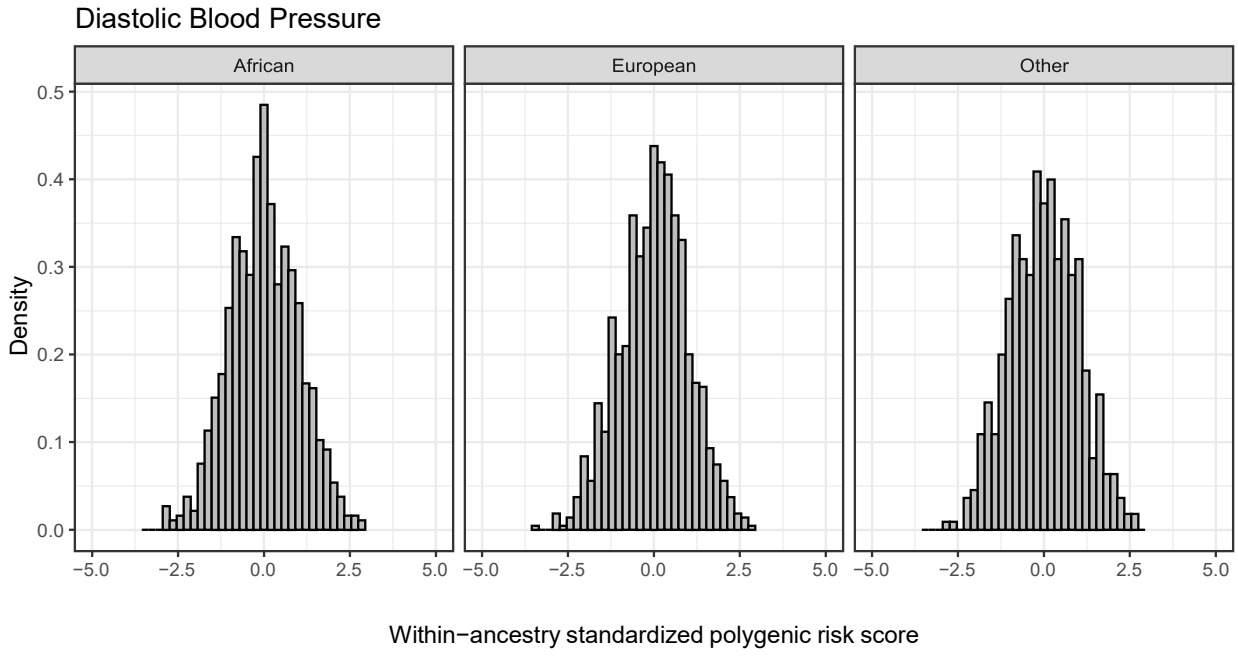
Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Mägi, R., … & Workalemahu, T. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature, 518(7538), 187.

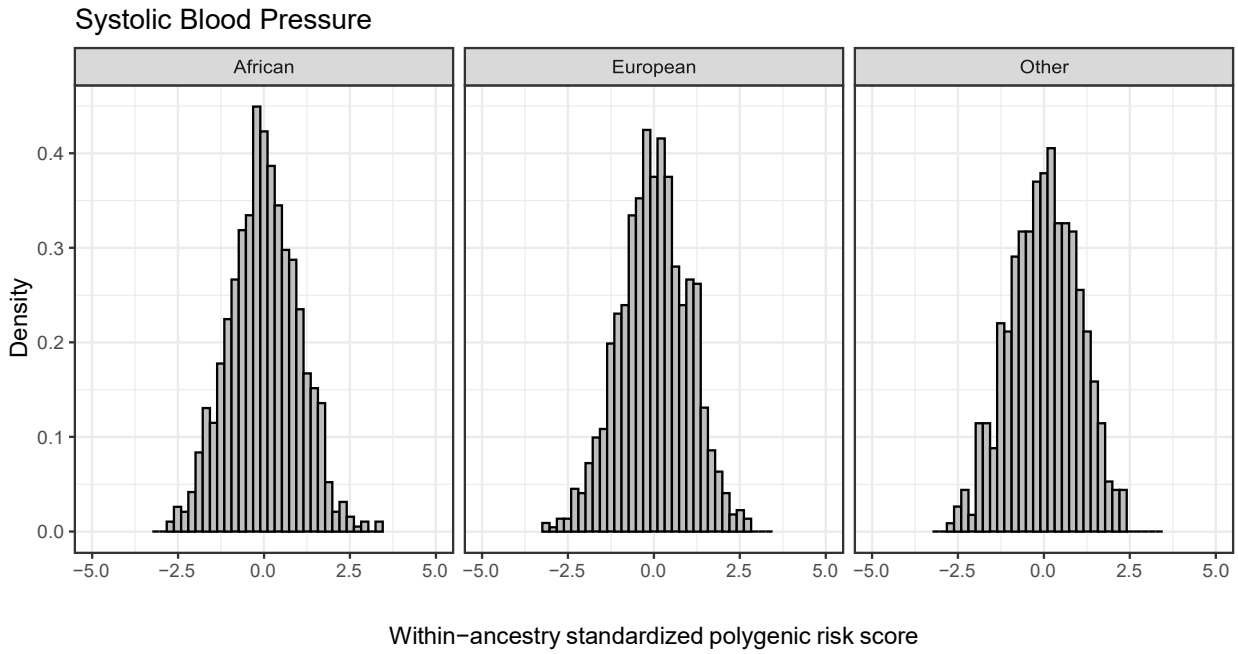
3.18 Blood Pressure – International Consortium of Blood Pressure-Genome Wide Association Studies (ICBP) + UKBiobank 2018

The diastolic blood pressure (DBP) and systolic blood pressure (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 UBKICBP DBP PGS contains 108,523 that overlapped between the CDS genetic data and the GWAS meta-analysis. The UBKICBP SBP PGS contains 108,470 that overlapped between the CDS 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 blood pressure results are based on 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.





References

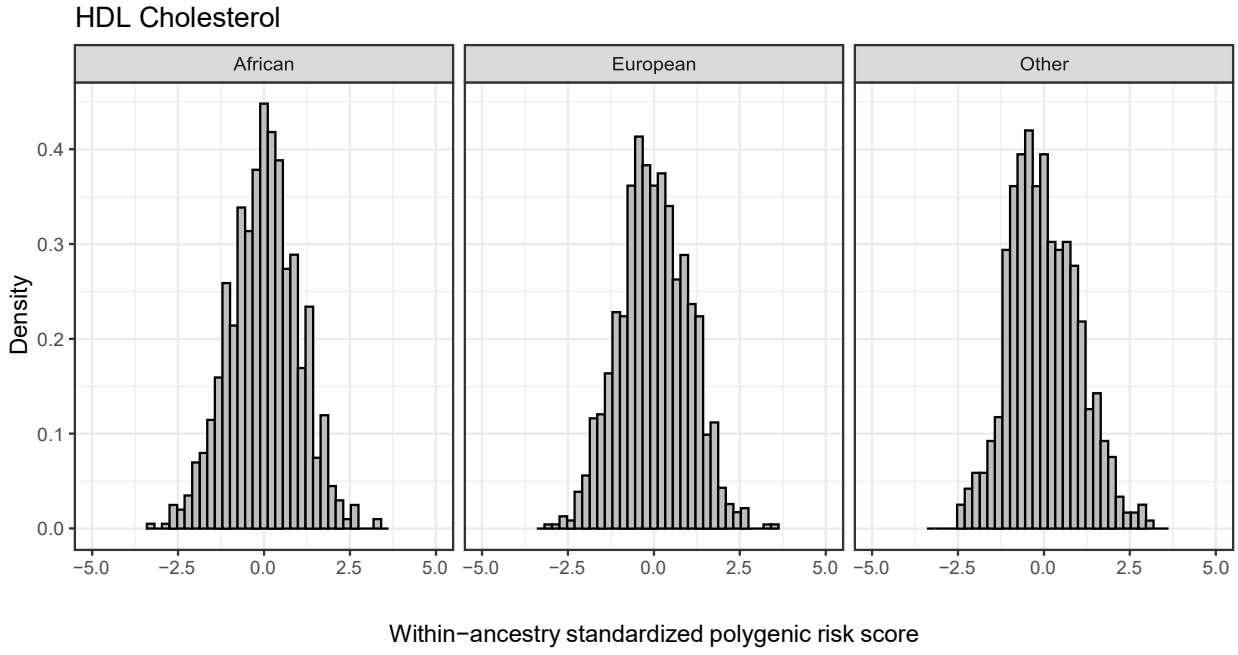
Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, &, Million Veteran Program. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet. 2018 Oct;50(10):1412-1425. doi: 10.1038/s41588-018-0205-x. Epub 2018 Sep 17. Erratum in: Nat Genet.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

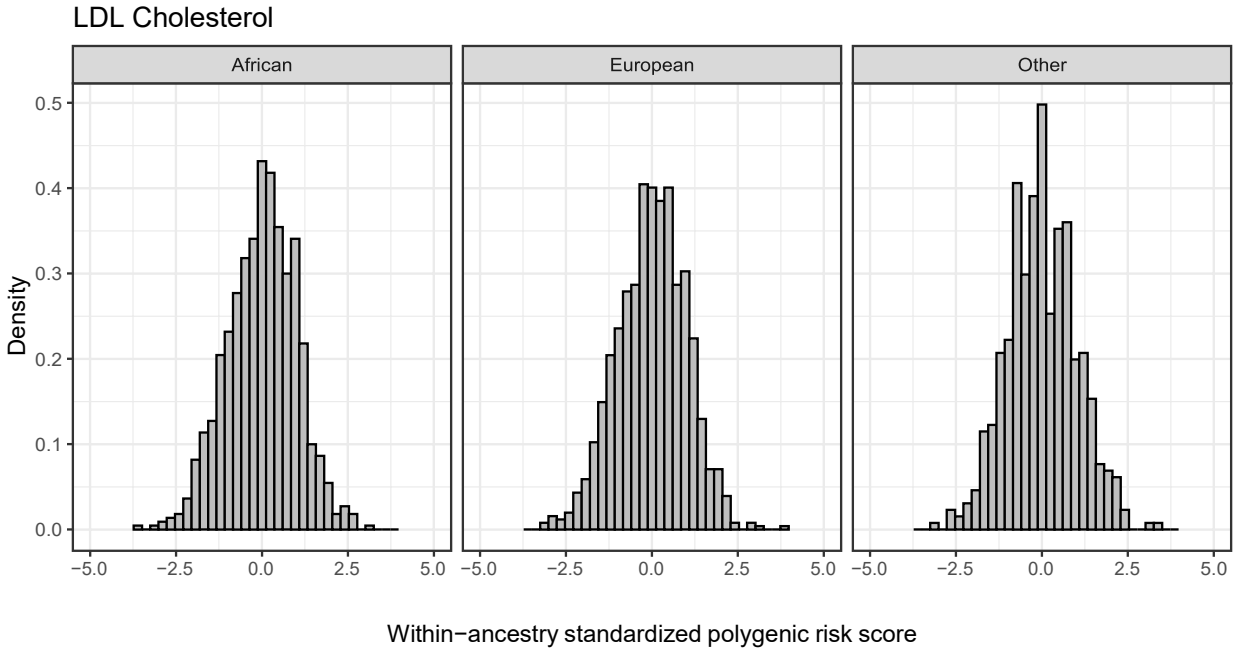
3.19 Lipid Traits (HDL, LDL, Total Cholesterol, Triglycerides) – Global Lipid Genetics Consortium 2013

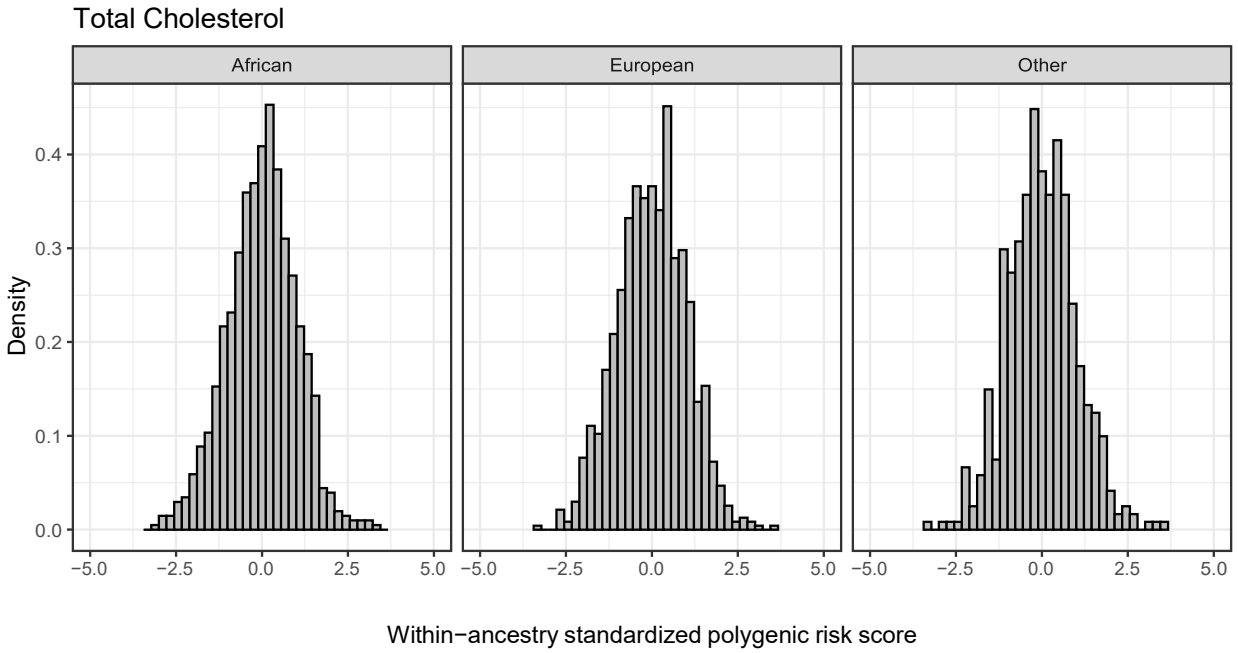
The HDL, LDL, and TC PGS were created using results from a 2013 study by the Global Lipid Genetics Consortium (Willer et al. 2013). Authors conducted separate GWAS for European (n=188,578) and non-European (n=7,898) ancestries followed by a meta-analysis of 7,168 individuals in a single ancestry group. 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. 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).

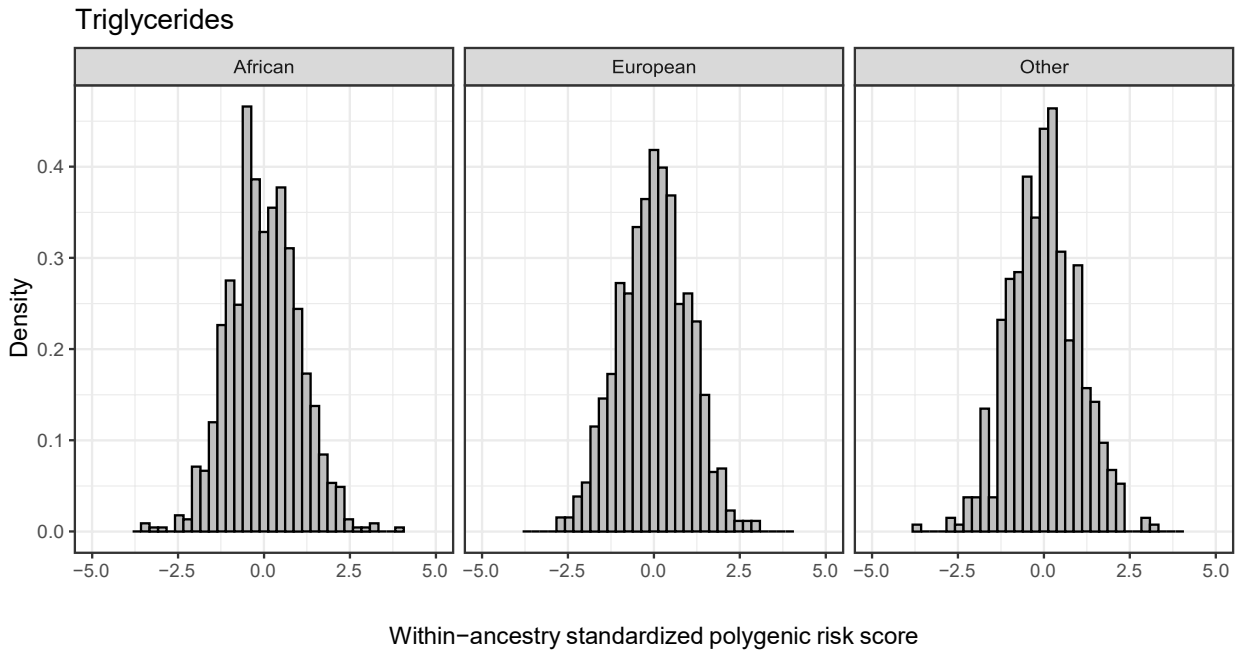
The GLGC HDL PGS contains 1,138,441 that overlapped between the CDS genetic data and the GWAS meta-analysis. The GLGC LDL PGS contains 1,135,295 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The GLGC TC PGS contains 1,138,230 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The GLGC TG PGS contains 1,133,346 SNPs that overlapped between the CDS 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.




LDL Cholesterol
African
European
Other
Density
0.0
0.1
0.2
0.3
0.4
0.5
−5.0
−2.5
0.0
2.5
5.0
Within−ancestry standardized polygenic risk score


Total Cholesterol
African
European
Other
Density
0.0
0.1
0.2
0.3
0.4
−5.0
−2.5
0.0
2.5
5.0
Within−ancestry standardized polygenic risk score



References

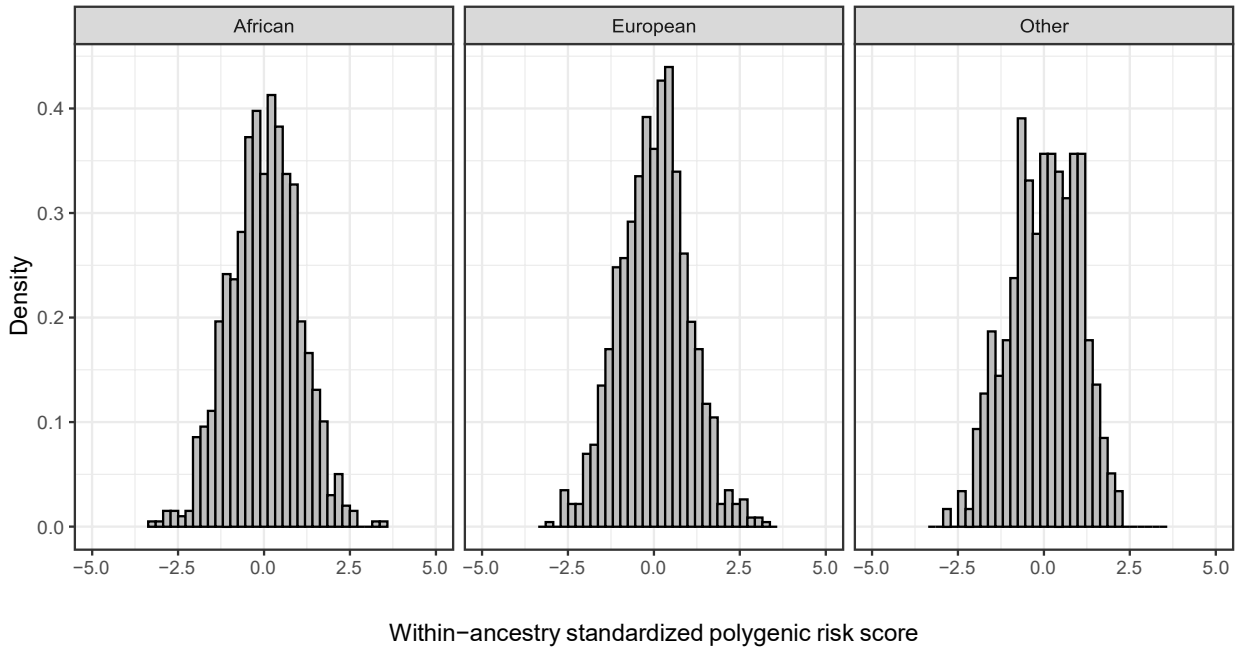
Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., & & Global Lipids Genetic Consortium. (2013) Discovery and Refinement of Loci Associated with Lipid Levels. Nat Genet. 45(11), 1274-1283. doi:10.1038/ng.2797.

3.20 Type II Diabetes – Diabetes Genetics Replication and Meta-analysis 2012

The PGSs for Type II Diabetes (T2D) were created using GWAS meta-analysis results from a 2012 study conducted by the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium. The GWAS meta-analysis files can be downloaded from the DIAGRAM Consortium website: http://www.diagram-consortium.org/downloads.html (DIAGRAMv3.2012DEC17.txt). The stage one (discovery) meta-analysis consists of 12,171 T2D cases and 56,862 controls across 12 GWAS from European descent populations. The stage two (replication) meta-analysis consisted of 22,669 cases and 58,119 controls, including 1,178 cases and 2,472 controls of Pakistani descent. The combined meta-analysis identified ten genome-wide significant loci (Table 1). HapMap-2 CEU was used as the imputation panel resulting in a common set of ~2.5 million SNPs across studies. Study-specific GWAS adjusted for age of onset (cases) or age of recruitment (controls), gender, and genetic principal components. The results of each GWAS were corrected for residual population structure using the genomic control inflation factor, as reported in Supplementary Table 1 of Morris et al. (2012).

The PGS contains 1,340,856 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity, to a standard normal curve (mean=0, standard deviation=1).

The effect estimates for SNPs come from the discovery stage I meta-analysis of European descent individuals. Note that the DIAGRAM 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.



References

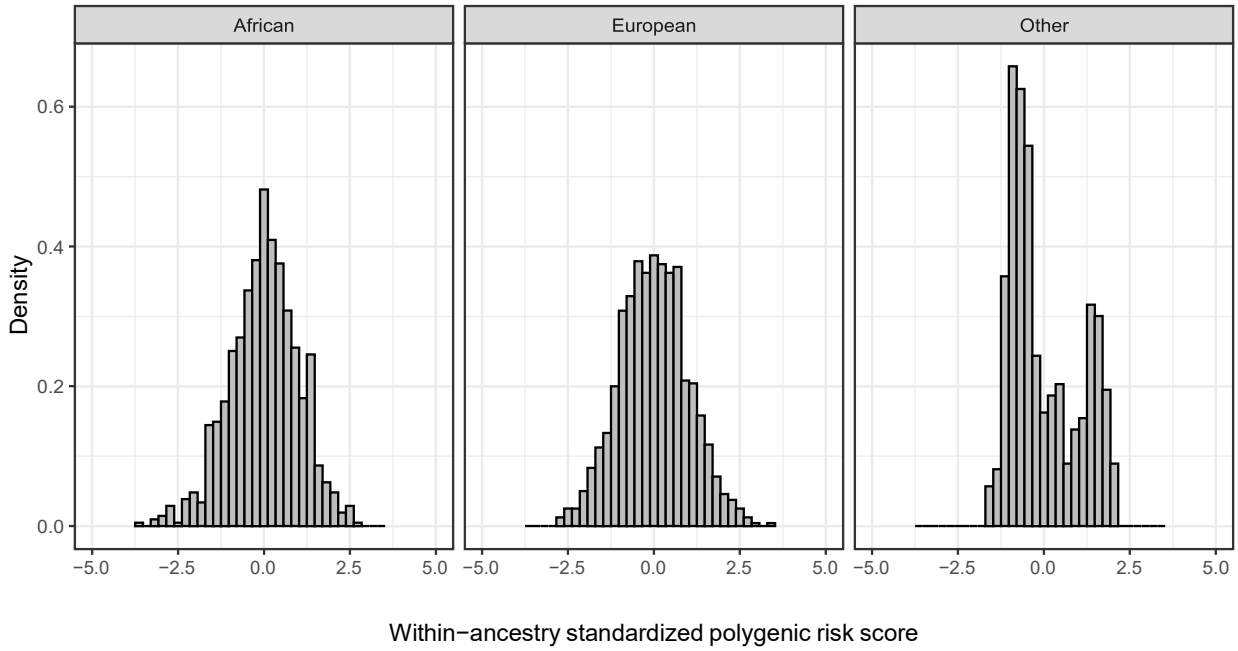
Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., … & Prokopenko, I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics, 44(9), 981-990.

3.21 Kidney Function – Chronic Kidney Disease Genetics Consortium 2019

PGSs for kidney function phenotypes were created using results from a 2019 study conducted by the Chronic Kidney Disease Genetics (CKDGen) consortium. The GWAS meta-analysis files are publicly available on the CKDGen data download page: http://ckdgen.imbi.uni-freiburg.de/ (20171016_MW_eGFR_overall_ALL_nstud61.dbgap.txt.gz, 20171017_MW_eGFR_overall_EA_nstud42.dbgap.txt.gz, BUN_overall_ALL_YL_20171017_METAL1_nstud_33.dbgap.txt.gz, BUN_overall_EA_YL_20171108_METAL1_nstud24 CKD_overall_ALL_JW_20180223_nstud30.dbgap.txt.gz, CKD_overall_EA_JW_20180223_nstud23.dbgap.txt.gz).
The CKDGen meta-analysis included GWAS on estimated glomerular filtration rate (eGFR), blood urea nitrogen (BUN) and chronic kidney disease (CKD) using a European ancestry only sample and also a trans-ethnic sample encompassing individuals of European, East Asian, African, South Asian, and Hispanic ancestry. The trans-ethnic GWAS of eGFR included 121 studies with an n of 765,348 and found 308 loci associated with eGFR. The European ancestry eGFR GWAS included 85 studies and an n of 567,460 with 256 discovered loci. The trans-ethnic BUN discovery analysis included an n of 416,178. The European ancestry BUN GWAS included an n of 243,029. The CKD trans-ethnic analysis included 625,219 individuals. The European ancestry CKD analysis included 480,698 individuals (41,395 cases and 439,303 controls). The GWAS meta-analyses included ~9 million imputed SNPs on NCBI Build 37/UCSC hg 19.

The PGS contains 1,367,353 SNPs that overlapped between the CDS genetic data and the GWAS meta-analysis. The posted PGSs have been standardized within ethnicity to a standard normal curve (mean=0, standard deviation=1).

Please note that the European ancestry based 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.



References

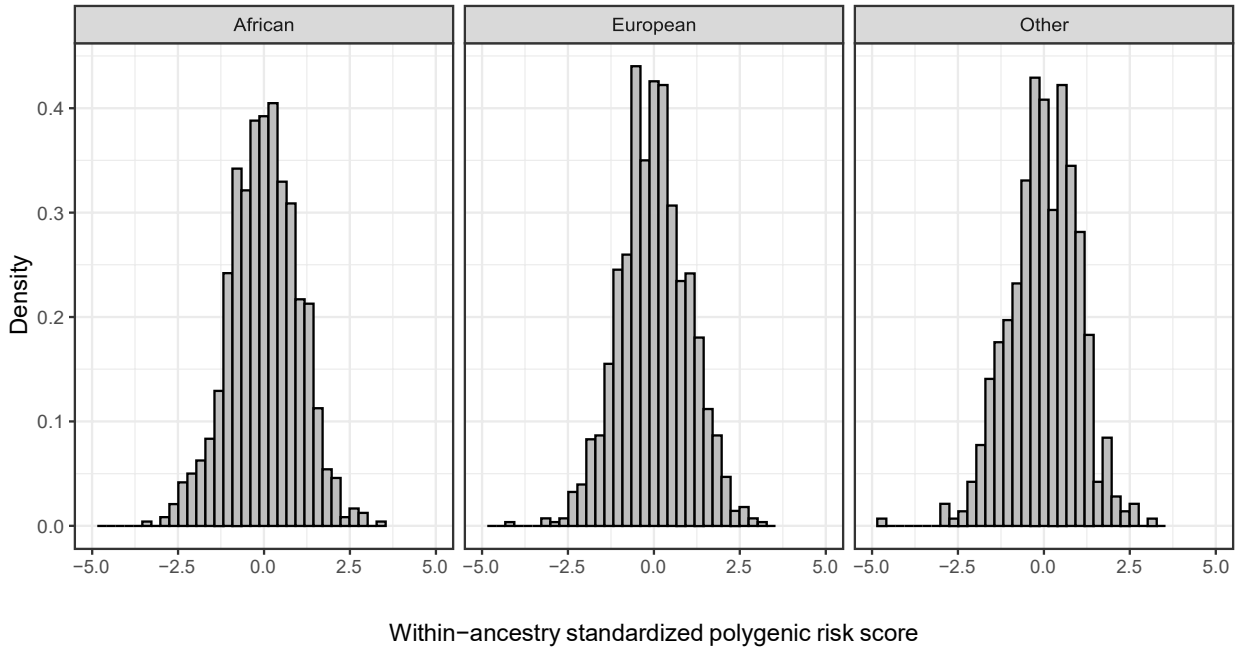
Wuttke M, Li Y, Li M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat Genet. 2019;51(6):957-972. doi:10.1038/s41588-019-0407-x

3.22 Myocardial Infarction

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 1,323,242 SNPs that overlapped between the CDS 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.



References

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