

Genetics

DOI: [10.15154/z563-zd24](#) (Release 5.1)

PUBLISHED
July 30, 2026

List of Instruments

Name of Instrument	Subdomain	Table Name
Tabulated Data		
Genetic Principal Components & Relatedness	Relatedness	gen_y_pihat
Twin Zygosity Rating	Relatedness	gen_y_zygrat

General Information

An overview of the ABCD Study® can be found at [abcdstudy.org](#) and detailed descriptions of the assessment protocols are available at [ABCD Protocols](#). This page describes the contents of various instruments available for download. To understand the context of this information, refer to the release note [Start Page](#).

This document contains information on zygosity, genetic and genetic derived data that is available for the ABCD sample for the 5.0 data release. In these release notes we describe both tabulated data that is available in release tables listed above, in addition to bulk genetic data that contains the following:

1. Curated genotyping data from Smokescreen array - set of PLINK files containing 11,666 unique subjects at ~500k variants
2. Imputed data based on TOPMED reference panel - set of PLINK files containing all imputed genotype data for 11,666 unique subjects at ~300 million variants
3. Genetic relatedness and $\hat{\pi}$ estimates across the full sample using methods correcting ancestry background
4. Genetic principal component weights to enable projection of other samples to ABCD genetic PC space

For compressive details of quality control steps performed and a description of the genetic data within the ABCD sample please refer to and cite the following work:

Chun Chieh Fan, Robert Loughnan, Sylia Wilson, John K. Hewitt & ABCD Genetic Working Group Behavior Genetics (2023). Genotype Data and Derived Genetic Instruments of Adolescent Brain

Cognitive Development Study® for Better Understanding of Human Brain Development. *Behavior Genetics*, 53, 159-168. [Find here](#)

Consideration of Population Descriptors

The use of population descriptors in genetic research can often be varied and inconsistent. We encourage users to review the following report for consideration of appropriate population descriptors for their analysis:

Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research; Board on Health Sciences Policy; Committee on Population; Health and Medicine Division; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine (2023). Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field. *Washington (DC): National Academies Press (US)* PMID: 36989389. [Find here](#)

Updates and Notes

In September, 2022, the ABCD DAIRC received a new batch of genotype data and updated the release accordingly. Here are some changes in the current data release:

1. Additional 567 samples were genotyped and passed QC
2. Subject ID issues from release 3.0 were fixed
3. Genetic principal components and genetic relatedness estimates were re-derived using methods specifically developed for genetic samples exhibiting large ancestral diversity and high genetic relatedness.
4. logRR and BAF are no longer provided due to failure to call CNV on the sex-chromosomes
5. There are some errors in the rel_relationship and rel_family_id variables, in cases where there are irresolvable inconsistencies between these variables and familial relationships derived from genetic data, we recommend to use the genetic_zygosity variables.

The inconsistencies fall into three groups:

- half siblings - these individuals do not meet the genetic relatedness threshold (0.35) to be included in the tabular data of genetic_zygosity variables (this corresponds to about 60% of the cases).
- missing genetic data (this corresponds to about 20% of the described cases).
- unresolvable inconsistencies between rel_relationship and genetic data i.e. genetic relatedness = 0. In these cases it's best to go with what is indicated by genetic_zygosity variables. We have picked up one or two errors with the genetics but we consider the rel_family_id variables to be less trustworthy and have really only left them there for legacy reasons. (This corresponds to about 20% of the cases.)

Instrument Descriptions (Tabulated Data)

Youth Instruments

Genetic Principal Components & Relatedness

Release 5.0 Data Table: gen_y_pihat

Measure Description: The data in this instrument are used to indicate genetic and related similarity. In this release we include a `rel_birth_id` which provides a unique ID for individuals that share the same birthdate. This can be used to indicate twins and triplets within the sample. This information can also be obtained through a combination of `rel_family_id` and `rel_group_id` (included for continuity with previous data releases), however we encourage users to use either `rel_birth_id` or the fields described below to identify twins and triplets. Many of the other fields in this instrument are derived from genetic data available from the Smokescreen array (non imputed data); these are indicated by a `genetic_` prefix. With this data release we have updated the process for computing genetic principal components and relatedness using methods appropriate for the genetic diversity and large number of related individuals in the ABCD sample. See section *GENESIS Derived Genetic Principal Component Weights and Relatedness Estimates* below for details. The first 32 principal components of genetic ancestry are available under fields `genetic_pc_{N}`. Genetic relatedness (as captured by $\hat{\pi}$) is available within this instrument for a subset of individuals who have $\hat{\pi} > 0.35$. These are indicated by fields `genetic_pi_hat_{N}` where `genetic_paired_subjectid_{N}` indicates the ID for whom this value corresponds. Additionally, a corresponding `genetic_zygosity_status_{N}` field indicates whether this relationship indicates monozygotic (1), dizygotic (2) or singleton siblings (3). Monozygotic relationships are those that have $\hat{\pi} > 0.8$, dizygotic relationships are those that have $0.8 > \hat{\pi} > 0.35$ and matching birth dates between pairs, and sibling relationships are those that have $0.8 > \hat{\pi} > 0.35$ and birth dates that are more than 3 months apart. Genetic relatedness across all pairs in the sample is available as part of the bulk genetic data (see section *GENESIS...* below).

ABCD Classification: Genetic

Number of Variables: 50

Summary Score(s): No

Measurement Waves Administered: Baseline

Modifications since initial administration: Genetic ancestry factors have been replaced by genetic principal components. Inclusion of a `rel_birth_id` variable indicates individuals that share the same birth (i.e. twins and triplets). Use of PC-Relate for relatedness computation, previous data releases used `PLINK --genome` for this calculation.

Notes and special considerations: None

References:

Gogarten, S. M. et al. (2019) Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* 35, 5346–5348.

Conomos, M. P., Miller, M. B. & Thornton, T. A. (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* 39, 276–293.

Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. (2016) Model-free Estimation of Recent Genetic Relatedness. *Am. J. Hum. Genet.* 98, 127–148.

Twin Zygosity Rating

Release 5.0 Data Table: gen_y_zygrat

Measure Description: The Twin Zygosity Rating study characterized photos of twins on physical characteristics to estimate zygosity.

ABCD Classification: Genetic

Number of Variables: 28

Summary Score(s): Yes

Measurement Waves Administered: Baseline

Modifications since initial administration: None

Notes and special considerations: None

Non-tabulated Genetic Data

For extraction of single genetic variants, users can use tools like [bed-reader](#) for python and [snpStats](#) for R to parse plink bed files described below. For generation of polygenic scores we have found that [PRSCs](#) has shown the best performance in ABCD data:

Ahern, J., Thompson, W., Fan, C.C. et al. Comparing Pruning and Thresholding with Continuous Shrinkage Polygenic Score Methods in a Large Sample of Ancestrally Diverse Adolescents from the ABCD Study®. *Behav Genet* 53, 292–309 (2023). <https://doi.org/10.1007/s10519-023-10139-w>

Plink files described do not contain family relatedness or sex information. For family relatedness please refer to “GENESIS Derived Genetic Principal Component Weights and Relatedness Estimates” section below, for sex please refer to [demo_sex_v2](#) in [abcd_p_demo](#).

Smokescreen Binarized PLINK Files

Files: * [ABCD_202209.updated.nodups.curated.cleaned_indivs.bed](#) *
[ABCD_202209.updated.nodups.curated.cleaned_indivs.fam](#) *
[ABCD_202209.updated.nodups.curated.cleaned_indivs.bim](#)

Measure Description: After dish quality control and profile checks, genotypes were called using Axiom Analysis Suite (apt version 2.11) on raw intensities from the Affymetrix Smokescreen array. Based on the Best Practices Analysis Workflow by Thermo Fisher, classification passed the final SNP quality controls were recommended, resulting in ~590K recommended probe sets in each genotyping batch. Blood and Saliva DNA samples were genotyped separately. We include one genotype result for each subject, using whichever sample has the best QC metrics (call rates and missingness). For release 5.0, there were nine genotyping batches, spanning across 149 plates (See [ABCD_202209.updated.nodups.curated.batch.info](#) in downloaded files). After obtaining the genotype batch, we mapped the probesets to SNPs using annotations derived from genome build hg19. After the mapping, we merged all nine batches into one study cohort and then performed additional study level QC to include missingness less than 10% in the SNP level, and less than 20% in the sample level. 515,270 variants and 11,666 people passed filters and QC. The subsequent imputation and relatedness inferences were based on the final curated genotype data. The batch information can be found in [ABCD_202209.updated.nodups.curated.batch.info](#).

ABCD Classification: Genetic

Genome Build: hg19

Number of variants: ~500k

Number of individuals: 11,666

Measurement Waves Administered: Baseline

Modifications since initial release: Genotyping of missing individuals from previous data releases due to sample mix up or failing quality control measures.

Notes and special considerations: None

Reference:

Baurley, J.W., Edlund, C.K., Pardamean, C.I. et al. (2016) Smokescreen: a targeted genotyping array for addiction research. *BMC Genomics* 17, 145.

Imputed PLINK Files using TOPMED imputation panel

Files: * [merged_chroms.bed](#) * [merged_chroms.fam](#) * [merged_chroms.bim](#)

Measure Description: The curated genotype data was used for the imputation, using the bioinformatic pipelines and recommendations of TOPMED Server, with TOPMED reference panel. The processing streams can be found at: https://github.com/robloughnan/TOPMED_Imputation_Scripts. Post processing steps to convert from [.vcf](#) files to binary PLINK files and assign RSID numbers removed triallelic variants from these files.

The TOPMED imputation scores and post-imputation quality report can be found on <https://nda.nih.gov/study.html?id=2313>

ABCD Classification: Genetic

Genome Build: GRCh38**Number of variants:** ~300 million**Number of individuals:** 11,666**Measurement Waves Administered:** Baseline

Modifications since initial release: Genotyping of missing individuals from previous data releases due to sample mix up or failing quality control measures. Previously these files were shared as vcf files, they are now available as binarized PLINK files with labeled RSID numbers for variants when available.

Notes and special considerations: None**References:**

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., ... Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10), 1284–1287

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11), 1443–1448

GENESIS Derived Genetic Principal Component Weights and Relatedness Estimates

Files: * [ABCD_202209_merged_pcair_loadings.tsv](#) *
[ABCD_202209.updated.nodups.curated_unrelateds.txt](#) *
[ABCD_202209.updated.nodups.curated_PCRelate_long.cleaned_indivs.tsv](#)

Measure Description: Accounting for Genetic Principal Components (PCs) in genetic studies (both GWAS and Polygenic Score analysis) is considered best practice to account for effects of population stratification that can lead to spurious results. Traditional approaches for calculating PCs (e.g. [FlashPCA](#)), although considered best practice for many genetic studies, may not be well suited to samples with large (known or cryptic) relatedness as is observed in ABCD. As such we have replaced these genetic PCs with ones calculated using PC-AiR – a method specifically developed and validated for samples with large family structure. Briefly, PC-AiR captures ancestry information that is not confounded by relatedness by finding a set of unrelated individuals in the sample that have the highest divergent ancestry, and computing the PCs in this set. The remaining related individuals are then projected into this space. This method is used by the Population Architecture through Genomics and Environment (PAGE) Consortium, which is principally concerned with conducting genetic studies in diverse ancestry populations.

PC-AiR was run using default suggested settings from the GENESIS package. We used non-imputed SNPs passing QCs from the 5.0 data release (~500k variants and 11,666 individuals). PC-AiR takes in kinship estimates for defining its unrelated set of individuals with divergent ancestry; this was computed

using `snpGdsIBDKING` as suggested by GENESIS authors. SNPs were LD pruned using `snpGdsLDpruning` with parameters: `method="corr"`, `slide.max.bp=10e6` and `ld.threshold=sqrt(0.1)`. This resulted in 114,707 SNPs remaining after pruning. Using the computed kinship matrix PC-Air was then run on this pruned set of SNPs. This resulted in 8,177 unrelated individuals from which PCs were derived – leaving 3,391 related individuals being projected onto this space. Subsequent analysis indicated a sample mix of 2 samples which were then removed from other genetic data. This is why the sum of unrelated and related individuals is more than the number of individuals in PLINK files ($8,177 + 3,391 > 11,668$). The weights, which can be used to project other samples into the same PC space, can be found in `ABCD_202209_merged_pcair_loadings.tsv`. The list of 8,177 unrelated individuals used for deriving P's is available in `ABCD_202209.updated.nodups.curated_unrelateds.txt`.

After Computing PCs from PC-Air, we then computed a genetic relatedness matrix (GRM) using PC-Relate. PC-Relate aims to compute a GRM that is independent from ancestry effects as derived from PC-Air. PC-Relate was run on the same pruned set of SNPs described above using the first two PCs computed from PC-Air. Identity by descent probabilities between individuals i and j were calculated as $\hat{\pi}_{ij} = \hat{k}_{ij}^{(2)} + 0.5 \times \hat{k}_{ij}^{(1)}$, where $\hat{k}_{ij}^{(2)}$ and $\hat{k}_{ij}^{(1)}$ represent the probabilities that individuals i and j share 2 or 1 alleles at a locus – calculated from PC-Relate. For all off-diagonal elements of the GRM we provide estimates of $\hat{k}_{ij}^{(0)}$, $\hat{k}_{ij}^{(1)}$, $\hat{k}_{ij}^{(2)}$, $\hat{\pi}_{ij}$ and genetic relatedness in `ABCD_202209.updated.nodups.curated_PCRelate_long.cleaned_indivs.tsv`. For a subset of related individuals ($\hat{\pi} > 0.35$) we include estimates in tabulated data in the `gen_y_pihat` instrument described above. Code used to perform the processes described in this section can be found here: https://github.com/robloughnan/ABCD_GeneticPCs_and_Relatedness.

ABCD Classification: Genetic

Number of individuals: 11,666

Measurement Waves Administered: Baseline

Modifications since initial release: Genotyping of missing individuals from previous data releases due to sample mix up or failing quality control measures. Use of PC-Relate for relatedness computation, previous data releases used `PLINK --genome` for this calculation.

Notes and special considerations: None

References:

Gogarten, S. M. et al. (2019) Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* 35, 5346–5348.

Conomos, M. P., Miller, M. B. & Thornton, T. A. (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* 39, 276–293.

Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. (2016) Model-free Estimation of Recent Genetic Relatedness. *Am. J. Hum. Genet.* 98, 127–148.

