Release Notes: Adolescent Brain Cognitive Development Study^{SM} (ABCD Study®) Data Release 4.0

Genetics

http://dx.doi.org/10.15154/1523041
October 2021

Change Log
October 2021 – Data Release 4.0

- Initial release

List of Instruments

Name of Instrument	Short Name
Genomics Sample*	genomics_sample_03
Experiment Description	omics_experiments
ABCD Youth Genetic Blood (RUCDR)	biocf01
ABCD Youth Genetic Saliva (RUCDR)	abcd_ygs01

General Information

The following information refers to the Adolescent Brain Cognitive Development Study^{SM} (ABCD) Data Release 4.0 available from https://nda.nih.gov/abcd. An overview of the ABCD Study® is at https://abcdstudy.org and detailed descriptions of the assessment protocols can be viewed at https://abcdstudy.org/scientists/protocols.

This document describes the contents of various instruments available for download. To understand the context of this information, see *Release Notes ABCD README FIRST* and *Release Notes ABCD Imaging Instruments*.

Summary of genotyping bulk release

- Genotype calls in PLINK format
    - Re-clustered data from Release 2.0
    - Sample size up to 11099 unique individuals.
- Release of LRR and BAF from well QCed ABCD genotype calls for researchers who want to do CNV calls
- Release of TOPMed Imputed ABCD data in VCF format

Genomics Data (SmokeScreen)

Genomics_sample_03: Each entry in this instrument references the same plink in a zip-file.

1. Genotyping Platform
    a. Affymetrix NIDA SmokeScreen Array
    b. Rutgers RUCDR performed sample preparation and genotyping, including:
        i. Extraction kit: Chemagen bead based/Chemagic STAR DNA Saliva4k Kit (CMG-1755-A)
        ii. Processing: DNA fragmentation, labeling, ligation, and hybridization
        iii. Equipment: Affimetrix GeneTitan Instrument.
    c. For additional information, see NIMH experiment description #1194.
    d. The smokescreen array contains 733,293 SNPs (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4769529/)

2. Genotyping Data
    a. We identified issues regarding the genotyping data following Release 2.0. Release 3.0 is based on genotyping re-clustering of all ABCD samples spanning six different batches (see ABCD_release3.0_batch_info.txt). To maximize the number of good QCed samples, we included samples from saliva and whole blood, whichever had: a) higher successful calls, b) higher none-missing, c) matched genetic sex, and d) less excessive IBS.
    b. Quality control was performed as follows:
        i. RUCDR performed DNA quality controls based on calling signals and variant call rates.
        ii. ABCD DAIRC performed the subsequent study-based QC process, following the recommendation of Ricopili pipeline (https://doi.org/10.1093/bioinformatics/btz633)

c. The QCed genotyping data is in binary PLINK format that contains 11,099 unique individuals with 516,598 genetic variants. All genetic variants referenced in positive strand. The data files contain:
   i. ABCD_release_3.0_QCed.bed
   ii. ABCD_release_3.0_QCed.fam
   iii. ABCD_release_3.0_QCed.bim

d. The FID is the sample collection ID while the IID is the ABCD study ID.

e. Summary of the difference comparing to release 2.0:
   i. Resolved sample issues and corrected assignments.
   ii. Genotypes were re-clustered.
   iii. Sample size increased to 11099 individuals.

3. Log R Ratio and B Allele Frequencies

a. To enable researchers to call CNVs using ABCD genotype data, we have generated the LRR and BAF from the intensity files of genotyping calls from Affymetrix Smokescreen array. The pipeline is identical to the process described in the supplement materials of Kendall et al. (2016; https://doi.org/10.1016/j.biopsych.2016.08.014).

b. The LRR and BAF were derived from five batches of genotype calls, including 11,088 well QCed individuals.

c. The batch information can be found in ABCD_release3.0_.batch_info.txt. The data files contain:
   i. *.sample
      1. The sample information
   ii. *.probe.info
      1. The probe information and AB definition
   iii. *_lrr.txt.gz
      1. LRR as a M by N numeric matrix, M as the probe and N as the sample.
   iv. *_baf.txt.gz
      1. BAF as a M by N numeric matrix, M as the probe and N as the sample.

d. There are five batches as described in the prefix, including four saliva batches and one whole blood batch.

4. ABCD imputed genotype data with TOPMED reference.

a. We provide imputed whole genome data to the research community in Release 3.0.

b. The released files include dosage files in VCF format and imputation INFO files.
   i. No post-imputation QC was performed.
   ii. We recommend removing two subjects from analyses due toa subject matching issue (see SUBJ_QC_BAD.txt).
   iii. The sample ID in the VCF format is the concatenation of Collection ID and Study ID.

c. Imputation was performed using QCed genotype data with the process described as the following:

  i. Imputation was performed using the TOPMed imputation server.
  ii. Pre-imputation steps were followed as instructed at https://topmedimpute.readthedocs.io/en/latest/prepare-your-data/. These steps involved:
     1. Calculating allele frequencies using PLINK v1.9,
     2. Executing the HRC-1000G-check-bim.pl script that checks bim files against HRC/1000G for consistencies,
     3. Conversion to VCF files using plink v1.9
     4. Running the checkVCF.py to verify that VCF conversion was successful. VCF files were next uploaded to the TOPMed Imputation Server and imputation was performed using mixed ancestry and Eagle v2.4 phasing.

NOTE

In our QC process, we found the plate 461 to be especially problematic. We recommend not using genotype data from plate 461, or at least including plate number as a covariate.

To use genomic data for population indices, such as genetic ancestry, population stratification, genetic relationships, and zygosity inference (described below), the user can employ the released QCed genetic data directly. Official release of these indices will be included in a future release.

Zygosity Calculation

The zygosity is determined by probability of identity-by-descent. We recommend the following steps to calculate the zygosity:

a. Selecting independent SNPs based on:
   a. allele frequency > 0.05,
   b. missing rate < 0.1.
   c. pruning to ensure independency.
b. We suggest using PLINK with default parameters to perform this task.
c. Based on selected SNPs, calculate probability of identity-by-descent.
d. We suggest using the PLINK command --genome to get the probability of identity-by-descent, PI_HAT
e. Determine zygosity based on calculated probability of identity-by-descent

Theoretically, P(IBD) =~ 0.5+/- 0.039 (Visscher et al, 2006) for full siblings or dizygotic twins, and 1 for monozygotic twins. However, given the estimations are based on identity-by-state and there is inherent noise in the genotyping data, we recommend using 0.4 and 0.8 as the thresholds. All the processing pipelines will be available on ABCD GitHub (https://github.com/ABCD-STUDY).

Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, et al. (2006) Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genet 2(3): e41. https://doi.org/10.1371/journal.pgen.0020041

ABCD Youth Genetic Blood (RUCDR) - Information collected by RAs at the time of venipuncture blood collection for genetics studies.

ABCD Youth Genetic Saliva (RUCDR) - Information collected by RAs at the time of saliva collection for genetics studies