NATIONAL HEALTH AND AGING TRENDS STUDY (NHATS)
Quality Control and Imputation Report for Genotypic Data

August 2023

Suggested Citation:

Ware EB, Battle SL, Skehan M, Arking DE, Freedman VA, Schrack JA, Kasper JD. 2023. National Health and Aging Trends Study Quality Control and Imputation Report for Genotypic Data. Baltimore: Johns Hopkins Bloomberg School of Public Health. Available at www.NHATS.org. This work is supported by U01AG032947 from the National Institute on Aging.

Contents

I. Summary and recommendations for NIAGADS users

This document outlines the genetic sample collection, quality control processing, and imputation decisions made by the NHATS team to produce genotyped and imputed data on the NHATS dried blood spot samples. There are minimally QCed genomic data on 4,015 participants available through an application to the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS). Access to the NHATS data in the NIAGADS storage site requires a research use statement, IRB letter, data use certification, NIAGADS data distribution agreement, NIA genomic data sharing policy, derived/secondary data return plan, and PI biosketch. Data may only be requested by a qualified principal investigator. Please see the NIAGADS website at https://www.niagads.org/ to review documentation and for the most updated data access requirements. When using genomic data from NIAGADS, please cite "NIAGADS: The NIA Genetics of Alzheimer's Disease Data Storage Site. Alzheimer's and Dementia, 12(11): 1200-1203". An in-text example acknowledgement statement is as follows: *The data used for the analyses described in this manuscript were obtained from the NIAGADS GenomicsDB on MM/DD/YY*.

Through a request to NHATS, the NHATS study team provides a NHATS sample filtering document that includes abbreviated information on the reason a sample was dropped during further quality control, sample plate and well location, and an indication of whether a sample is in one of the analytic groups. These groups are defined as non-Hispanic white with European ancestry (n=2791), non-Hispanic Black with African ancestry (n=667), other (n=393), or NA (n=164), indicating the sample was dropped during quality control steps. We strongly recommend against analyzing heterogeneous ancestry groups together.

To link the genetic data from NIAGADS to NHATS study data, users must complete an application on the NHATS website https://nhats.org/researcher/data-access/sensitive-data-files?id=restricted_data_files.

II. Overview of NHATS Data Collection

Begun in 2011, the National Health and Aging Trends Study (NHATS) conducts annual in-person interviews with a nationally representative sample of approximately 8,000 Medicare beneficiaries ages 65 or older. The study supports research on late-life disability trends and trajectories and ways to reduce disability, maximize independent functioning, and enhance quality of life at older ages [1].

Content areas include: the physical, social, technological and service environment; tests and self-reports of physical and cognitive capacity; use of assistive devices and rehabilitation services; help received with daily (self-care, mobility, household, and medical) activities; participation in valued activities; and wellbeing. Other topics focus on chronic conditions, symptoms, sensory impairments, transportation, subjective and economic wellbeing, and demographic factors. A last month of life interview focuses on quality of end-of-life care and a facility interview is conducted for those living in residential care settings. Links to Medicare records are also available. Caregivers of NHATS participants are interviewed occasionally in the companion National Study of Caregiving.

NHATS is led by the Johns Hopkins Bloomberg School of Public Health and the University of Michigan's Institute for Social Research, with data collection by Westat. Support for NHATS is provided by the National Institute on Aging (U01AG032947).

NHATS is designed to provide a nationally representative cross-section of the Medicare population ages 65 and older at regular intervals. Oversamples by age and for Black non-Hispanic persons are embedded in the sample design [2]. In Round 5 (2015), a new sample was introduced to restore the sample to original size by age and race groups [3].

A dried blood spot (DBS) collection in Round 7 (2017) provided the biological material for genotyping. All participants in that round with a completed Sample Person interview were considered eligible for the dried blood spot (DBS) collection. However, self-response was required for the DBS consent process so a small percentage who had proxy respondents, although eligible, were not invited to participate in the DBS collection; 4,903 (93.1%) provided consent to collect.

In all, 4,691 persons (95.7%) of those who consented had at least 1 card with DBS sample available for assaying [4]. After several assays were conducted, 4,091 NHATS participants had genetic material available and had given permission for genotyping to be conducted.

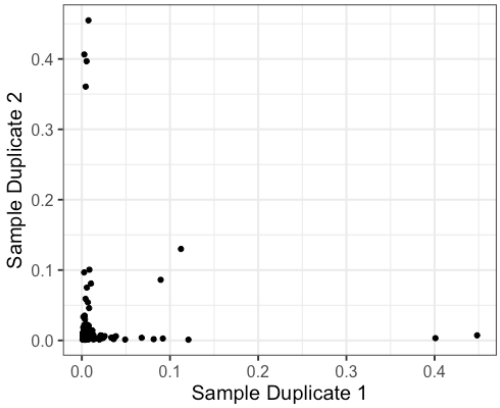
III. Genotyping process

The 4,091 samples were genotyped at Erasmus Medical Center in Rotterdam, Netherlands. Samples were genotyped on the Illumina Infinium Global Screening Array v3.0. The array contains clinical and rare variants ideal for multiethnic populations. Additional information is provided through the Erasmus MC Human Genomics Facility HUGE-F website: http://glimdna.org/global-screening-array.html. In total 725,831 SNPs are included in the dataset. Internal quality control (QC) methods for SNPs are shown in Table 1 and described in the section on *SNPs with high missingness and samples with low call rates*.

IV. Duplicate samples

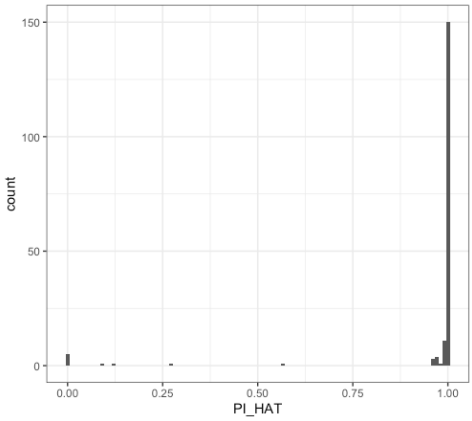
Of the 4,091 samples, 178 were genotyped twice (duplicate samples) for a total of N=4,269 samples in this study. Concordance rate for duplicates was 0.98. Duplicated samples were assessed for missingness, missing SNP genotype data (**Figure 1**). Among the duplicate pairs, samples with higher missingness rates were removed from further analysis. Most samples duplicate pairs have close to 100% shared identity (**Figure 2**) as determined by identity-by-descent analysis. Nine sample duplicate pairs had shared identity of less than 60%, as reported in the Duplicate Sample QC file.

Figure 1. Missing rates between duplicates, n=178 samples



Scatterplot of the 178 duplicate samples in the dataset and their data missingness (missing genotype data) rate.

Figure 2. Proportion identity-by-descent in duplicate samples, n=178



Proportion identity-by-descent (PI_HAT) value for sample duplicates.

V. Quality control process and participants

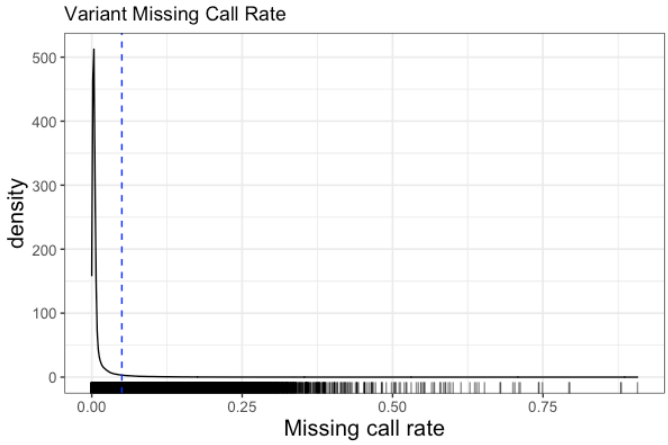
Quality control was performed at the Arking Lab at the Johns Hopkins University and validated independently at the University of Michigan. Analysis programs used to generate the results presented here include PLINK v1.9 [5] and R packages ggplot2 [6] for visualization. Any additional tools used for analysis are described in their relevant sections.

The full dataset is comprised of 4,269 genotyped samples. There are 178 duplicate samples and nine samples with poor duplicate concordance that were removed from the dataset (see section IV. Duplicate samples) prior to the quality control steps described below resulting in 4,082 non-duplicate samples that moved through the quality control process.

VI. SNPs with high missingness and samples with low call rates

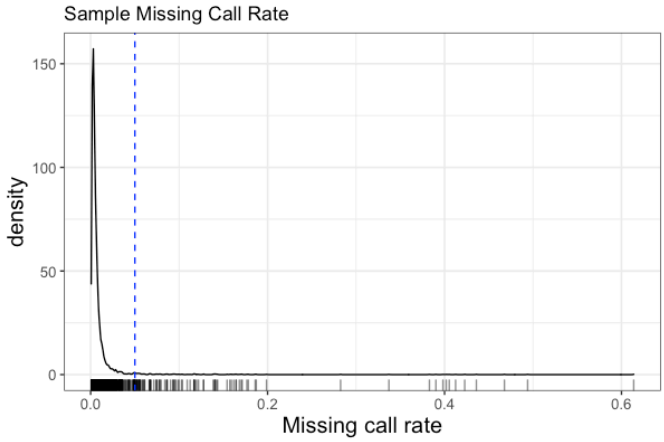
Individual *variants* with high rate of missing data are potentially low quality and should be filtered out. Similarly, *samples* with high rates of missing data are indicative of poor quality DNA and/or assay failure and should be removed. PLINK was used to identify low quality variants by calculating their missing call rate. Erasmus MC Human Genomics Facility returned 725,831 variants. Of those variants, 25,822 (3.6%) were missing from 5% or more of samples (**Figure 3**). These variants were excluded. After removing poor quality variants, samples that were missing 5% or more variant genotype data were also excluded. There were 76 (2%) samples with 5% or more missing variant data (**Figure 4**). This left a total of 700,009 variants and 4,006 samples. Of these 4,006 samples, self-reported reported primary race/ethnicity with missing values assigned the modal category indicated 729 non-Hispanic Black, 2,962 non-Hispanic White, 223 Hispanic, and 92 other race/ethnicity samples. For detail about each self-reported race/ethnicity group see the NHATS User Guide [https://nhats.org/researcher/nhats/methods-documentation?id=user_guide].

Figure 3. Variant missing call rates, n_{SNPs}=725,831



Each vertical line along the x-axis is a variant (n_{SNPs} = 725,831). Variants missing from 5% or more samples were excluded from further analysis. Blue dashed line indicates 5% cutoff.

Figure 4. Sample missing call rate, N=4,082

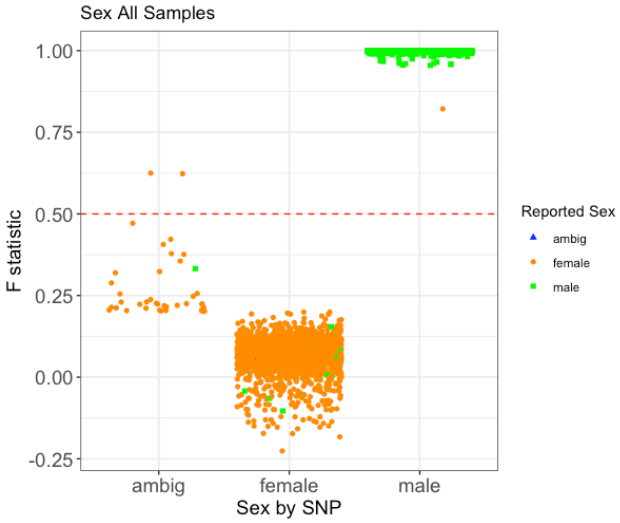


Each vertical line along the x-axis is one sample (N = 4,082). Samples missing 5% or more variants were excluded from further analysis (n = 76). Blue dashed line indicates 5% cutoff.

VII. Sex check

Sex checks were performed to test that the reported survey sex and the genotype determined sex agree using the PLINK command 'check-sex'. All samples with genotype data and call rates above 95% (n = 4,006) are included (**Figure 5**). We identified 57 samples with errors in their reported sex. Those samples are flagged as "Sex_mismatch" in the NHATS sample filtering document.

Figure 5. Sex check F statistic by reported sex, N=4,006



Sex F-statistic for each sample is plotted. Samples are group by their genotype-determined sex and colored by their reported sex. Red line shows 0.5 cutoff for females (< 0.5) and males (>0.5).

VIII. Minor allele frequency and Hardy-Weinberg equilibrium

We calculated variant minor allele frequency within each self-reported race/ethnicity group and removed variants with minor allele frequencies less than 5% within each self-reported race/ethnicity group for quality control analyses. To identify variants that are distributed as expected (e.g., in Hardy-Weinberg equilibrium) in each population, we calculated Hardy-Weinberg Equilibrium p-values for each self-reported race/ethnicity group. After removing samples with discordant sex, our sample consisted of 2,934 non-Hispanic White, 724 non-Hispanic Black, 211 Hispanic, and 80 identify as other individuals. The number of variants dropped due to minor allele frequency less than 5% and Hardy-Weinberg disequilibrium within each race/ethnic group can be seen in **Table 1**.

Table 1. National Health and Aging Trends Study genetic sample single nucleotide polymorphism filtering

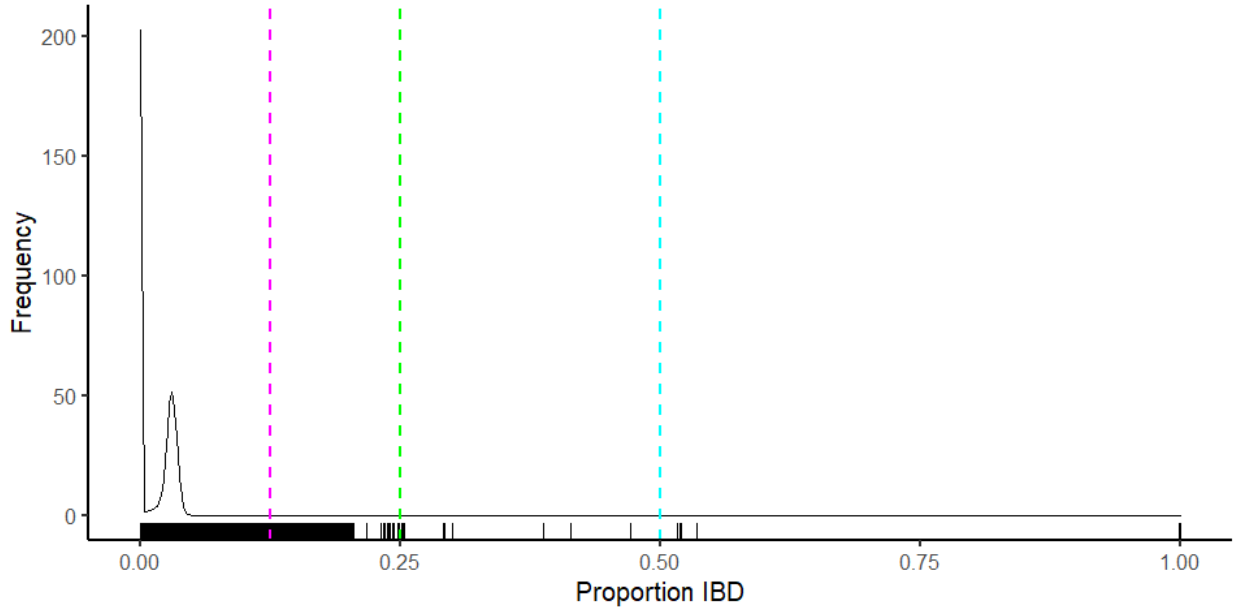
SNP filtering by self-reported race/ethnicity		N	SNPs lost to MAF <5%	SNPs lost due to HWE P<0.0001	SNPs retained
SNPs retained for quality control analysis and relatedness analysis, by race/ethnicity	Non-Hispanic Black	724	376,506	1474	322,029
	Non-Hispanic White	2,934	411,463	1131	287,415
	Hispanic	211	381,041	623	318,345
	Other	80	371,349	717	327,943

SNP: single nucleotide polymorphism; MAF: minor allele frequency; HWE: Hardy-Weinberg Equilibrium

XI. Relatedness check

Relatedness checks were performed to identify cryptic relatedness among participants. We removed the 57 sex mismatches and 76 individuals with high missing call rates. An initial pass to identify identical samples and first-degree relatives (pi-hat > 0.35) across the entire sample was performed using variants in Hardy-Weinberg Equilibrium (P>0.0001) and minor allele frequency >5% in the set of 3,949 samples (103,566 variants). Additionally variants in high linkage disequilibrium regions (HLA) on chromosome 6 and inversion areas on chromosome 8 and 17 were excluded from analyses. In this first pass, a total of 36 (22 unexpected identical samples in 11 pairs, and 14 unexpected first-degree samples) samples were removed. These samples are denoted in the NHATS sample filtering document.

Figure 6. Proportion identity-by-descent for pairs of samples in National Health and Aging Trends Study, N=3,913.



Each line along the x-axis represents a pairwise comparison between samples. Relatedness cutoffs are shown with vertical lines. Magenta marks 0.125 cutoff (3rd degree relatives), green marks 0.2 cutoff (2nd degree relatives), and cyan marks 0.5 cutoff (1st degree relatives).

Relatedness analyses were re-run on a total of 3,913 samples, stratified by self-reported race/ethnicity, with variants selected as above. Variants were pruned within self-reported race/ethnicity group to a subset that are in approximate linkage equilibrium with a window size of 50, a SNP window shift of 5, and a variance inflation threshold of 2. A total of 168,625 variants in 718 non-Hispanic Black participants; 92,481 variants in 2,907 non-Hispanic white participants; 109,492 variants in 210 Hispanic participants; and 104,710 variants in 78 participants of "other" race/ethnicity were used in relatedness calculations.

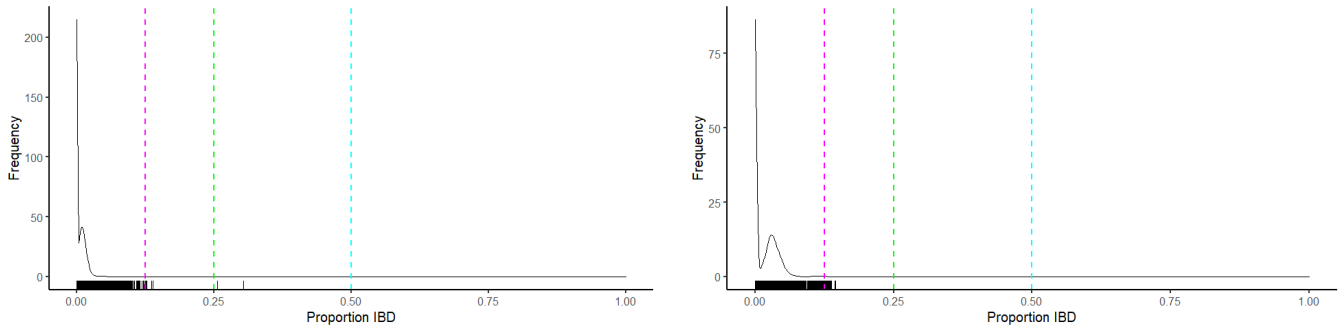
Relatedness between samples was assessed by plotting the overall identity-by-descent proportion, or PI_HAT, in a pairwise manner, stratified by self-reported race/ethnicity (**Figure 7**). We additionally examined the fraction of shared alleles between pairs (**Figure 8**). Sample pairs with a PI_HAT between 0.2 and 0.35 were flagged as "Relatedness_SecondDegree" in the NHATS sample filtering document. Two non-Hispanic black samples and 6 non-Hispanic white samples were flagged as unexpected second-degree relatedness. Sample pairs with a PI_HAT between 0.125 and 0.2 were flagged as "Relatedness_ThirdDegree" in the NHATS sample filtering document. Nine non-Hispanic black samples, 24 non-Hispanic white samples, and 31 Hispanic sample were flagged as unexpected third-degree relatedness.

After removing all individuals with unexpected relatedness, 3,839 samples ($n_{\text{non-Hispanic black}}$=705; $n_{\text{non-Hispanic white}}$=2,877; n_{Hispanic}=179; n_{other}=78) were included in genetic principal component analysis to identify non-Hispanic black/African genetic ancestry and non-Hispanic white/European ancestry analytic samples. A full accounting of samples lost due to unexpectedly high genetic relatedness can be seen in **Table 2**.

Figure 7. Proportion identity-by-descent for pairs of samples stratified by National Health and Aging Trends Study self-reported race/ethnicity.

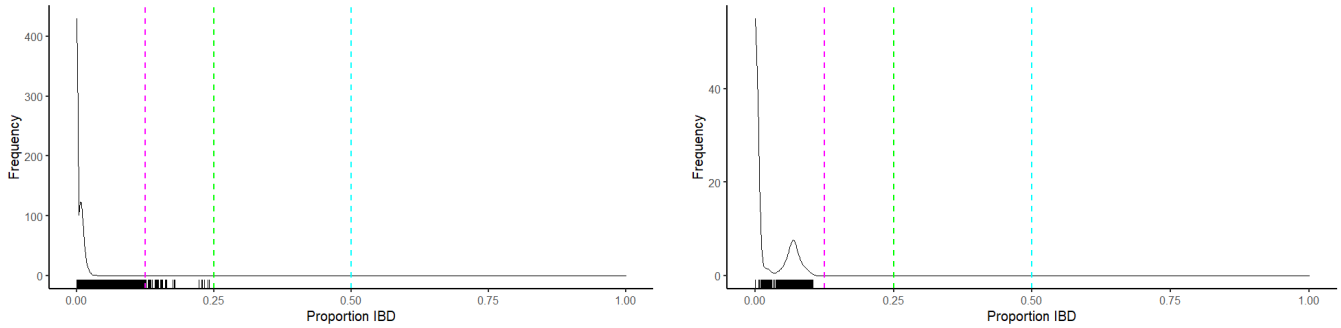
A. Non-Hispanic Black, n=718

C. Hispanic, n=210



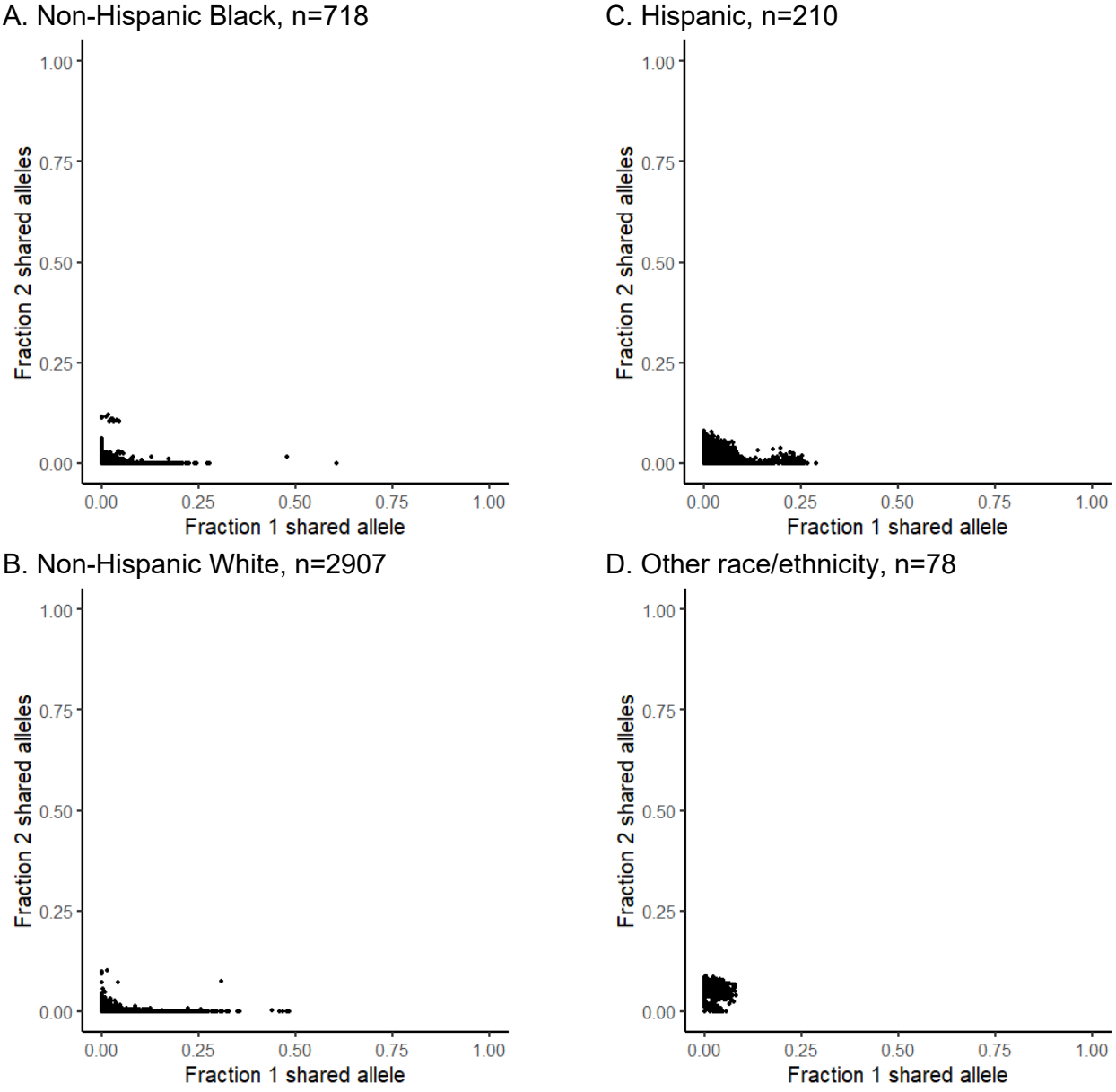
B. Non-Hispanic White, n=2,907

D. Other race/ethnicity, n=78



Each line along the x-axis represents a pairwise comparison between samples. Relatedness cutoffs are shown with vertical lines. Magenta marks 0.125 cutoff (3rd degree relatives), green marks 0.2 cutoff (2nd degree relatives), and cyan marks 0.5 cutoff (1st degree relatives).

Figure 8. Proportion of one or two shared alleles for pairs of samples, stratified by National Health and Aging Trends Study self-reported race/ethnicity.



Proportions of one and two shared alleles are plotted on the X and Y axes using the Z1 and Z2 metrics from PLINK's 'genome' function.

Table 2. Genetic relatedness analysis National Health and Aging Trends Study

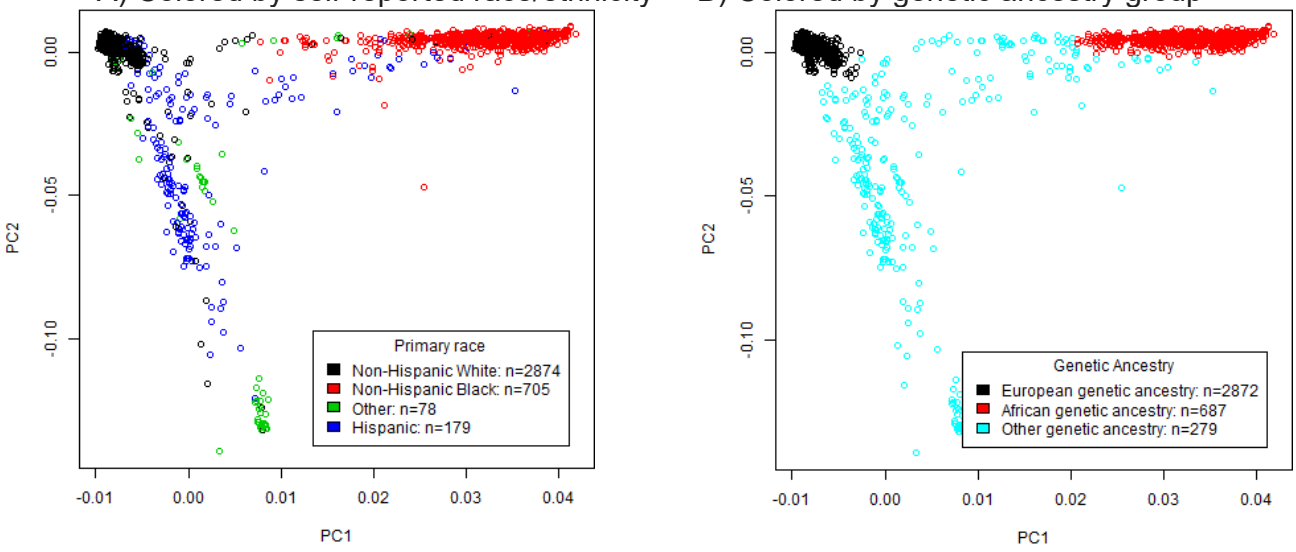
	Non-Hispanic black	Non-Hispanic white	Hispanic	Other	**Total kept**	**Total flagged**
Relatedness analysis sample	724	2,934	211	80	3,949	-
After identical and first-degree samples removed,	718	2,907	210	78	3,913	36
After second-degree samples removed	716	2,901	210	78	3,905	8
After third-degree samples removed	705	2,876	179	78	3,838	67
Total removed for unexpected relatedness	19	58	32	2	-	110

X. Population structure and homogenous analytic groups

We calculated global genetic principal components on 3,839 samples after the removal of samples flagged for low call rate (<95%), sex mismatches, and unexpected relatedness. Variants with Hardy-Weinberg Equilibrium test p-values greater than 0.0001 and minor allele frequency greater than 5% in the sample of 3,839 were included in population structure analysis. Variants from a list of known high linkage disequilibrium (LD) regions [7] were first removed. Additional LD pruning was performed using Plink (--indep-pairwise 50 5 0.2). This final pruned set of variants (n_{SNPs}=102,939) were then used to perform principal component analysis on 3,838 samples (**Figure 9, panel A**). We color the global genetic principal component plot by self-identified race/ethnicity. Relatively homogenous genetic ancestry groups were defined as: European genetic ancestry as having a PC1 value ± 2 standard deviations from the mean PC1 and a PC2 value ± 2 standard deviations from the mean PC2 for all those reporting non-Hispanic white as their race/ethnicity (-0.014≤PC1≤-0.003 and -0.011≤PC2≤0.016); African genetic ancestry is defined as having a PC1 value ± 2 standard deviations from the mean PC1 and a PC2 value ± 2 standard deviations from the mean PC2 for all those reporting non-Hispanic black as their race/ethnicity (0.021≤PC1≤0.044 and -0.002≤PC2≤0.010). All individuals who fall outside those defined ranges are not included in the genetically defined European or African ancestry groups (**Figure 9, panel B**).

Figure 9. Scatter plot of the first two global genetic principal components

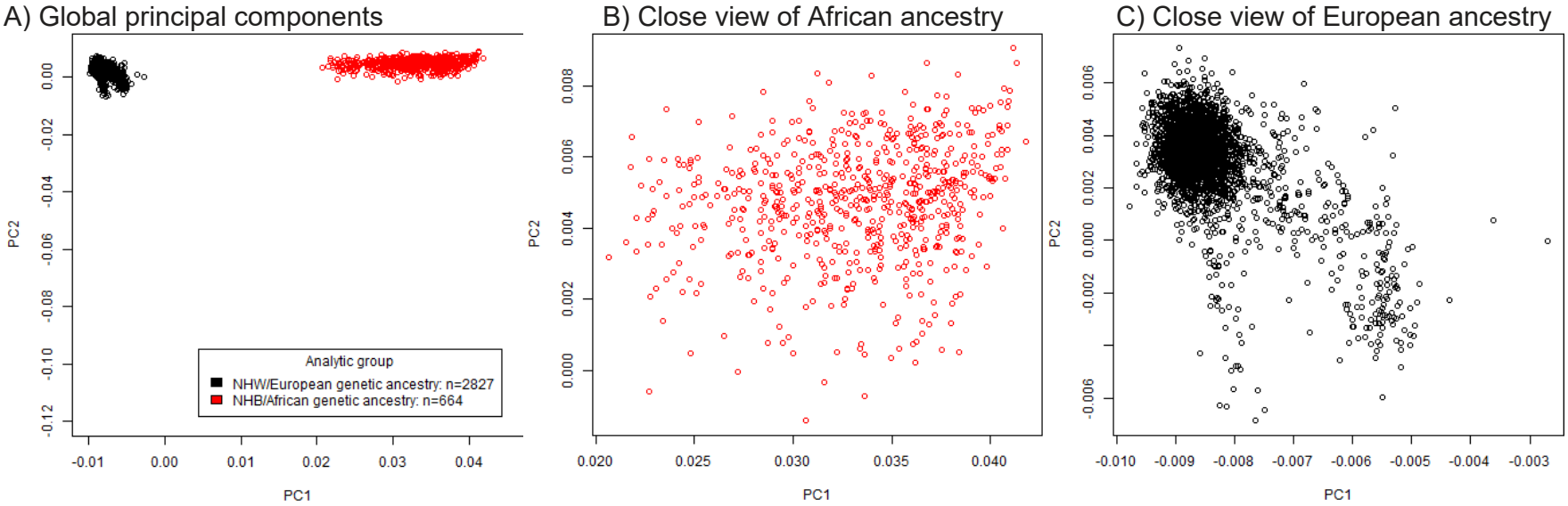
A) Colored by self-reported race/ethnicity B) Colored by genetic ancestry group



Points represent sample coordinates on global genetic principal component 1 and 2 and are colored by A) self-reported race/ethnicity and B) genetic ancestry. European genetic ancestry is defined as having a PC1 value ± 2 standard deviations from the mean PC1 and a PC2 value ± 2 standard deviations from the mean PC2 for all those reporting non-Hispanic white as their race/ethnicity; while African genetic ancestry is defined as having a PC1 value ± 2 standard deviations from the mean PC1 and a PC2 value ± 2 standard deviations from the mean PC2 for all those reporting non-Hispanic black as their race/ethnicity. All individuals who fall outside those defined ranges are not included in the analytic European or African genetic ancestry.

To identify analytic groups, we take the intersection of those samples in the European genetic ancestry group and those who reported race/ethnicity as non-Hispanic White (n=2827) and separately the union of those in the African genetic ancestry group who reported race/ethnicity as non-Hispanic Black (n=664) (**Figure 10)**. These two groups are defined as “European” and “African” in the NHATS sample filtering document. We recommend analyzing these two groups separately. The decision flow chart to identify the analytic samples can be seen in **Figure 11**.

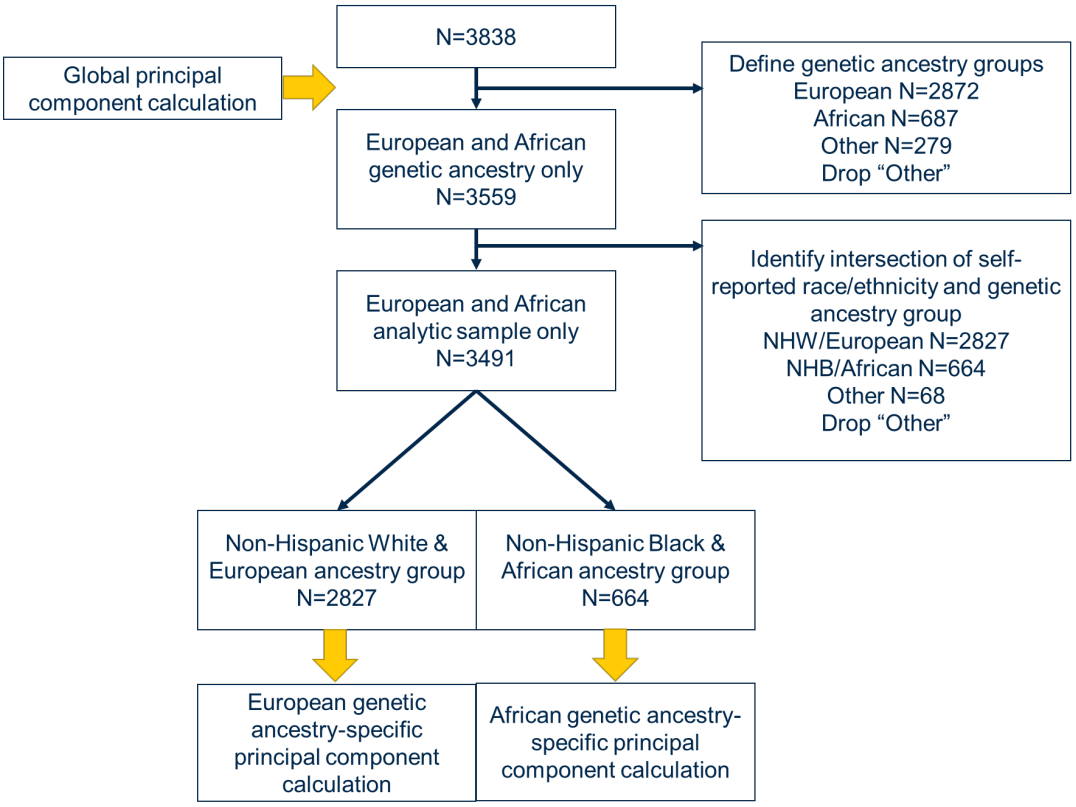
Figure 10. Scatter plot of the first two global genetic principal components for the analytic samples



NHW: Non-Hispanic White, NHB: Non-Hispanic Black

Panel A represents the analytic groups on the scale seen in the previous figure. Panel B is a close view of the Non-Hispanic Black/African genetic ancestry cluster (n=664). Panel C is a close view of the Non-Hispanic White/European genetic ancestry cluster (n=2,827). Points represent sample coordinates on global genetic principal component 1 and 2 and are colored by analytic sample (black: Non-Hispanic White and European genetic ancestry, red: Non-Hispanic Black and African ancestry)

Figure 11. Decision flow chart to identify genetic analytic samples, National Health and Aging Trends Study



When performing analyses, we recommend doing so separately by analytic group (African, European) and controlling for local genetic principal components (those calculated within analytic group). In the African analytic group we recommend adjusting for at least two genetic principal components (**Figure 12**) in analyses and at least the first five European analytic group genetic principal components in the European analytic group (**Figure 13**). An overall summary of the quality control steps can be found in **Figure 14**.

Figure 12. Scatter plot matrix of the first five genetic principal components calculated within the analytic non-Hispanic Black/African genetic ancestry sample in the National Health and Aging Trends Study, n=664.

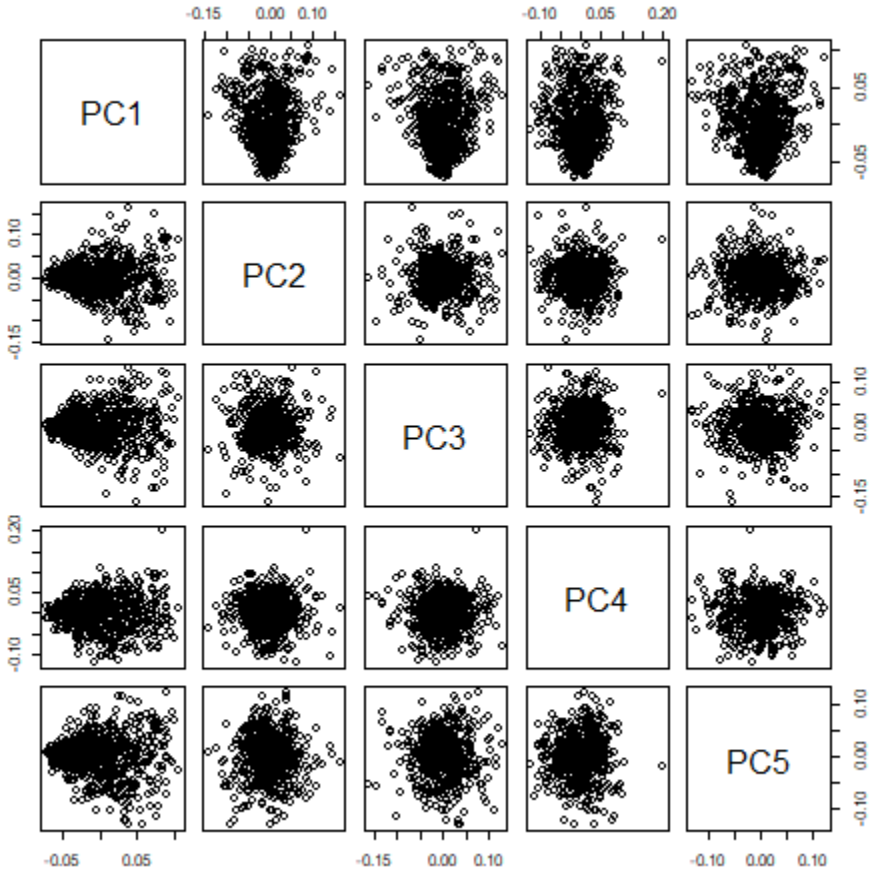


Figure 13. Scatter plot matrix of the first seven genetic principal components calculated within the analytic non-Hispanic White/European genetic ancestry sample in the National Health and Aging Trends Study, n=2827.

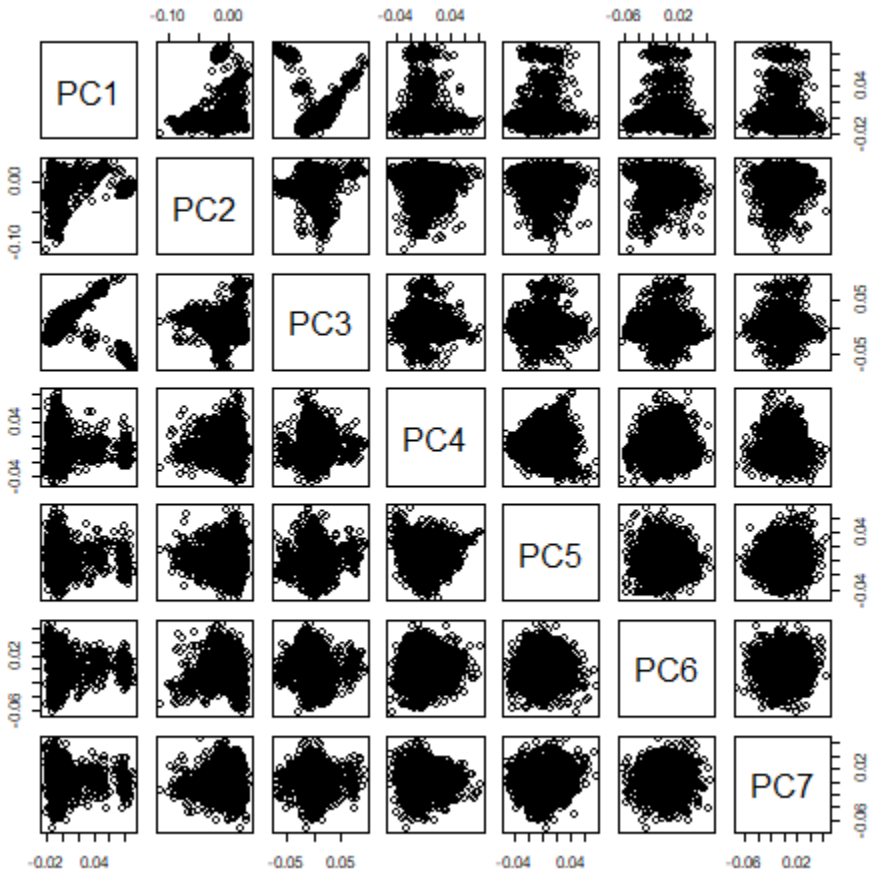
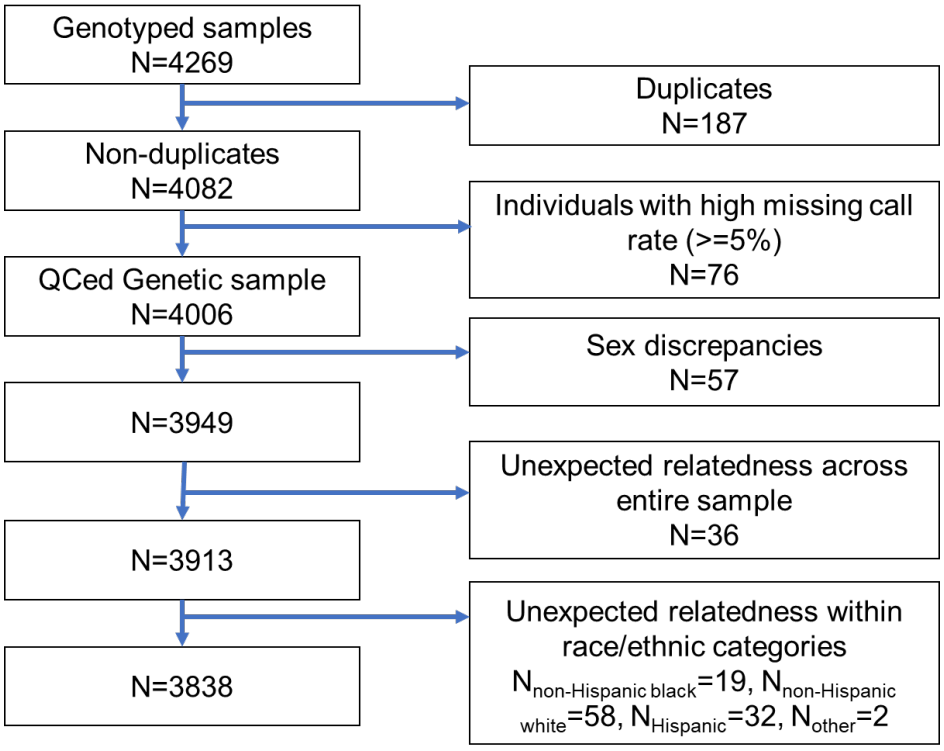


Figure 14. Summary of the quality control process steps, National Health and Aging Trends Study genetic sample



XI. TOPMed Imputation

Genotype imputation is the process of inferring unobserved genotypes in a study sample based on the haplotypes observed in a more densely genotyped reference sample [8,9]. Imputation was performed on the sample of 4,006 participants (729 non-Hispanic blacks, 2,962 non-Hispanic white, 223 Hispanic, and 92 other race/ethnicity). The Trans-Omics for Precision Medicine (TOPMed) Imputation Reference panel is a diverse reference panel including information from 97,256 deeply sequenced human genomes. The panel is available to the community through a collaboration between TOPMed Study Investigators, the National Heart Lung and Blood Institute and the University of Michigan Imputation Server team [10, 11]. Observed genotypes (which have a probability of 1) are included in the imputation output. Where an observed study SNP had sporadic missing data, the missing genotypes were imputed in the same manner as the completely unobserved SNPs and should be treated with the same caveats. Additionally, SNPs genotyped in the study but failing pre-imputation quality filters may also appear in imputed results, when available in the reference panel.

The final pruned set of SNPs used for imputation are in Table 3. This final set was derived after filtering out samples and SNPs with 5% or more missingness, Hardy-Weinberg disequilibrium and minor allele frequency less than 5%. Prior to imputation, we used the HRC or 1000G Imputation preparation and checking developed by Will Rayner to check input variant data for accuracy relative to expected TOPMed SNPs. This process identifies errors in the 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. Problematic SNPs are excluded and aligned SNPs are flipped to the + strand, producing the final SNP set to be used for imputation. The tool and detailed procedures are available here: https://www.well.ox.ac.uk/~wrayner/tools/.

Table 3. Number of SNPs Excluded or Strand-Flip Prior to Imputation

	Excluded SNPs	Strand-Flip SNPs
Non-Hispanic Black	91,954	34,957
Non-Hispanic White	11,932	34,028
Hispanic	17,891	35,375
Other	27,144	35,478

Table 4. TOPMed Imputation Server Parameters

Parameter	TOPMed Server Setting
Genome Build	Hg19 liftover to hg38
Include ChrX	yes
Reference Panel	TOPMed Reference Panel
Phasing Algorithm	Eagle v2.4
r2 value threshold	none

Table 5. Number of additional SNPs Excluded by TOPMed Imputation Server

	TOPMed Server Excluded SNPs
Non-Hispanic Black	407
Non-Hispanic White	256
Hispanic	338
Other	373

The final pruned set of variants from Section XI was used for the imputation. Pre-phasing and imputation was performed stratified by self-reported race/ethnicity on the TOPMed Imputation Server Cloud based web server: https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home [10, 11, 12]. The TOPMed reference panel consists of 308,107,085 genetic variants across 22 autosomes and chromosome X. The data was stratified by race and genotyped SNPs were submitted as 23 separate VCF files, one for each chromosome (1-22, X). The genotype data is in human genome build hg19 and were updated (liftOver) to genome build hg38, to be compatible with the TOPMed reference panel. The server performs additional QC checks (as described here https://topmedimpute.readthedocs.io/en/latest/pipeline.html) and further excluded SNPs if, for example, the liftOver failed, reference allele does not match panel reference allele, or the SNP call rate is less than 90%. A file containing this list of excluded SNPs is also available. Phasing was performed by Eagle v2.4 algorithm. The liftOver and phasing steps were performed as part of the TOPMed Imputation Server program. The TOPMed Imputation server splits chr X into three regions, pseudoautosomal regions 1, non-pseudoautosomal region, and pseudoautosomal region 2 for phasing and imputation. No r2 value threshold for imputation quality was implemented. Output SNPs are provided in VCF format. Note genotype SNPs are denoted as “TYPED” in the INFO field. The VCF FORMAT field contains the following genotype information in the following format “GT:DS:HDS:GP” where GT is Genotype, DS is Estimated Alternate Allele Dosage, HDS is Estimated Haploid Alternate Allele Dosage, and GP is Estimated Posterior Probabilities for Genotypes 0/0, 0/1, and 1/1. Results are provided for each chromosome separately and by race/ethnic group. The recommended sample filtering document can be used to subset to different analytic groups.

References

[1] Freedman VA, Kasper JD. Cohort Profile: The National Health and Aging Trends Study (NHATS). Int J Epidemiol. 2019;48(4):1044-1045g. doi:10.1093/ije/dyz109

[2] Montaquila J, Freedman VA, Edwards B, Kasper JD. *National Health and Aging Trends Study Round 1 Sample Design and Selection. NHATS Technical Paper #1*. Baltimore, MD: Johns Hopkins University School of Public Health, 2012

[3] DeMatteis J, Freedman VA, Kasper JD. National Health and Aging Trends Study Round 5 Sample Design and Selection. NHATS Technical Paper #16. Baltimore, MD: Johns Hopkins University School of Public Health, 2016

[4] Kasper JD, Skehan ME, Seeman T, Freedman VA. Dried Blood Spot (DBS) Based Biomarkers in the National Health and Aging Trends Study User Guide: Final Release. Baltimore, MD: Johns Hopkins University Bloomberg School of Public Health, 2019

[5] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

[6] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

[7] Price AL, Weale ME, Patterson N, et al. Long-range LD can confound genome scans in admixed populations. Am J Hum Genet. 2008;83(1):132-139. doi:10.1016/j.ajhg.2008.06.005. https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)

[8] Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 2008;124(5):439-450. doi:10.1007/s00439-008-0568-7

[9] Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406. doi:10.1146/annurev.genom.9.081307.164242

[10] Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-299. doi:10.1038/s41586-021-03205-y

[11] Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-1287. doi:10.1038/ng.3656

[12] Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2015;31(5):782-784. doi:10.1093/bioinformatics/btu704