2018 Add Health Documentation

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Quality Control Analysis of Add Health GWAS Data

Report prepared by

Heather M. Highland

Christy L. Avery

Qing Duan

Yun Li

Kathleen Mullan Harris

UNC CAROLINA POPULATION CENTER

CAROLINA POPULATION CENTER | CAROLINA SQUARE - SUITE 210 | 123 WEST FRANKLIN STREET | CHAPEL HILL, NC 27516

This research was supported by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Further information may be obtained by contacting Add Health at adhealth@unc.edu

Quality Control Analysis of Add Health GWAS Data

Heather M. Highland, Christy L. Avery, Qing Duan, Yun Li, and Kathleen Mullan Harris
University of North Carolina at Chapel Hill
March 26, 2018

Add Health design

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is an ongoing, nationally-representative longitudinal study of the social, behavioral, and biological linkages in health and developmental trajectories from early adolescence into adulthood. The Add Health cohort was drawn from a probability sample of 132 middle and high schools and is representative of American adolescents in grades 7-12 in 1994-1995. The adolescent cohort has been followed for 20+ years with in-home interviews in 1995 (Wave I, 79% response rate), 1996 (Wave II, 89% response rate), 2001-02 (Wave III, 77% response rate), and 2008 (Wave IV, 80% response rate). Wave V is currently underway in 2016-2018 when respondents are aged 32-42. The Add Health design included an embedded genetic sample of pairs of identical twins, fraternal twins, full sibs, half sibs, and adolescents who grew up in the same household but have no biological relationship. Add Health is a multiracial and multiethnic sample with substantial numbers of individuals with Hispanic and Asian ancestry. For more information about the design of Add Health see Harris 2010 and Harris et al. 2013.

Genome-wide Genotyping

At Wave IV, Add Health collected Oragene saliva samples from consenting participants (96% of n=15,701), and requested a second consent to archive their samples for future genomic studies. Approximately 80% consented to archive and were thus eligible for genome-wide genotyping. Genotyping was completed over three years funded by R01 HD073342 (PI Harris) and R01 HD060726 (PIs Harris, Boardman, and McQueen). Add Health utilized two Illumina platforms for genotyping: the Illumina Human Omni1-Quad BeadChip for the majority of samples and the Illumina Human Omni-2.5 Quad BeadChip for the remainder. The two platforms utilized tag SNP technology to identify and include over 1.1 million and 2.5 million genetic markers respectively from Omni1 and Omni2.5 derived from the International HapMap Project and the most informative markers from the 1000 Genomes Project (1KGP). The genetic markers include known disease-associated SNPs from multiple sources, ancestry-informative markers, sex chromosomes, and ABO blood typing markers. The platforms also included probes for the detection of copy number variation (CNV) covering all common CNV regions and more than 5,000 rare CNV regions. After quality control procedures (described below), genotype data were available for 9,974 individuals: n=7,917 from the Illumina HumanOmni1-Quad chip and for 2,057 individuals from the Illumina HumanOmni2.5-Quad chip (Figure 1). After filtering, the Add Health genotype GWAS data contained n=609,130 single-nucleotide polymorphisms (SNPs) common to both chips to enable joint imputation to the entire Add Health population (see below).

Quality Control (QC) report for Illumina HumanOmni1-Quad. Below we describe our approach to participant- and SNP-level exclusions (Table 1).

SNP Marker Set. Initially, 9,344 individuals (Add Health samples, duplicates, and HapMap controls) were genotyped by Illumina HumanOmni1-Quad with 1,140,419 markers each. A total of 1,058,111 markers are annotated by Illumina annotation file HumanOmni1-Quad-v1-0_H (82,308 markers excluded). The following markers were removed: 142,695 A/T or G/C markers due to alignment uncertainty; 1,985 markers that were triallelic or did not map to chromosome 1-22 or chromosome X. After these exclusions, 913,431 markers entered the first-pass marker level QC. Next, we removed markers with call rates < 90% (N=20,008) and minor allele frequency (MAF) < 0.5% (N=33,333). After the first-pass marker level QC, 860,090 markers remained.

Sample-level QC. Next we conducted sample level QC by removing 308 individuals with < 90% call rate. After this step, the total genotyping rate in the remaining 9,036 individuals reached 0.995. We then converted IDs assigned during genotyping to Add Health ID or HapMap ID, dropping three individuals who could not be mapped (n=9,033 remained). After excluding 106 HapMap controls, a total of 8,927 Add Health samples (with duplicates) remained.

HapMap controls concordance check. A total of 106 HapMap controls were genotyped together with the Add Health samples, 99 of which were HapMap sample NA12003 and 7 of which were HapMap sample NA18856. We then compared the strand of HapMap controls and the external HapMap sample, NA12003, and performed the concordance check using the 518,630 overlapping and bi-allelic markers. The mean genotype matching error was 0.0045 and the mean allele matching error was 0.0023. Out of 518,630 markers, 5600 (1%) had mismatch rate > 0.01.

Duplicate Concordance. A total of 288 duplicate pairs were genotyped. Using PLINK to estimate IBD, 281 of the pairs had a PI_HAT >0.99. The seven pairs with PI_HAT<0.99 were dropped and the two plates on which these pairs resided were flagged for additional QC. For matching pairs, the genotyping instance with the highest call rate was retained.

Sex check and heterozygosity check. We inferred genetic sex based on X chromosome heterozygosity as implemented in PLINK separately by self-reported race/ethnicity. Autosomal heterozygosity was also calculated and compared to X chromosome estimates. Using the heterozygosity estimate (F), genetic sex was defined to be female if F < 0.2 and male if F > 0.8 (PLINK default); F values 0.2-0.8 indicated an ambiguous sex. A total of 94 genetic sex-discordant samples were dropped and the plates on which the samples resided were flagged for additional QC. An additional 44 individuals with higher than expected autosomal heterozygosity or ambiguous X heterozygosity that may indicate a sex mismatch also were dropped; sex mismatches did not favor a particular direction, indicating that sample swaps, not contamination, was the likely problem (Table 2; Figure 2).

Second-pass marker QC, HWE by ancestry. Marker level QC was repeated on the remaining individuals. Six markers were removed due to a call rate <90%; 120 markers were removed due to a MAF<0.5%; and 7,953 were removed due an ancestry specific Hardy-Weinberg Equilibrium $P<5x10^{-5}$. Further we removed 4,375 variants that shows low concordance between duplicate pairs.

Summary. The five preceding QC steps identified 55 plates with no evidence of sample switching through sex-mismatches or duplicate mismatch and 50 plates with evidence of at least one sample swap. Our strategy to identify participants with high certainty of sample identity on the 50 plates with evidence of sample misidentification was performed in conjunction with the data genotyped on the Omni 2.5 and therefore follows our description of Omni 2.5 QC.

QC report for Illumina HumanOmni2.5-Quad. Below we describe our approach to participant- and SNP-level exclusions (Table 1).

SNP Marker Set. Initially, 2,461 individuals (Add Health samples, duplicates, and HapMap controls) were genotyped by Illumina HumanOmni2.5-Quad with 2,369,541 markers each. All markers were in the annotation file and none were removed due to alignment uncertainty. We then removed 9,328 markers not on chromosome 1-22 and chromosome X, leaving 2,360,213 markers entering the first-pass marker level QC. Further, for first pass QC, we removed markers with call rate < 90% (N=56,182) and MAF < 0.5% (N=571,200). After the first-pass marker level QC, 1,732,831 markers remained.

Sample level QC. Next we conduct sample level QC by, first, removing 65 individuals with <90% call rate. After this step, the total genotyping rate in the remaining 2,396 individuals reached 0.9941. Next, we converted IDs assigned during genotyping to Add Health ID or HapMap ID; eight individuals could not be mapped (n=2,391 individuals remaining). Finally, after excluding the 30 HapMap controls, a total of 2,361 Add Health samples (with duplicates) remained.

HapMap controls concordance check. A total of 30 HapMap controls were genotyped together with the Add Health samples, all of which were from HapMap sample NA12003. For example, we compared the strand of HapMap controls and the external HapMap sample, NA12003, and performed the concordance check using the 468,501 overlapping and bi-allelic markers. Out of 464,508 markers, 2,166 (0.5%) had mismatch rate > 0.01, with remaining markers' mean mismatch rate =0.

Duplicate Concordance. A total of 48 duplicate pairs were genotyped. Using PLINK to estimate IBD, all of the pairs had a PI_HAT >0.99.

Sex check and heterozygosity check. We inferred genetic sex based on X chromosome heterozygosity, as implemented in PLINK, separately within each self-reported race/ethnicity. Autosomal heterozygosity was also calculated and compared to X chromosome estimates. Genetic sex was again defined to be female if F < 0.2 and male if F > 0.8 (PLINK default), with F values 0.2-0.8 indicative of ambiguous sex. Samples that were the opposite of self-reported sex (N=23) were dropped and plates on which these samples resided were flagged for additional QC. In addition, individuals with higher than expected autosomal heterozygosity, or ambiguous X heterozygosity that may indicate a sex mismatch (N=16) were also dropped. Sex mismatches did not favor a particular direction, indicating sample swapping, rather than contamination as the likely problem (Table 2, Figure 2).

Second-pass marker QC, HWE by ancestry. Marker level QC was repeated on the remaining individuals. One hundred ninety three markers were removed due to a call rate <90%; 6,508 markers were removed due to a MAF<0.5%; and 4,607 were removed due an ancestry specific Hardy-Weinberg Equilibrium $P<5x10^{-5}$. Further we removed 5,581 variants that shows low concordance between duplicate pairs.

Summary. In total, the five preceding QC steps identified 13 plates with no evidence of sample switching through sex-mismatches or duplicate mismatch and 14 plates with evidence of at least one sample swap. Below we describe our strategy to identify participants with high certainty of sample identity on the 14 plates with evidence of sample misidentification. This was done in conjunction with the data genotyped on the Omni 1.

Add Health Exome Chip data. Given the non-negligible proportion of plates with at least one potential instance of sample swapping, we performed a third round of QC by comparing all individuals genotyped on Omni 1 or 2.5 to previously performed Exome Chip genotyping, funded by R01 HD057194 (PIs: Gordon-Larsen and North). Because the samples used for the exome genotyping were the first specimens selected from the Add Health DNA archive, they are considered to be of higher quality than those used for subsequent genotyping. (Consistent with this expectation, QC of Exome Chip data indicated genotypes of high fidelity (e.g. >99% concordance in genetic vs. self-reported sex; 97% of participants with SNP missingness <3%).

Exome Pairs. First, data from the Exome Chip and both Omni arrays were aligned and combined. Only markers well-genotyped on all three arrays were retained (n=2,961markers). A common dataset was then constructed, which included n= 20,150 individuals (with duplicates) and relatedness between these 20,150 individuals was calculated using PRIMUS[4] (https://primus.gs.washington.edu/primusweb/). Of the 8,662 expected pairs (i.e. individuals genotyped on the Exome Chip and one of the Omni arrays), 8,585 matched (99%). Although the majority individuals that did not match their Exome Chip genotype resided on plates already flagged for additional follow up, an additional 11 plates were identified as containing a mismatch, and flagged for additional QC.

Sample Retention. All samples that resided on any of the 67 plates without evidence of sample swaps were retained (n=5,539), as were all Omni 1 and 2.5 array individuals who matched their ExomeChip pair (n=4,164) (Figure 1). For individuals that were not genotyped on the Exome Chip, they were retained if they showed a relationship as expected by self-report (i.e. if A and B reported they were full siblings and they showed IBD estimates consistent with that relationship or a one degree further away relationship (i.e. half-sibs), we retained both individuals) (n=338) (Figure 3). In total, 10,041 samples were retained.

Merging Omni1 and Omni2.5 arrays and imputation. After initial QC, we merged the two genotyped SNP data sets. There were a total of 1,954,448 markers on the two arrays, of which 609,130 were common to both arrays and passed the above QC steps (call rate in merged data > 80%). As strand files were not available, all palindromic (i.e G/C or A/T) SNPs were dropped. Two imputation releases are provided: imputation of European ancestral participants using the HRC imputation panel and the GSCAN protocol (see below); and imputation of all (i.e. multiethnic) Add Health participants to the 1000 Genomes Phase 3 imputation panel using the GLGC/GIANT protocol (see below). No filtering was performed after imputation.

Relatedness

More precise relatedness was calculated in all remaining individuals using PRIMUS[4] (https://primus.gs.washington.edu/primusweb/). Sixty-six (n=10,041-66; 0.66%; 9,974 unique participant retained) participants appeared to be closely related to more individuals than is plausible. Because excessive “relatedness” can reflect genotyping error, these participants were removed from the genotype data. We also removed one unintentional duplicate. In the remaining 9,974 participants, the genetic relatedness matrix (GRM) was calculated using PLINK’s --make-rel command. We used the identity by descent (IBD) estimates to identify 831 full-sibling pairs irrespective of self-reported kinship or demographic information. No other pairs (e.g. half-siblings, avuncular etc.) were included in the pedigree file. Users should be aware that half-siblings, cousins, and avuncular relationships were not included in the pedigree file.

Population structure. To evaluate population structure, we calculated principal components in the unrelated subset in eigenstrat[5]. The remaining individuals and HapMap3 reference samples were projected into this space (Figure 4). These eigenvectors may be used as covariates in the statistical model used for association testing to adjust for possible population stratification.

Table 1. SNP- and sample-level exclusions yielding n=10,778* unique Add Health participants for second-pass QC.

	Omni1		Omni 2.5	
Procedure	**N. markers**	**N. individuals**	**N. markers**	**N. individuals**
Start	1,140,419	9,344	2,369,541	2,461
Markers in the annotation file	1,058,111	9,344	na	na
Removing A/T G/C markers and markers with both alleles missing/alignment uncertainty	915,416	9,344	na	na
Removing triallelic markers and markers not on chromosomes 1-23	913,431	9,344	2,360,213	2,461
Removing markers with call rate < 90%	893,423	9,344	2,304,031	2,461
Removing markers with call rate MAF < 0.5%	860,090	9,344	1,732,831	2,461
Removing individuals with call rate < 90%	860,090	9,036	1,732,831	2,396
Updating EAID to AID	860,090	9,033	1,732,831	2,391
Exclude HapMap controls	860,090	8,927	1,732,831	2,361
Duplicate concordance check	860,090	8,645	1,732,831	2,314
Sex check	860,090	8,551	1,732,831	2,291
Het check + sex check	860,090	8,507	1,732,831	2,275
2nd-pass call rate< 90%	860,084	8,507	1,732,638	2,275
2nd-pass MAF	859,964	8,507	1,726,130	2,275
HWE	852,011	8,508	1,721,523	2,275
Poor duplicate concordance	847,636	8,508	1,715,942	2,275

*Five Add Health participants were genotyped on both Omni 1 and Omni 2.5 chips, yielding a combined sample size of 10,783; n=10,778 unique participants.

Table 2: Direction of Sex Mismatches.

	F < 0.2, female; F > 0.8, male			
	Omni1		Omni2.5	
	Reported sex M; genetic sex F	Reported sex F; genetic sex M	Reported sex M; genetic sex F	Reported sex F; genetic sex M
White	40	40	1	5
Black	6	6	NA	NA
Native American	0	0	0	0
Asian	1	0	3	5
Hispanic	1	0	6	4

Figure 1. Second-pass QC sample retention strategy. Of the initial 10,778 samples retained after SNP- and sample-level exclusions (Table 1), 5,539 samples were genotyped only on plates with no evidence of sex mismatches or duplicate errors. In order to maximize our sample size, we then performed third pass QC, which examined independently genotyped exome chip data and previously reported kinship. Using these independent sources, we retained a total of n=4,164 whose Omni and Exome Chip data matched, n=338 participants that showed a biologically plausible relationship (within 1 degree of self-reported relationship) with another genotyped participant that passed QC (either Exome or Omni data). After these QC steps, we identified a sample size of n=10,041 participants.

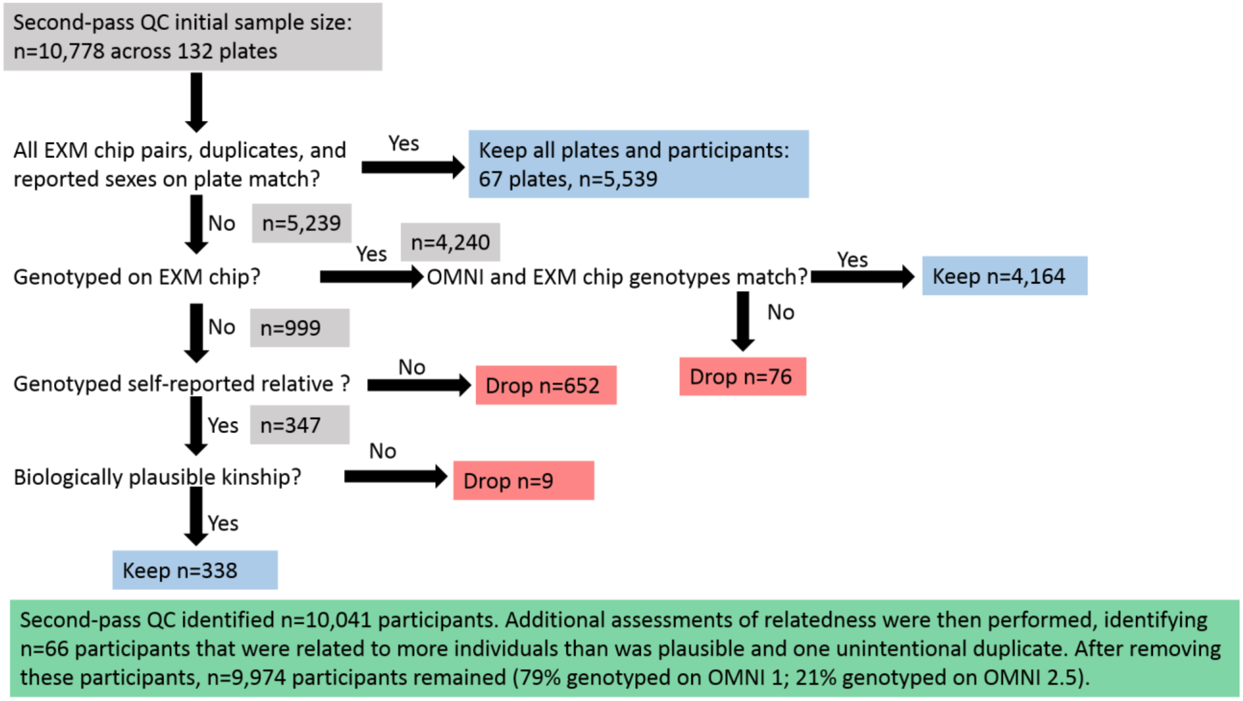

Second-pass QC initial sample size:
n=10,778 across 132 plates
All EXM chip pairs, duplicates, and reported sexes on plate match?
Yes
Keep all plates and participants: 67 plates, n=5,539
No
n=5,239
Genotyped on EXM chip?
Yes
n=4,240
OMNI and EXM chip genotypes match?
Yes
Keep n=4,164
No
Drop n=76
No
n=999
Genotyped self-reported relative ?
No
Drop n=652
Yes
n=347
Biologically plausible kinship?
No
Drop n=9
Yes
Keep n=338
Second-pass QC identified n=10,041 participants. Additional assessments of relatedness were then performed, identifying n=66 participants that were related to more individuals than was plausible and one unintentional duplicate. After removing these participants, n=9,974 participants remained (79% genotyped on OMNI 1; 21% genotyped on OMNI 2.5).

Figure 2. Autosomal and sex heterozygosity, by race/ethnicity and genotyping platform.

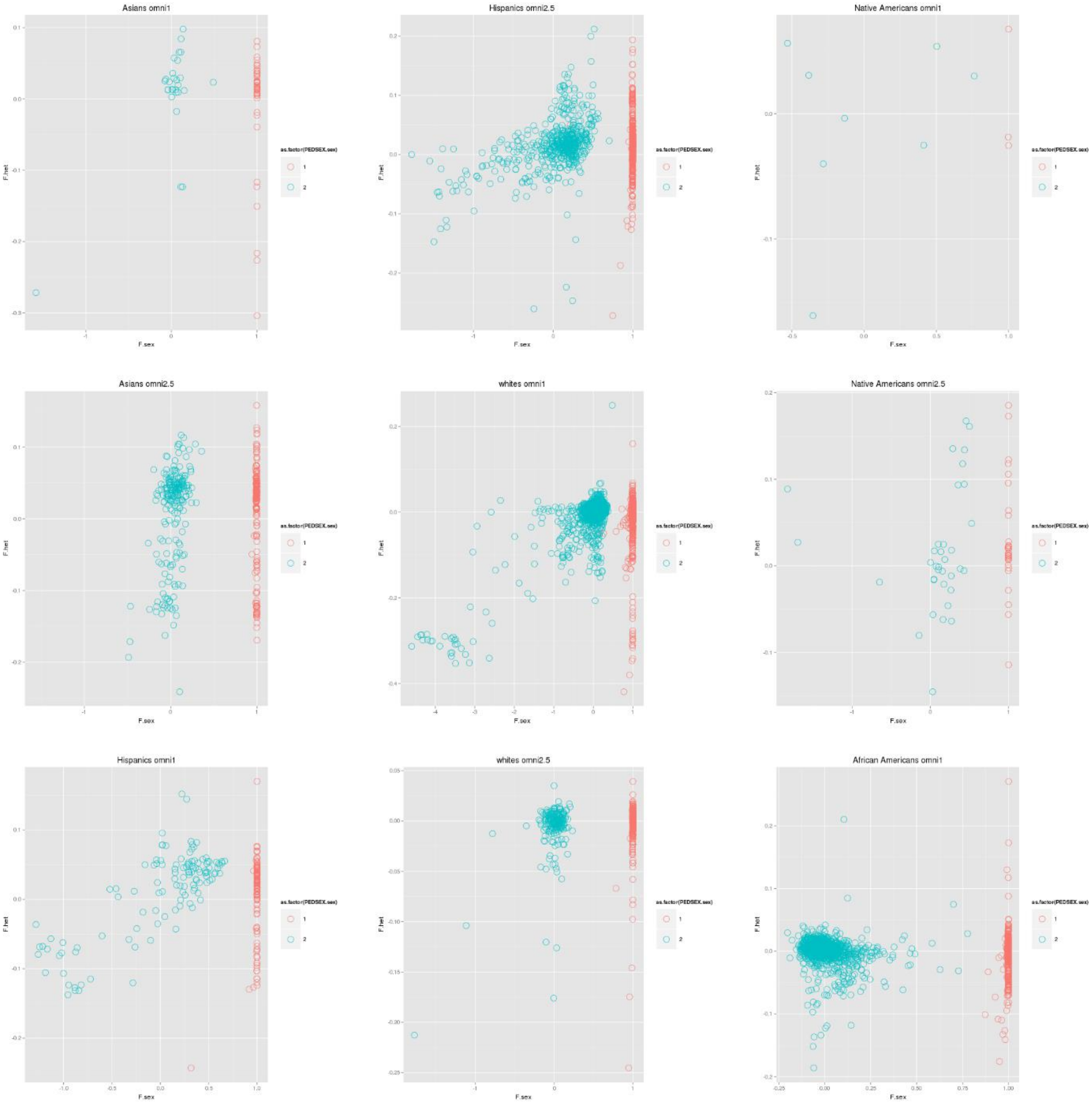


Figure 3. Allele sharing color coded by self-reported relationship

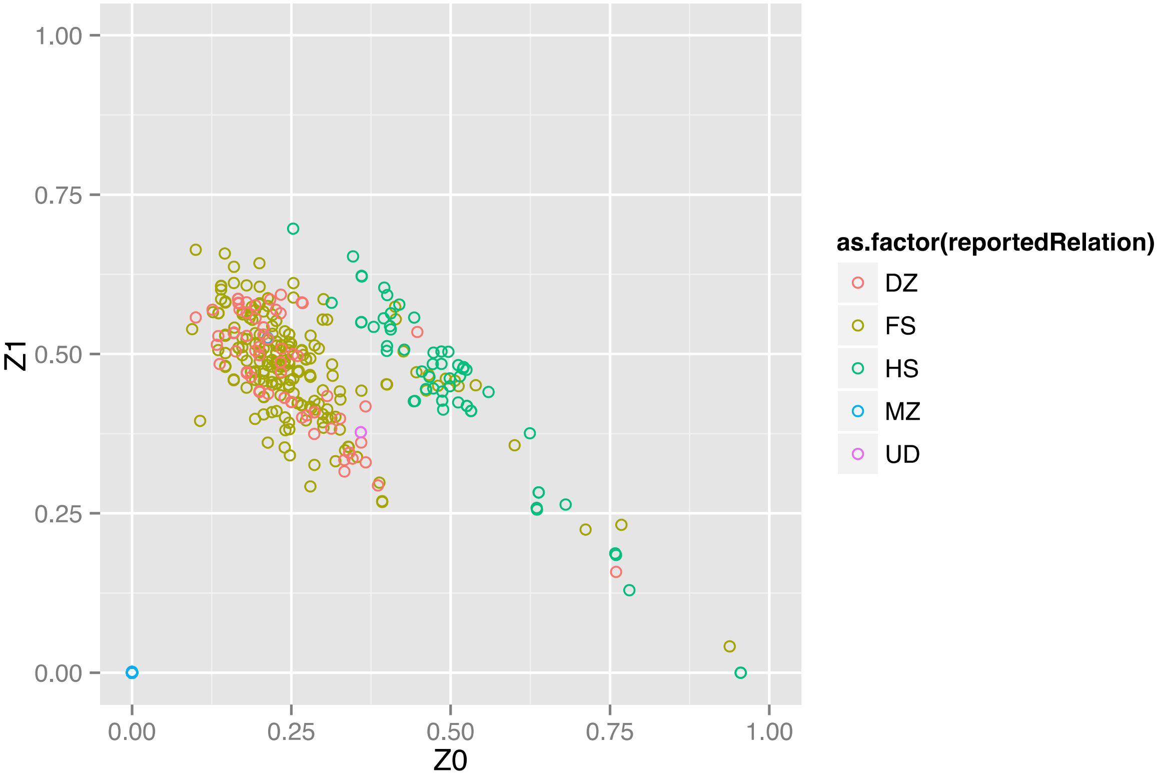
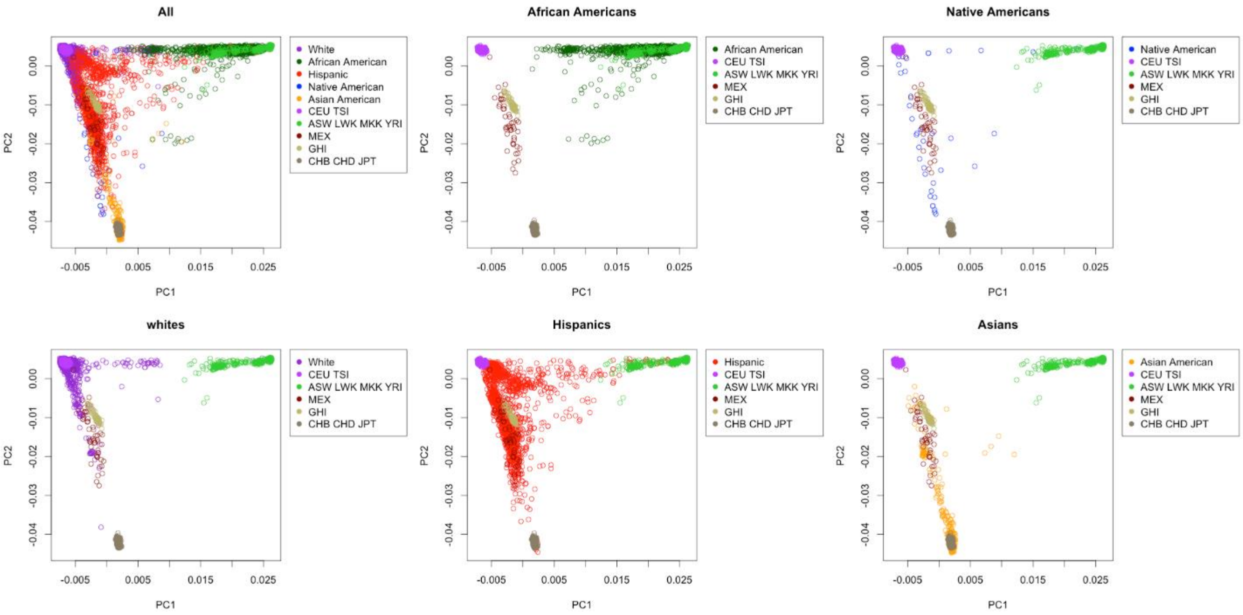


Figure 4. PCA plots overall and by self-reported race/ethnicity.



LITERATURE CITED

1. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).
2. Altshuler, D. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
3. Bahlo, M. *et al.* Saliva-Derived DNA Performs Well in Large-Scale, High-Density Single-Nucleotide Polymorphism Microarray Studies. *Cancer Epidemiology Biomarkers & Prevention* **19**, 794-798 (2010).
4. Staples, J. *et al.* PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am J Hum Genet* **95**, 553-64 (2014).
5. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

Add Health Haplotype Reference Consortium (HRC) Imputation Protocol

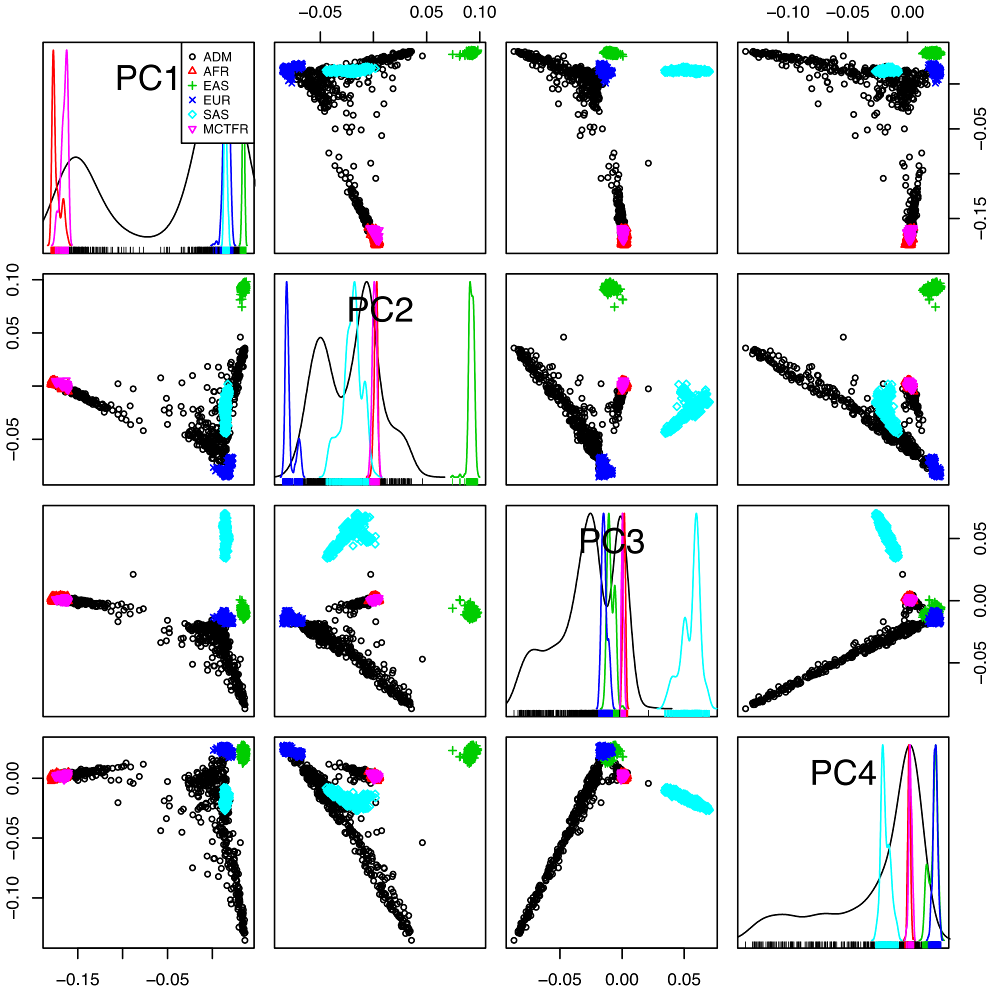
Summary: Europeans in the final release of the Add Health genotyped data were imputed with Release 1 of the Haplotype Reference Consortium reference panel (HRCr1.1).[1] The HRC panel currently provides the most comprehensive imputation panel for European ancestry. Amongst other uses, HRC imputed data may commonly be used both for genome-wide association studies or for the construction of polygenic scores using imputed (rather than genotyped) data. For imputation of non-European samples, the 1000 Genomes Phase 3 reference panel[2] is currently more appropriate.

SNP-Level Filters: Before imputation, some SNP-level filters were applied. Of 606,673 variants in the final Add Health genotyped data, 13,721 were removed with a per-variant missing call rate filter of 0.02; 245,589 were removed with a Hardy-Weinberg Equilibrium filter of 0.0001, and 609 were removed with a minor allele frequency filter of 0.01, leaving 346,754 SNPs carried through to imputation.

Individual-Level Filters: Some individual-level filters were also applied. First, the sample was reduced to only individuals of genetically-determined European ethnicity. European ethnicity was ascertained through the protocol developed for the common variant association analysis of the GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN).

The GSCAN website is available here: http://gscan.sph.umich.edu, and the ethnicity protocol can be found in the GSCAN GWAS analysis plan available here: https://ibg.colorado.edu/mediawiki/index.php/GSCAN#Phenotype_definitions_and_analysis_plan_2.

Results of that Principal-Components based analysis are plotted here:



In total, of 9,974 individuals in the final Add Health genotyped data, 4,187 non-European individuals were dropped.

Next, individuals with high rates of genotyping failures were removed with a per-sample missing call rate of 0.05. In total, of 9,974 individuals in the final Add Health genotyped data, 40 non-individuals were dropped.

Finally, individuals with excessively high or low rates of heterozygosity were removed by dropping those with an F-statistic lower than -0.3 or higher than 0.3. In total, of 9,974 individuals in the final Add Health genotyped data, 58 individuals were dropped.

Ancestral outliers were screened for using the Plink Identity-By-State (IBS) binomial test with a threshold of 0.05, though no outliers were identified.

In summary, 5,690 individuals were carried through to imputation.

Imputation: Imputation was performed with the Michigan Imputation Server[3], using the HRC r1.1 2016 reference panel and ShapeIt v2.r790 phasing software. Options were specified for a European (“EUR”) population and for “Quality Control & Imputation” mode.

The free Michigan Imputation Server is available here:
https://imputationserver.sph.umich.edu/index.html.

Imputation was completed and imputed .vcf files downloaded on July 25, 2017; imputation execution time lasted 38 hours, 40 minutes, and 19 seconds. All imputation procedures were conducted by Robbee Wedow (rwedow@alumni.nd.edu).

Add Health 1KG phase 3 Imputation Protocol

Summary: As described above, 1000 Genomes imputation is more appropriate for non-European samples. Thus, Add Health investigators used the Global Lipids Genetics/GIANT Consortium Imputation and Analysis Plan, version 2 (updated March 17th, 2017) to guide imputation to the 1000 Genomes reference panel. Imputation was performed on the University of Michigan imputation server (https://imputationserver.sph.umich.edu) All n=9,974 Add Health participants were imputed using this analysis plan.

There were two major components to this analysis plan:

1) Genome-wide genotypes must be on the correct build (37/hg19) and correct strand (forward).
2) Imputation of genotypes is performed using the 1KG phase 3

Standard pre-imputation QC criteria were applied:

- Sample call rate (cut off >95% threshold recommended)
- Exclude samples with heterozygosity > median + 3*IQR
- Remove gender mismatches
- Remove duplicates
- Remove PCA outliers using a PCA projection of the study samples onto 1KG reference samples.
- Hardy-Weinberg $p>10^{-6}$, SNP call rate ≥98%
- Remove monomorphic markers

Prepare files for imputation

The “HRC/1KG Imputation Preparation and Checking Tool” developed by Will Rayner was used to check input data for accuracy relative to expected HRC or 1000G inputs prior to imputation. This process was used to identify errors in input data, including incorrect REF/ALT designations, incorrect strand designations, extreme deviations from expected allele frequencies, and palindromic (A/T and G/C) SNPs with allele frequency near 0.5 that are often the source of imputation errors, and generates commands to make files that have fixed or removed these problematic variants. The tool can be downloaded here: http://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim.v4.2.5.zip

Impute to 1KG phase 3 panel

The 1KG phase 3 samples has both European and non-European samples and includes SNPs, indels, and CNVs.

The following options at the Michigan Imputation Server were used:

- **Reference Panel: 1000G Phase3 v5**
- **Phasing:** Eagle v2.3
- **Population:** as appropriate for the study (used only for quality control purposes)
- **Mode:** Quality Control & Imputation

Note: Chromosome X will need to be imputed separate from other chromosomes after reformatting plink code 23 to X in the VCF file. Currently, SHAPEIT is the only available method for phasing chrX.

Imputation was completed and imputed .vcf files downloaded in November, 2017. All imputation procedures were conducted by Robbee Wedow (rwedow@alumni.nd.edu).

References

1. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** (2016).
2. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
3. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48,** 1284–1287 (2016).