Centre intégré universitaire de santé et de services sociaux de l'Estrie – Centre hospitalier universitaire de Sherbrooke
Québec

Centre de recherche sur le vieillissement

NuAge Database and Biobank
1036 Belvédère St. South
Sherbrooke, Quebec, J1H 4C4
819-780-2220, ext.: 45613
NuAge-cdrv@usherbrooke.ca

NuAge

GENOTYPING QUALITY CONTROL REPORT FOR NUAGE PARTICIPANTS *(FINAL REPORT)*

TABLE OF CONTENTS

December 5, 2023

1. CONTEXT

In 2016-2017, fruitful discussions have taken place between Professor Mark Lathrop, director of the McGill University Genome Center, and Pierrette Gaudreau, Professor at the department of Medicine of Université de Montréal, director of the Quebec Network for Research on aging (RQRV), one of the five founding members of the NuAge cohort study, and director of the NuAge biobank at the *Centre hospitalier de l'Université de Montréal* (CHUM). These discussions led to a scientific and financial partnership between the McGill Genome Center, the CHUM and the principal investigators of the NuAge study. Steps were subsequently taken by the NuAge Steering Committee and the RQRV to develop a collaborative framework between NuAge, Professor Mark Lathrop and Professor Jiannis Ragoussis, Head of Genome Sciences at the McGill University Genome Center, to obtain genotyping data from NuAge participants. This initiative was undertaken as a partnership to enrich the NuAge database (genotypes) and biobank (DNA, RNA) and was therefore not a secondary research project as normally seen when access to NuAge samples are requested.

The collaborative agreement came into effect in April 2017 and stipulated that the NuAge biobank should send peripheral lymphocytes from NuAge participants to the McGill Genome Center (Professor Ragoussis' laboratory) to extract DNA and RNA and carry out genotyping using the Affymetrix UK Biobank Axiom™ array. The costs of sample retrieval from freezers, preparation for transportation, forms completion and transport between the NuAge biobank (located at the CHUM Research Center) and McGill were covered by funds from the RQRV. Costs for DNA and RNA extraction, as well as for chip genotyping were graciously covered by discretionary funds from Professor Mark Lathrop.

2. OBJECTIVES

The main objective of this report is to describe the steps for obtaining good quality and scientifically-relevant genotyping data for the NuAge Database and Biobank participants who agreed to participate in genetic studies.

This objective can be separated into five sub-objectives aiming to describe:

A) The usage of NuAge biological samples for DNA extraction and genome-wide genotyping with the Affymetrix UK Biobank Axiom™ array;
B) Additional sample-based and marker-based genotyping quality control (QC) steps;
C) The imputation and QC steps of additional genetic markers to increase genome coverage;
D) The genotyping of two SNPs in the *APOE* gene defining the ε2, ε3 and ε4 alleles using the TaqMan RT-PCR technology;
E) The final preparation and availability of genotyping datasets for secondary studies.

3. METHODOLOGY

3.1. Study participants

Overall, 1,405 of the 1,753 NuAge participants have agreed to participate in genetic studies (see Attachment A) and have been considered in this initiative. When biological samples, DNA or RNA extracts were available for participants who refused to participate in genetic studies at the time of

the analyzes (2017-2019), they were definitively destroyed according to the procedures described in the NuAge Database and Biobank Management Framework. For genetic data obtained from excluded participants, they were removed from all files used for quality control and for the preparation of the final genotype files integrated in the NuAge Database and Biobank.

3.2. NuAge data

Only two variables were extracted from the NuAge database to complete QC steps. These variables are self-reported sex (men, women) and self-reported race (Caucasian, Asian, Black, Hispanic, and Metis).

3.3. Sub-objective A) Biological samples and genotyping[1]

3.3.1. Sample collection and storage

Fresh blood samples from overnight-fasted participants were collected into a 10 ml sodium heparin tube following an overnight fast during the NuAge study (2003-2008; annual collection: T1, T2, T3, T4). Peripheral blood mononuclear cells (PBMCs) were isolated from fresh blood (≤2 hours) by transferring 6 ml of fresh blood into an Accuspin System-Histopaque-1077 tube (Sigma) and centrifuged at 1000 x g for 20 min at 25°C (IEC Centra MP4R Centrifuge). After removal of the plasma layer, PBMCs (about 1 to 1.5 mL) were collected and transferred into a 15 ml sterile tube, washed once with 10 mL of PBS RNAse free 1X (Ambion) and a second time with 5 mL, and centrifuged at 360 x g during 15 min at 25°C after each washing. The PBMC pellet was then treated with 1 ml TRIzol (Biobar Invitrogen), resuspended (pipet), lightly homogenized (one tissue homogenizer per participant) and then aliquoted into RNase free barcoded cryogenic tubes (UltiDent) kept on ice and then at -20°C for a short period. Samples were then rapidly stored at -80°C until their use for RNA and DNA extraction and genotyping.

3.3.2. DNA and RNA extraction and storage

One PBMC sample per participant was selected at follow-up T2, T3 or T4. All samples were shipped on dry ice to the laboratory of Dr. Jiannis Ragoussis (McGill Genome Centre, McGill University, Montréal, Canada) in 2018 and kept at -80°C until RNA and genomic DNA extraction. A total of 200 µl chloroform was added to PBMC samples, inverted and centrifuged at 12,000 x g at 4°C for 15 minutes. The upper aqueous phase containing the RNA was transferred into a 2.0 ml deep 96 well plate and used for RNA extraction using the protocol “Chemagic RNA Tissue 96 prefilling drying MATRIX VD110916” on the Chemagen instrument (Perkin-Elmer cat# CMG-1212). RNA concentration and integrity (RIN) was obtained for each RNA samples obtained using the LabChip GX Touch nucleic acid analyser from Perkin-Elmer (Reagent kit cat# CLS960010). The remaining interphase and lower organic (phenol-chloroform) phase obtained after previous centrifugation were used for manual DNA extraction. To do so, 300 µl 100 % ethanol was added to the remaining phases, and then inverted, incubated 5 minutes on ice, and centrifuged at

[1] Based on CLSA (chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fwww.clsa-elcv.ca%2Fdoc%2F2748, accessed December 21, 2021) and UK Biobank report (chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fbiobank.ctsu.ox.ac.uk%2Fcrystal%2Fcrystal%2Fdocs%2Fgenotyping_qc.pdf&clen=7603821&chunk=true, accessed December 21, 2021).

4,000 x g at 4°C for 5 minutes. The supernatant (organic phase) was discarded and the pellet resuspended in 800 µl of 0.1 M sodium citrate / 10 % ethanol (1.47 g sodium citrate + 5 ml anhydrous ethanol + Axiom™ H_2O to 50 ml) and incubated 20 minutes at room temperature with constant agitation (800 rpm). Samples were then centrifuged at 4,000 x g at 4°C for 5 minutes, the supernatant discarded and pellet resuspended with 1 ml 75 % ethanol and incubated 5 minutes at room temperature, following final centrifugation at 4,000 x g at 4°C for 5 minutes. The supernatant was discarded and the pellet air-dried for 5 minutes. Single-stranded DNA (ssDNA) samples were then eluted in 300 µl of 8 mM NaOH pH 11. Concentration and purity (260 nm / 280 nm ratio) were obtained by absorbance assay for each ssDNA sample using TECAN Spark 10M Microplate reader. RNA and ssDNA extracts were aliquoted in barcoded tubes and stored at -20°C until genotyping and shipment to the NuAge biobank. All RNA, remaining ssDNA (not used for genotyping) and remaining PBMC samples were shipped on dry ice to the NuAge biobank for long term storage at -80°C or destruction if participants needed to be excluded. Data on the concentration and purity/integrity of ssDNA and RNA samples provided by McGill are kept by the NuAge biobank and can be provided to secondary studies aiming to use these samples.

Among the 1,500 PBMC samples sent by the NuAge biobank, 1,499 successfully provided RNA and ssDNA extracts. The extracts from 183 samples were finally destroyed by the biobank (February 2022) because participants were excluded from the NuAge Database and Biobank or did not consent to participate in genetic studies (see consents in Attachment A). It thus remained 1,316 samples with RNA and ssDNA extracts (including replicates), which correspond to **1,303 unique NuAge participant RNA and ssDNA extracts that can be used in genetic studies.**

3.3.3. Genotyping and calling

Among the 1,316 NuAge ssDNA samples, 1,312 **(1,299 unique NuAge participants) had a concentration of at least 10 ng/µl necessary for genotyping on the chip**. Each genotyping 96-wells plate contained NuAge ssDNA samples, two control DNA samples (NA24385, Caucasian male; CEPH control 1463-02, Caucasian female) and one negative control of deionized water. The standard Affymetrix protocol was applied by the laboratory of Dr. Ragoussis for ssDNA sample preparation, genome amplification, fragmentation, precipitation and re-suspension, and then for hybridization to the UK Biobank Axiom™ genotyping arrays (Thermo Fisher Catalog # 902502[2]). This genotyping array covers about 800,000 genetic variants and was designed to target known disease-associated and coding single nucleotide polymorphisms (SNPs). It also target a set of SNPs enabling good genome-wide imputation of additional SNPs with common (>5%) and relatively low (1-5%) minor allele frequency (MAF) in individuals of European ancestry. Hybridized plates were processed in a single batch on the Affymetrix Instrument and analyzed using the Affymetrix Axiom™ Analysis Suite software version 2.0. Following initial sample and plate quality control (QC) steps (Dish QC ≥ 0.82, QC call rate ≥ 95.0, percent of passing samples per plate ≥ 70.0, average call rate for passing samples per plate ≥ 95.0), genotype calling was performed for a total of 17 plates based on their clustering position in the signal intensity space

[2] See the content of this array here: https://www.thermofisher.com/order/catalog/product/000854?SID=srch-srp-000854.

(one dimension for each allele). A final round of genotype clustering and calling was done after re-genotyping samples and/or plates not passing initial QC filters.

There were a total of 1,539 samples successfully genotyped and identified "pass" in the "SampleReport_NuAge_20200302" file provided by McGill, which include unique NuAge samples, replicates and positive controls. Replicated samples (tech duplicates of the same sample ID and NuAge participant replicates with different sample ID) were detected using the PLINK software version 1.9 or following familial relationships inference using the KING software program (identified as "Dup/MZ"; see section 3.4.2). Independent validation was done for any replicate observed. Genetic sex (male, female, unknown) was determined using both the Affymetrix Axiom Analysis Suite algorithm and PLINK v1.9. A Support Vector Machine (SVM) model based on UK Biobank samples was used to correct the miscalling of males by Affymetrix. An empirical threshold was used to recall the sex of samples miscalled by PLINK (corrected) through setting X-chromosome F estimate < 0.3 as female, F estimate > 0.8 as male, and F estimates between 0.3-0.8 as unknown sex. Independent validation was done for any sex discordance observed. The genetic sex data and F estimates were provided by McGill in their Sample-based QC report file "SampleReport_NuAge_20200302". The genotype data files provided by McGill are in PLINK binary format, which are .bed (biallelic genotype table), .bim (variant information), and .fam (sample information) files. Original PLINK binary files provided by McGill contain data for all 1,539 successfully genotyped samples detailed above (identified as "pass") and for 722,976 unique genetic markers identified as "BestandRecommended" markers by the Axiom Analysis Suite (i.e. unique best probeset). More details on PLINK file formats are available here (https://www.cog-genomics.org/plink/1.9/formats).

Among the 1,539 successfully genotyped samples, there were a total of **1,276 unique NuAge participants who consented to participate in genetic studies and successfully genotyped at 722,976 unique genetic markers (no multi-allelic markers).** For further NuAge usage, new PLINK binary files (--keep --make-bed; PLINK v1.9) and a revised Sample-based QC report file ("SampleReport_NuAge_1276ID_finalv1.txt") were created by the NuAge team by keeping only the 1,276 unique NuAge participants and choosing the sample with the best call rate for each replicate. This step removed a total of 183 unique NuAge participants excluded from the NuAge Database and Biobank (since there creation in March 2019) or who did not consent to participate in genetic studies. Original files (Sample report, PLINK binary files) sent by McGill will not be used for further analyses by the NuAge team, but will be kept internally by NuAge and McGill's team for QC history purposes only.

3.3.4. Discrepancies between genetic sex calls and self-reported sex

The original Sample Report sent by McGill contains the sex determined genetically (genetic sex calling; section 3.3.3). The NuAge team analyzed the discrepancies observed between the PLINK corrected genetic sex calls obtained for the 1,276 eligible unique NuAge participants with their self-reported sex in the NuAge database. There were 596 men and 679 women with concordant sex. Only one participant self-reported as women and had an unknown genetic sex, but with an F estimate (0.3255) close to the threshold used to call sex as women (<0.3). Thus, self-reported sex was considered as a valid variable that can be used in all future studies using genotypes from the

1,276 NuAge participants (596 men and 680 women). Only self-reported sex is thus provided in the Sample-based QC report "SampleReport_NuAge_1276ID_finalv1.txt". Furthermore, the PLINK binary files were updated to include the self-reported sex in the third column of the .fam file (--update-sex <filename> –make-bed; PLINK v1.9).

All downstream analyses were done using the PLINK binary files restricted to the 1,276 unique NuAge participants who consented to genetic studies (with sex information).

3.4. Sub-objective B) Additional QC steps[3]

Additional basic QC steps were completed by McGill (laboratory of Dr. Ragoussis) and the NuAge team in order to "flag" genetic markers and samples which may be of lower quality and/or would need to be removed for further analyzes. These flagged markers and samples were not removed from the PLINK binary files, but can be excluded in a case-by-case basis. Information about the flagged markers and samples are available in the revised Marker-based QC report ("NuAge_AX001toAX017_markerQC_finalv1.txt") and Sample-based QC report ("SampleReport_NuAge_1276ID_finalv1.txt"). Details on the content of these reports are provided in Attachments B and C, respectively.

3.4.1. Marker-based QC

For the 722,976 unique genetic markers genotyped, three tests were realized to check for genotype consistency across different conditions. The statistics and flagged markers are provided in the revised Marker-based report. The two last tests are based on hypothesis testing. For instance, an adjusted *P*-value threshold was defined by dividing the UK Biobank single test *P*-value (0.005) by the number of tests realized; number of markers (722,976) * number of hypothesis testing tests (2) = 1,445,952 tests. The *P*-value threshold was set at 3.458×10^{-9}.

- Discordant genotype across control replicates: Each genotyping plate had two DNA positive controls (CEPH146302 and NA24385). It is thus expected that the genotypes obtained for each control in a plate is fully concordant with those obtained in the other plates. A concordance metric (*d*) was calculated and provided by McGill for each genetic marker and control sample. The NuAge team noticed that 6,349 markers had a control replicate discordance metric >0.05 (below 95% concordance) in at least one of the DNA positive control and were thus flagged in the revised Marker-based QC report file.

- Departure from Hardy-Weinberg equilibrium: Genotype frequencies at a given marker are normally in tight relation with its allele frequencies, called Hardy-Weinberg equilibrium (HWE). Observed genotype frequencies for a given marker may deviate from the expected genotype frequencies under certain conditions, such as inbreeding, population stratification (systematic ancestry differences in allele frequencies between strata [e.g. case-control study]), evidence for disease-association in affected individuals, and genotyping problems (Wigginton et al.

[3] Based on CLSA (https://www.clsa-elcv.ca/stay-informed/new-clsa/2018/clsa-releases-first-genomics-data, accessed January 2022) and UK Biobank report (chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fbiobank.ctsu.ox.ac.uk%2Fcrystal%2Fcrystal%2Fdocs%2Fgenotyping_qc.pdf&clen=7603821&chunk=true, accessed December 21, 2021).

2005). The Fisher's exact test (Wigginton et al. 2005) was applied by NuAge to test for departure from HWE for each genetic marker using PLINK (--hardy) and the revised PLINK binary files for the 1,276 included NuAge participants. Only diploid regions were analyzed (i.e., autosomes [chr1-22], pseudo-autosomal regions on chr-X, and females on the sex-specific region of chr-X). The NuAge team noticed 336 genetic markers with a HWE *P*-value $<3.458 \times 10^{-9}$ (departed from the expected genotype frequencies) and were thus flagged in the revised Marker-based QC report file.

Among the 722,976 unique markers genotyped, there were 715,462 remaining after the exclusion of markers flagged by one of the three QC tests (Table 1). Complementary filters were then applied to target markers that will be used in downstream analyzes, which left 515,077 markers. Markers filtered out at this step are also detailed in Table 1. **These 515,077 remaining markers are thus considered as having good quality genotyping data, and restricted to relatively frequent/common (MAF ≥1%) SNPs (no indel) located on autosomes (chr-1-22) and with a good SNP-wise call rate (missingness <0.01).** The 515,077 SNPs were then pruned to a set of 149,004 independent SNPs based on an r^2 <0.10 using PLINK (--indep-pairwise 5000kb 1 0.1). Markers flagged by the three QC steps, complementary filters and during the pruning process are identified in the revised Marker-based QC report file.

Table 1. Tabulation of the markers flagged by the three QC tests (in bold) and by complementary filtering

	Control	**HWE***	**Sex***	Mono	Indel**	Chr-X,Y,MT	MAF<1%	Miss≥1%
Control	**6,349**	**10**	**1**	31	29	50	312	1,766
HWE*		**336**	**0**	–	8	1	0	211
Sex*			**17**	–	1	16	0	8
Mono				39,833	n.a.	1,339	39,833	525
Indel					4,937	76	2,299	588
Chr-X,Y,MT						19,395	2,654	2,344
MAF<1%							107,420	2,876
Miss ≥1%								81,969

Indel, insertion or deletion; MAF, minor allele frequency; Miss, marker-wise missingness; Mono, monomorphic (minor allele count [MAC] = 0); MT, mitochondrial; n.a., not available.
*Apart from 823 markers outside diploid regions (chr-Y,MT; without statistics).
** Monomorphic indels were not included in the counts.

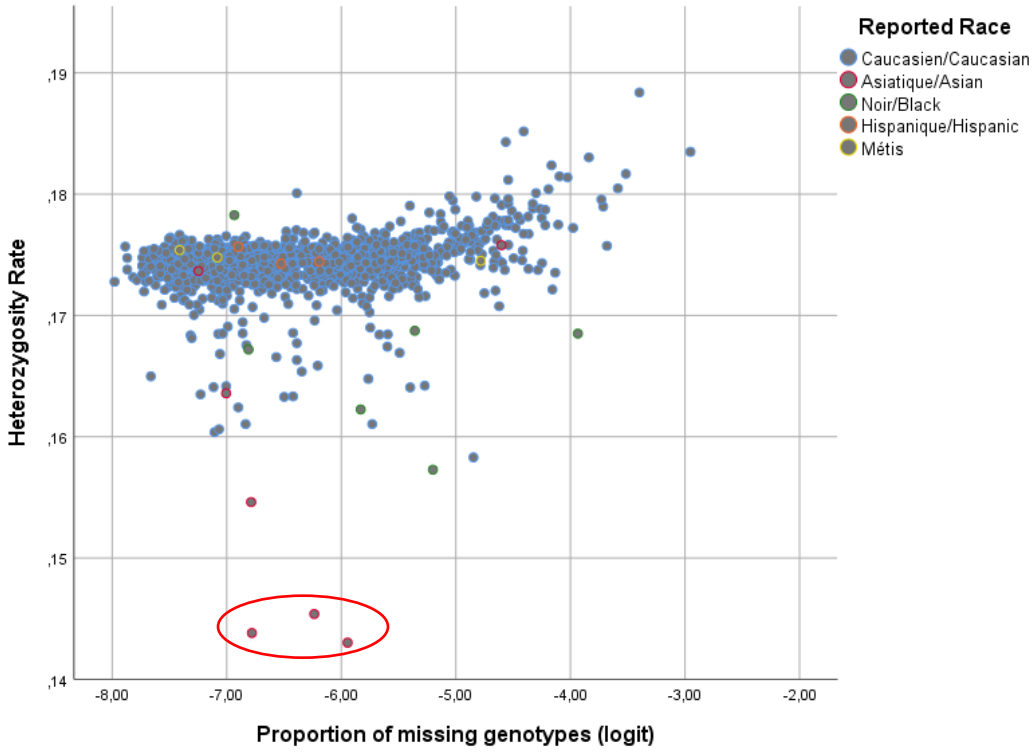
3.4.2. Sample-based QC

Two sample-based QC analyses were performed by NuAge and McGill, respectively, in order to identify low-quality genotyped samples and familial relatedness between NuAge participants.

- Detection of outliers in heterozygosity and missing rates: Samples having extreme heterozygosity or missing genotypes can indicate poor sample quality or cross-contamination of samples. However, other conditions outside of sample quality can explain extreme heterozygosity rate, such as population structure, where non-Caucasians tend to have lower

heterozygosity while mixed ethnicity tend to have higher heterozygosity. Individuals with closely related parents also tend to have lower heterozygosity. In order to flag only samples with quality issues, the NuAge team calculated the heterozygosity rate (--het) and sample-wise missingness rate (--missing) with PLINK for the 1,276 included NuAge participants based on the list of 149,004 independent SNPs (pruned) that passed the QC tests and complementary filters from section 3.4.1. The Figure 1 plots the logit of the sample-wise missingness (logit(miss) = ln(miss / (1 – miss)) against the heterozygosity rate. Points were color-coded by the self-reported ancestry (1,257 [98.5%] Caucasian, 7 [0.5%] Asian, 6 [0.5%] Black, 3 [0.2%] Hispanic, and 3 [0.2%] Metis). As seen in Figure 1, three outliers with lower heterozygosity were noticed (circled), all with Asian reported ancestry but without clustering with the other self-reported Asian participants.

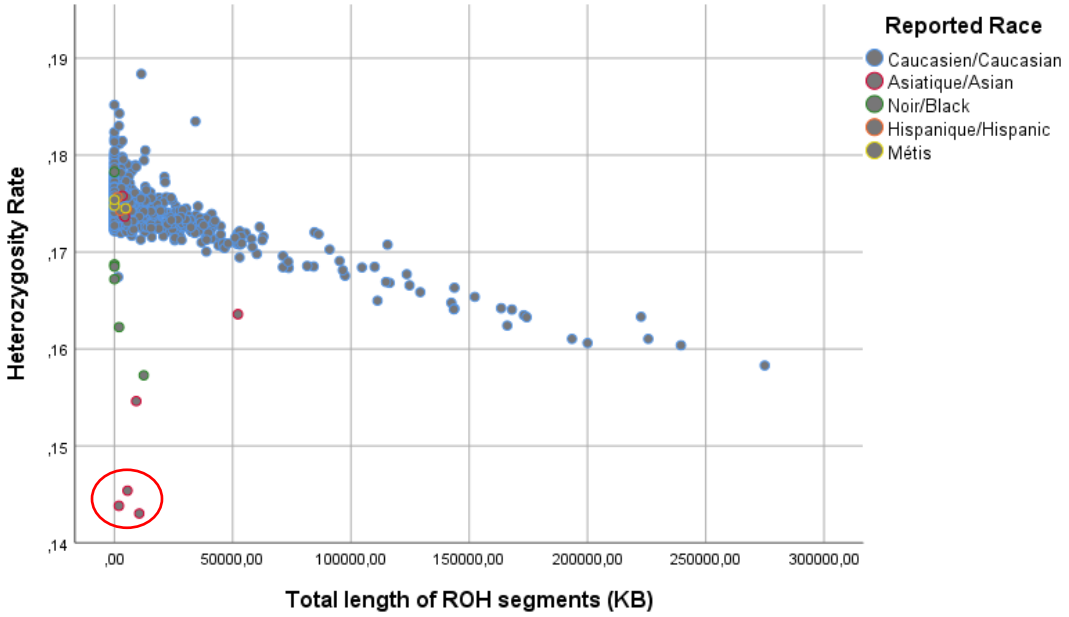
Figure 1. Heterozygosity versus genotype missingness



In order to remove the possibility of closely related parents for these three samples, the NuAge team looked at long runs of homozygosity (ROH) (long stretches of DNA lacking genetic variation; Ringbauer et al. 2021) by calculating the total length of long ROH with PLINK (--homozyg-kb 1000)[4] and using the same set of 149,004 independent SNPs. As seen in Figure 2, total ROH length is short for these three outliers, which discard the possibility of closely related parents to explain their low heterozygosity rate. These three samples were thus flagged in the Sample-based QC report file.

[4] Based on the UK Biobank report (chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fbiobank.ctsu.ox.ac.uk%2Fcrystal%2Fcrystal%2Fdocs%2Fgenotyping_qc.pdf&clen=7603821&chunk=true, accessed December 21, 2021).

Figure 2. Heterozygosity versus total length of ROH segments (KB)



- Identify familial relatedness: Some genetic association studies may need to take into account the degree of relatedness between participants in order to minimize bias in effect estimates. Since this information was not recorded in NuAge participants, it was inferred by analyzing kinship coefficient, identity-by-descent segments (IBD) and proportion of IBD between pair of individuals using the KING software (https://www.kingrelatedness.com/; Manichaikul et al. 2010). This analysis was performed by McGill using the 1,276 unique NuAge participants and the 149,004 pruned markers. All pairs with inferred relatedness of 3rd degree or closer were outputted by the software (kinship coefficient >0.0442). Table 2 shows a total of 104 pairs having a 3rd degree familial relationship or closer. Which means that a total of 1,112 participants were unrelated while 164 were related with at least another participant and flagged in the Sample-based QC report file.

  The kinship coefficient estimator implemented in KING is robust to population structure, but is not reliable for samples with high heterozygosity or high missing rate. Thus, a single poorly genotyped sample could lead to a cluster of inflated relationship. To minimise false positives related samples, we would need to remove self-reported mixed ancestry before the kinship analyses, which was not a response category offered to NuAge participants (no 'mixed ancestry' category). After inferring pairs that are related to the 3rd degree or closer, we verified if there was any pair with a sample having extreme heterozygosity rate (proxy of probable mixed ancestry), which don't seem to be the case based on Figures 1 and 2, or having a sample-wise missing rate >0.02. For instance, all reported relationship had a sample-wise missing rate below 0.02, except two near this threshold (0.021 and 0.023). Following this verification, we finally kept all kinship pairs identified in Table 2.

Table 2. Count of kinship pairs per type of inferred relationship

Inferred relationship	Number of pairs (164 unique participants)
Parent-Offspring	0
Full siblings	34
2nd degree	5
3rd degree	65

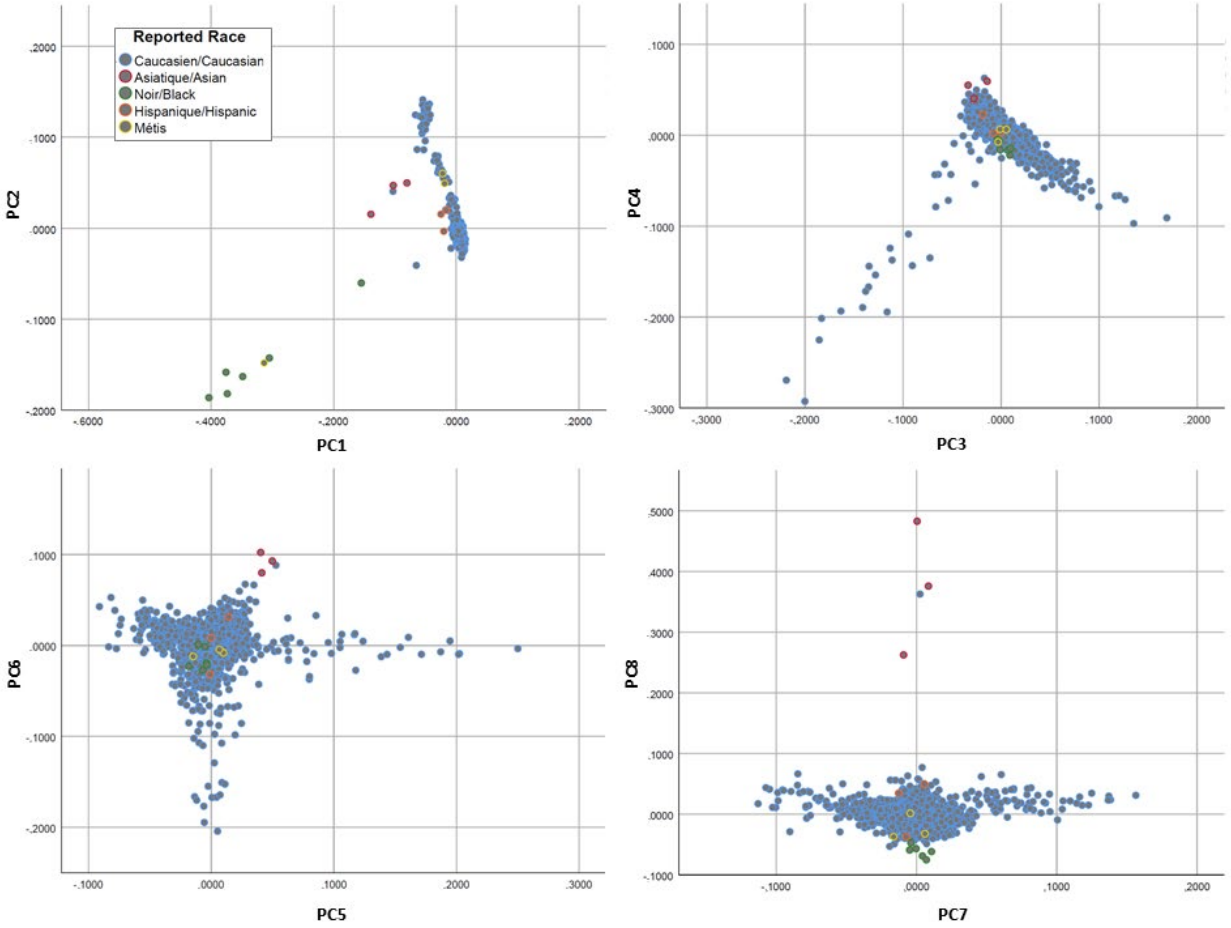
This sample-based QC step tagged a total of 167 individuals (74 men, 93 women) that should be removed for analyses restricted to unrelated individuals, which were identified in the Sample-based QC report file. **In this context, a total of 1,109 unrelated NuAge participants (522 men, 587 women) is obtained and with good quality genotypes.**

3.4.3. Population structure and definition of the Caucasian subset

Controlling for population structure in GWAS statistical models is a common procedure to avoid biased estimates (Price et al. 2006; Balding, 2006) and can be done by extracting the first principal components (PCs) from the genotyping dataset. We thus computed the 20 first PCs for the 1,109 unrelated NuAge participants that passed sample-based QC tests (section 3.4.2). Only the 149,004 independent SNPs that passed QC tests and filters (section 3.3.3) were kept for PC analysis. The method developed by Galinsky et al. (2016) and implemented in PLINK v1.9 (--pca) was used. These 20 PCs were added in the revised Sample-based QC report file. Figure 3 presents pairwise comparison of the first four pairs of PCs (PC 1 to 8) while individuals are color-coded based on their self-reported ancestry.

In order to select a homogenous Caucasian cultural background among NuAge participants, we selected the 1,094 participants (among the 1,109) who self-reported their ancestry as “Caucasian”. We then selected individuals present in the largest cluster throughout the three first PC pairs (PC1/PC2, PC3/PC4 and PC5/PC6). To do so, we identified extreme outliers from bivariate linear regression standardized residuals for the first PC pair (PC1-PC2) using the 3 * interquartile range as the cut-off (boxplot output, IBM SPSS Statistics 25). We then kept the remaining samples (non-outliers) and identified the extreme outliers in the second PC pair (PC3-PC4), and then in the third PC pair (PC5-PC6), using the same procedure as the first PC pair. Finally, the distribution of this Caucasian cluster was further verified by plotting the remaining PC pairs (PC7-PC8 to PC19-PC20). Four additional evident outliers were identified visually in PC11-PC12 and PC13-PC14 pairs. Thus, a total of 985 NuAge participants were kept for the Caucasian subset. Attachment D presents all PC pairs with the Caucasian subset and the extreme outliers identified at each step (color-coded).

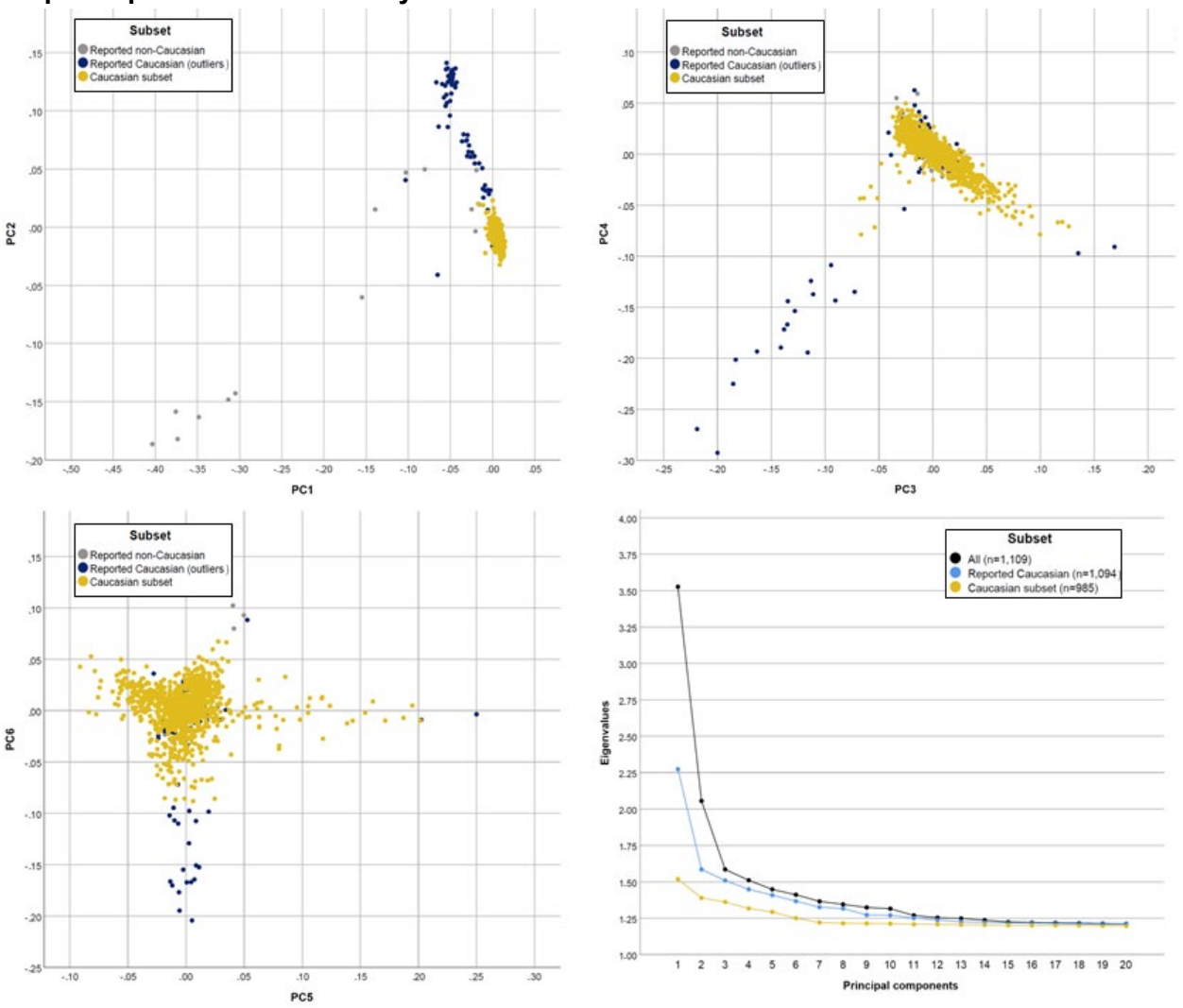
Figure 3. Pairwise plots of the top 8 PCs obtained for the 1,109 unrelated NuAge participants using 149,004 independent SNPs passing all QC tests and complementary filters



As the final picture, Figure 4 presents the three first PC pairs of all NuAge samples (n=1,109) using color-coded classification for the Caucasian subset, reported Caucasian outside of the subset (outliers), and reported non-Caucasian. A second round of PC analysis (PLINK --pca) was then performed to compute the 20 first PCs restricted to the 985 NuAge Caucasian subset (Attachment E). As seen in Figure 4 (bottom right), running the PC analysis using the Caucasian subset reduced the genetic ancestry variance in the first PCs (lower eigenvalues) as compared to PC analysis using all self-reported Caucasian individuals (n=1,094) or all unrelated NuAge participants (n=1,109). **The 985 unrelated individuals included in the Caucasian subset** were identified as "yes" in the Sample-based QC report and their 20 first PCS were also added to this file for further analyses (e.g. in Caucasian subset analyses).

Of note, all 164 participants that were flagged because of familial relationship with at least another participant at the 3rd degree or closer were not part of the Caucasian subset. They were excluded from this subset because of reported non-Caucasian ancestry (n = 1) or were identified as outliers in the first PC pairs (PC1-PC2). Thus, no additional participant can be included in the Caucasian subset if related participants may be considered in future association analyses restricted to this ancestry.

Figure 4. Plots of the top 6 PCs and eigenvalues obtained for the 1,109 unrelated NuAge participants as classified by their inclusion or exclusion from the Caucasian subset



3.5. <u>Sub-objective C) Genotype imputation using the HRC r1.1 reference panel</u>

The genotyping array used in NuAge covers a subset of known genetic markers of the human genome. These directly genotyped markers can then be used to computationally predict genotypes of other genetic markers not covered by the array based on known correlation strength (linkage disequilibrium) existing between markers in the genome, which builds haplotypes (blocks of two or more correlated markers). This imputation process needs a reference panel of densely sequenced individuals from the same ancestry (e.g., Caucasian) in order to refer these haplotypes and then computationally predict genetic markers in a subset of individuals that was partially genotyped. Increasing genotyping data density in NuAge participants is highly relevant for future collaborations in multi-cohort studies and international consortiums in the genomic field. This prerequisite helps to increase power in GWAS and to improve fine-mapping of causal genes and markers.

To run this imputation process on the directly genotyped markers of NuAge participants, we used the Sanger Imputation Service (https://imputation.sanger.ac.uk/) which offers free genotype imputation services provided by the Wellcome Sanger Institute (https://www.sanger.ac.uk/). The Haplotype Reference Consortium (HRC) version r1.1 served as the reference panel for the imputation. The steps realized for this imputation and downstream QC are detailed bellow.

3.5.1. Secure transfer of directly genotyped data into a Compute Canada account

Compute Canada[5,6] services is used for the storage and management of our directly genotyped and imputed data. The transfer of files in and out of the Compute Canada account is securely managed by Globus Connect[7,8] services which is already installed on Compute Canada infrastructure. The PLINK binary files created in section 3.5.1 was transferred into our Compute Canada account before completing the following steps.

3.5.2. Check allele strand and creation of a genotype VCF file

The PLINK binary files were checked for accurate alleles assignment to the TOP (forward) strand against the HRCr1.1 human reference sequence panel (build GRCh37). This verification was done with Will Rayner's HRC preparation checking tool which excluded indels and markers on chrY, MT and pseudo-autosomal regions of chrX, and then outputted a set of PLINK commands to update alleles on the TOP strand (https://www.well.ox.ac.uk/~wrayner/tools/; only three markers needed to be flipped) and to force REF alleles to match those from GRCh37. This tool finally created one .bim and .VCF file per chromosome, which were then combined into a single VCF file (concat) and sorted (sort) using the bcftools. From the 722 976 initial markers, the tool kept a total of 667 291 markers.

3.5.3. Selection of samples and genetic markers and final VCF formatting

We will use directly genotyped data for all 1,276 unique NuAge participants that were successfully genotyped and included in genetic studies, as well as markers that passed the three marker-based QC tests (i.e., excluding those tagged in section 3.4.1; control, HWE, sex), plus these specific exclusion criteria:

- Having a minor allele count (MAC) <2 (corresponding to a MAF <0.078%);
- Having a SNP-wise missingness >=5%;
- Indels, chrY, MT, pseudo-autosomal region of chrX (already removed in section 3.5.3)

We used the vcftools to remove these markers from the VCF file (--exclude-positions –remove-indels –mac 2 –max-missing 0.95 --recode), which remained a total of 652 911 markers. We then compressed the VCF file (bgzip –c), assigned the right chromosome naming scheme as Ensembl (bcftools annotate –rename-chrs; i.e., 1, 2, 3,…, X) and finally indexed the file (bcftools index) to make sure the gzip file is adequately sorted. These steps are detailed in the Sanger Imputation Service resources (https://imputation.sanger.ac.uk/?resources=1). The gzip VCF file was then

[5] For an account request: https://alliancecan.ca/fr/services/calcul-informatique-de-pointe/portail-de-recherche/gestion-de-compte/demander-un-compte.

[6] For technical information: https://docs.alliancecan.ca/wiki/Technical_documentation/fr

[7] https://docs.alliancecan.ca/wiki/Globus/fr

[8] https://docs.globus.org/faq/security/#how_does_globus_ensure_my_data_is_secure

transferred to the Sanger Imputation Service (https://imputation.sanger.ac.uk/) using Globus Connect.

3.5.4. Pre-phasing and imputation on the Sanger Imputation Service website

A preliminary pre-phasing step was carried out using the Eagle2 algorithm (Loh et al. 2016). This first step simply consists of separating the sections of chromosomes from the maternal and paternal side. The markers were then imputed with the PBWT (Positional Burrows_Wheeler Transform) method from the HRCr1.1 reference panel (McCarthy et al. 2016). This panel contains 69,940 haplotypes created from approximately 40 million SNPs (minor allele number >5) on chromosomes 1 to 22 and X in 32,470 individuals mainly of European descent (http://www.haplotype-reference-consortium.org/). The imputation process produced a gzip VCF file for each chromosome along with their respective index file. All files were transferred into our Compute Canada account using Globus Connect. There are thus a total of 40,359,612 directly genotyped and imputed markers available in these VCF files. Note that no additional imputation process was performed for the Human Leukocyte Antigen (HLA) region.

In order to efficiently filter or check genetic markers based on their imputation quality score, we used the vcfparse.pl perl script developed by Will Wrayner which extracts the first 8 columns of the VCF files; these files are also available with the imputed datasets (see section 3.7). The imputation quality score ('INFO') is provided in the 8th column. See this website for more information on this script (https://www.well.ox.ac.uk/~wrayner/tools/Post-Imputation.html).

3.6. Sub-objective D) Genotyping of APOE ε2, ε3 and ε4 alleles

The APOε4 allele is known to be an important risk factor for the early and late onset Alzheimer disease, and with cognitive decline during normal aging, while APOε2 allele decreases this risk (Liu et al. 2013; Lumsden et al. 2020). We can define APOε4 and APOε2 carriers by genotyping two missense SNPs located in the coding sequence of the *Apolipoprotein A* (*APOE)* gene, which are rs429358 T>C (chr19:45411941; GRCh37) and rs7412 C>T (chr19:45412079) (see Table 3). The APOε3 allele is more frequently seen worldwide (~78%), as opposed to the APOε4 (~8%) and APOε2 (~14%) alleles, and varies depending of the ethnicity (Liu et al., 2013).

Due to its strong association with cognitive function, it becomes necessary to adjust for this covariable in secondary studies aiming to identify determinants of cognitive status and decline. The SNPs rs429358 and rs7412 are among the list of SNPs genotyped on the UK Biobank Axiom™ Array, but both markers failed during the genotyping calling process (section 3.3.3). To overcome this issue, we decided to use TaqMan RT-PCR genotyping assays as provided by McGill Genome Center services. These analyses were financially supported by the Quebec Network for Research on Aging (RQRV).

Genotyping was done using the same ssDNA extracts obtained for the 1,303 unique NuAge participants included in genetic studies (see section 3.3.2). The two APOE SNPs were genotyped using TaqMan SNP Genotyping Assays with validated primers and probes from Applied Biosystems™ (rs429358, Cat.no. 4351374, Assay.ID C_3084793_20; rs7412, C_904973_10, 4351379). PCR reactions were prepared using the TaqPath™ ProAmp™ Master Mix with ROX™

dye (ThermoFisher, Cat.no. A30866), following manufacturer instructions for 5 μL per reaction and standard real-time PCR thermal cycling. Genotypes were determined using a LightCycler® 480 System (Roche). Among the 1,303 ssDNA samples, we successfully genotyped rs429358 for 1,296 participants (MAF = 11%) and rs7412 for 1,302 participants (MAF = 10%). These MAF are near those observed in the European ancestry (7-8%; https://www.ncbi.nlm.nih.gov/snp/). APOE ε2, ε3 and ε4 alleles (and rarely ε1) were then defined based on allele calls displayed in Table 3.

Table 3. Definition and approximate frequency of APOE alleles in NuAge

APOE alleles	SNPs		APOE allele frequency (1296 participants) (2592 alleles)
	rs429358 T>C	rs7412 C>T	
ε4	C	C	10%
ε3	T	C	79%
ε2	T	T	10%
ε1	C	T	<0.5%

Adapted from Lumsden et al. 2020.

3.7. Sub-objective E) Final preparation and availability of genotyping datasets for secondary studies

All the imputed genotyped datasets available for usage in secondary studies will be stored and managed in our Compute Canada account. The directly genotyped datasets and the APOE will be stored both, in our Compute Canada account and in the CIUSSS de l'Estrie-CHUS Research Center on Aging's server. The NuAge database Coordinator will supervise the preparation of the relevant genotyping dataset, and complementary files, for each secondary study within the Compute Canada account. The format of the genotyping file will depend on the project's protocol and the researcher's experience with genetic data, and will need to be discussed with the Database Coordinator. Available formats can be VCF (.vcf), PLINK binary files (.bed, .bim, .fam), and PLINK standard plain text format (.map, .ped), as detailed here: https://www.cog-genomics.org/plink/1.9/formats. Other formats can be requested depending on the statistical software that will be used for the project (e.g. SAS, SPSS, and CSV formats).

The participant's identifier code in the genotyped datasets will always remain the same in all secondary projects. However, once the files are prepared for each project, the original participants' identifier code in the NuAge datasets (e.g. data from tests and questionnaires) will be changed (double coded). Therefore, a key file that links the participant's identifier in the genetics data to the one created for each project for the other NuAge datasets will be sent along the transferred data. The key that links the double coded identifier to the NuAge's participant identifier will be kept in a secure file on the NuAge server.

The genotyping files will then be transferred to the research team via Globus Connect and stored securely by them, as requested in our NuAge Database and Biobank Guidelines and the Data Transfer Agreement. In some cases, we will strongly suggest that researchers use the PLINK software to manipulate genotyping files and consider using Compute Canada services, depending

on the number of SNPs requested, the size of the files and the computation capacity needed to run the analysis.

Below is the list of genotyping and complementary files available for secondary studies:

Directly genotyped markers with Affymetrix UK Biobank Axiom™ array (final version prepared by NuAge containing 722,976 markers for 1,276 unique NuAge participants):

- NuAge_AX001toAX017_1276ID_finalv1.bed – PLINK binary biallelic genotype table
- NuAge_AX001toAX017_1276ID_finalv1.bim – PLINK extended map file
- NuAge_AX001toAX017_1276ID_finalv1.fam – PLINK sample information file (with reported sex; see section 3.3.4)

More information regarding the binary PLINK format (.bed, .bim, .fam) files can be found on the PLINK website: https://www.cog-genomics.org/plink/1.9/formats. The .bim and .fam files list the order of markers and genotyped individuals. We recommend using PLINK to manipulate these files.

Imputed genotypes on HRCr1.1 (for 40,359,612 markers for 1,276 participants):

- {1,2,…,X}.vcf.gz
- {1,2,…,X}.vcf.gz.csi
- {1,2,…,X}.vcf.cut.gz

More information regarding the VCF format (.vcf.gz; .vcf.gz.csi) can be found on the Sanger Imputation Service website (https://imputation.sanger.ac.uk/?about=1#pipeline) and here[9]. Note that the individual IDs correspond to the merging of the two first columns of the binary PLINK files (FID_IID). We recommend using bcftools and vcftools to manipulate the gzip VCF files. PLINK can also accept gzip VCF files for running association analyses and run other common functions (see https://www.cog-genomics.org/plink/1.9/input#vcf). As detailed in section 3.5.4, the imputation quality score for each genetic marker can be more efficiently accessible using the parsed files (.vcf.cut.gz; first 8 columns of VCF files). Note that the directly genotyped markers are also part of the imputed genotypes datasets.
bcftool: (https://github.com/samtools/bcftools/releases/tag/1.17)
vcftools: (https://vcftools.sourceforge.net/man_latest.html)

Marker-based QC report file (final version prepared by NuAge):

- NuAge_AX001toAX017_markerQC_finalv1.txt – Tab delimited (Attachment B)

Sample-based QC report file (final version prepared by NuAge):

- SampleReport_NuAge_1276ID_finalv1.txt – Tab delimited (Attachment C)

APOE genotyping via TaqMan RT-PCR (rs429358, rs7412) for 1,303 participants:

- APOE_NuAge_1303ID_finalv1.txt – Tab delimited (Attachment F)

[9] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://samtools.github.io/hts-specs/VCFv4.2.pdf

4. FUNDING SOURCES

All fees for biological samples preparation and transportation between the NuAge biobank and McGill University, as well as for the APOE TaqMan RT-PCR genotyping assays, were covered by funds granted by the Quebec Research Network on Aging (Réseau québécois de recherche sur le vieillissement), a thematic network funded by the Fonds de recherche du Québec – Santé (FRQS). All cost for DNA and RNA extraction, as well as the genotyping of NuAge participants using the UK Biobank Axiom™ Array, were covered by discretionary funds from Professor Mark Lathrop from McGill University. Data analysis was also supported by Canada Foundation for Innovation – Major Science Iniative Fund #3544 (Professor Mark Lathrop and Professor Jiannis Ragoussis).

The NuAge Database and Biobank are supported by the Fonds de recherche du Québec (FRQ; 2020-VICO-279753), the Quebec Network for Research on Aging, and by the Merck-Frosst Chair funded by La Fondation de l'Université de Sherbrooke. These funding sources must be acknowledged in each scientific communication (e.g. manuscript, presentation).

5. AUTHORSHIP AND ACKNOWLEDGMENT REQUIREMENTS

Due to the important contribution of the McGill partners, the NuAge Database Coordinator, and the director of the NuAge biobank, all researchers that will use UK Biobank Axiom™ Array genotyping data in a NuAge secondary research project, are requested to invite these contributors as co-researchers of the study and co-authors of any scientific communications (including -but not limited to- manuscripts, posters, oral presentations).

- Funding:
  Professor Mark Lathrop, McGill University Genome Centre, Department of Human Genetics, McGill University, QC, Canada; mark.lathrop@mcgill.ca
- Sample preparation and array analysis:
  Professor Jiannis Ragoussis, McGill University Genome Centre, Department of Human Genetics, McGill University, Department of Bioengineering, McGill University, QC, Canada; ioannis.ragoussis@mcgill.ca
- NuAge final QC, imputation and datasets preparation:
  Valérie Turcot, Research Center on Aging, CIUSSS de l'Estrie-CHUS, Sherbrooke, QC, Canada; valerie.turcot@usherbrooke.ca
- NuAge biobank management
  Professor Pierrette Gaudreau, Department of Medicine, Université de Montreal, QC, Canada; pierrette.gaudreau@umontreal.ca

Along with the funding sources listed above, we strongly recommend to acknowledge in manuscripts, the valuable assistance of all staff members who have prepared the PBMC samples, extracted DNA and RNA from PBMCs, genotyped NuAge participants, and run preliminary QC steps, namely Corinne Darmond, Alexandre Bélisle, Ariane Boisclair, Antoine Paccard, and Rui Li from McGill University, as well as Patricia L'Archer from the CHUM Research Center.

6. REFERENCES

Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006 Oct;7(10):781-91.

Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am J Hum Genet. 2016 Mar 3;98(3):456-472.

Liu CC, Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol. 2013 Feb;9(2):106-18.

Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, L Price A. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016 Nov;48(11):1443-1448.

Yengo L, Vedantam S, Marouli E, …, Okada Y, Wood AR, Visscher PM, Hirschhorn JN. A Saturated Map of Common Genetic Variants Associated with Human Height from 5.4 Million Individuals of Diverse Ancestries. bioRxiv 2022.01.07.475305; doi: https://doi.org/10.1101/2022.01.07.475305

Lumsden AL, Mulugeta A, Zhou A, Hyppönen E. Apolipoprotein E (APOE) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the UK Biobank. EBioMedicine. 2020 Sep;59:102954.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867-2873

Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. The American Journal of Human Genetics, 76(5):887-893, 2005.

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, …, Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279-83.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904-9.

Ringbauer H, Novembre J, Steinrücken M. Parental relatedness through time revealed by runs of homozygosity in ancient DNA. Nat Commun 12, 5425 (2021). https://doi.org/10.1038/s41467-021-25289-w

Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005 May;76(5):887-93.

ATTACHMENT A – CONSENT FOR GENETIC STUDIES AMONG THE 1,753 NUAGE DATABASE AND BIOBANK PARTICIPANTS

Status	Justification	Number
Voluntary	**Signed Banking ICF* – positive response for genetic studies**	**1200**
Presumed voluntary	**Signed Banking ICF – unspecified response**	**204**
Presumed voluntary	**Signed Banking ICF – "yes" and "no" checked**	**1**
Not voluntary	Signed Banking ICF – negative response for genetic studies	1
Presumed not voluntary	Absence of signed Banking ICF Initial ICF does not include equivalent consent	347

* ICF, Informed Consent Form.

ATTACHMENT B – MARKER-BASED QC REPORT FORMAT

This file contains information about the QC of the directly genotyped markers. It is a plaintext file separated by tabs:

- NuAge_AX001toAX017_markerQC_finalv1.txt

It contains the following columns:

Column order Column header name — Comments if applicable [Datatype]

1 Affy_probeID — Affymetrix array probeset identifier [string]

2 Affy_SNPID — Affymetrix array marker identifier [string]

3 dbSNP_rsID — dbSNP identifier, version 142, if available [string]

4 Chr — Chromosome number 1-26 — 23=X, 24=Y, 25=XY (pseudo-autos.), 26=MT [numeric]

5 Position — Chromosomal position, build GRCh37 (hg19) [numeric]

6 A1_minor_allele — Minor allele based on PLINK (--freq) [string][10]

7 A2_major_allele — Major allele based on PLINK (--freq) [string]

8 Ref_allele — Reference allele as in annotation file Axiom_UKB_WCSG.na35.annot.csv [string]

9 Alt_allele — Alternative allele as in annotation file Axiom_UKB_WCSG.na35.annot.csv [string]

10 MAF_cat — Minor allele frequency category based on PLINK (--freq) [numeric]

11 QC_2ctr_disc — Marker failed positive control genotype discordance test [0/1=no/yes]

12 QC_hwe_disc — Marker failed HWE discordance test based on PLINK (--hardy) [0/1=no/yes]

13 QC_sexgeno_disc — Marker failed sex genotype frequency discordance test [0/1=no/yes]

14 QC_mono_MAC0 — Monomorphic marker (MAC: C1=0) based on PLINK (--freq count) [0/1=no/yes]

15 QC_indel — Indels based on A1 and A2 alleles from PLINK .bim file [0/1=no/yes]

16 QC_not_CHR1_22 — Markers in autosomes (chr1-22) [0/1=yes/no]

17 QC_low_maf1prct — Very low minor allele frequency (<1%) based on PLINK (--freq) [0/1=no/yes]

18 QC_high_miss1prct — High missingness (>=1%) based on PLINK (--missing) [0/1=no/yes]

19 PASSING_markers — Marker passing QC tests and complementary filtering (n=515,077) [0/1=not passing/passing]

20 PRUNED_markers — Marker pruned in or out among those passing QC tests and complementary filtering (n=149,004 pruned in markers) [0/1=pruned in/pruned out]

[10] All PLINK analyses were run with 1,276 unique NuAge participants included in genetic studies with reported sex included in binary files.

ATTACHMENT C – SAMPLE-BASED QC REPORT FORMAT

This file contains information about the QC of the NuAge participant samples. It is a plaintext file separated by tabs:

- SampleReport_NuAge_1276ID_finalv1.txt

It contains the following columns:

Column order Column header name — Comments if applicable [Datatype]

1 IID — NuAge participant identifier as shown in the PLINK .fam file and specific for each secondary project (double coding) [string]

2 genetic_sex_PLINKcorr — Genetic sex as reported by PLINK corrected version [0/1/2=unknown/men/women]

3 reported_sex — Self-reported sex from NuAge database [1/2=men/women]

4 het_miss_outliers — Sample identified as outlier in heterozygosity and missing rates (section 3.4.2) [0/1=no/yes]

5 in_kinship — Sample (unique participant) which have a least one 3rd degree are closer familial relatedness with another genotyped sample (unique participant) (section 3.4.2) [0/1=no/yes]

6 in_caucasian_subset — Sample selected in the Caucasian subset (section 3.4.3) [0/1=no/yes]

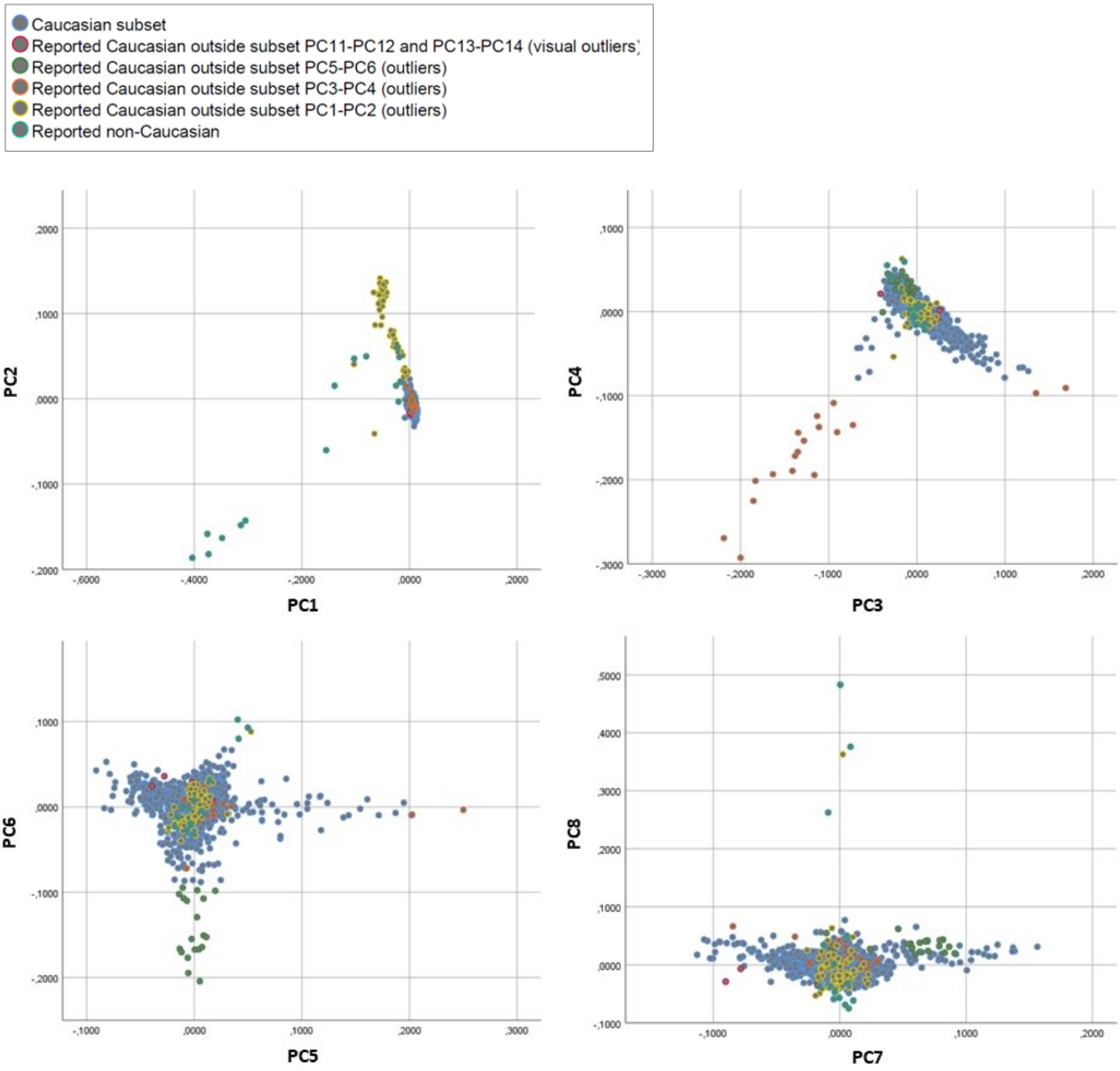
7 PC1_1109ID — Principal component 1 score for the 1,109 unrelated NuAge participants with good quality genotypes (section 3.4.2) [numeric]

8-26 PC2-20_1109ID — Principal component 2-20 scores for the 1,109 unrelated NuAge participants with good quality genotypes (section 3.4.2) [numeric]

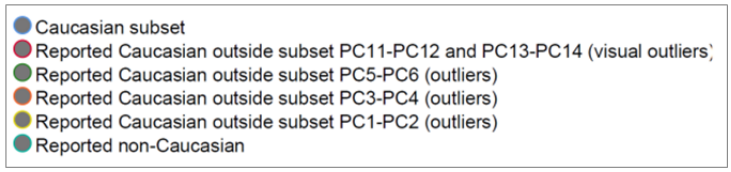
27 PC1_985ID_cauc — Principal component 1 score for the 985 unrelated NuAge participants with good quality genotypes and in the Caucasian subset (section 3.4.3) [numeric]

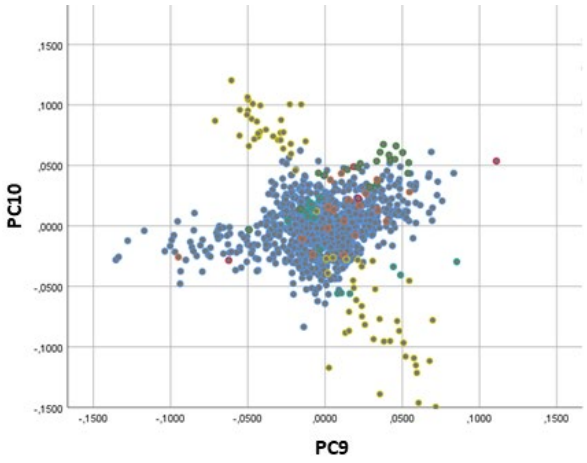
28-46 PC2-20_985ID_cauc — Principal component 2-20 scores for the 985 unrelated NuAge participants with good quality genotypes and in the Caucasian subset (section 3.4.3) [numeric]

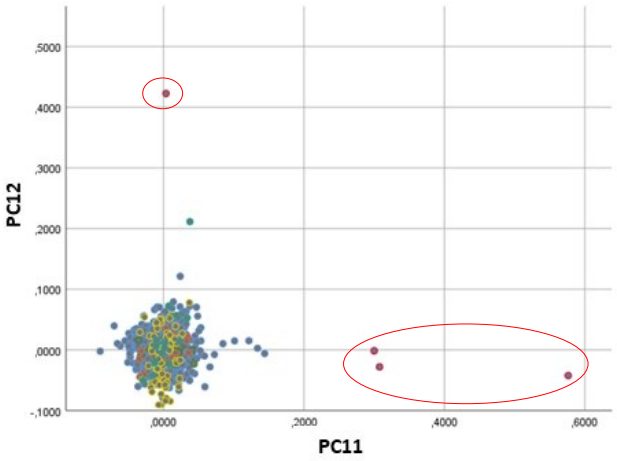
ATTACHMENT D – PAIRWISE PLOTS OF THE 20 FIRST PRINCIPAL COMPONENTS OBTAINED FOR THE 1,109 UNRELATED NUAGE PARTICIPANTS CLASSIFIED BY THEIR INCLUSION OR EXCLUSION FROM THE CAUCASIAN SUBSET

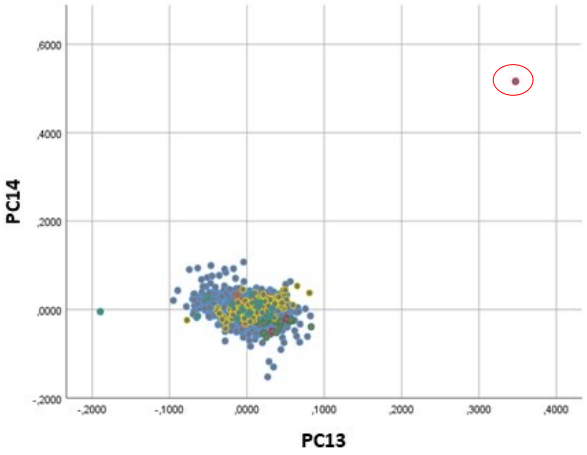


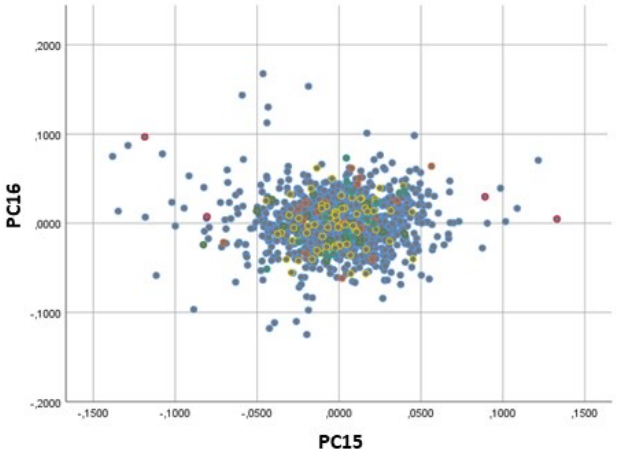
ATTACHMENT D (CONTINUED)

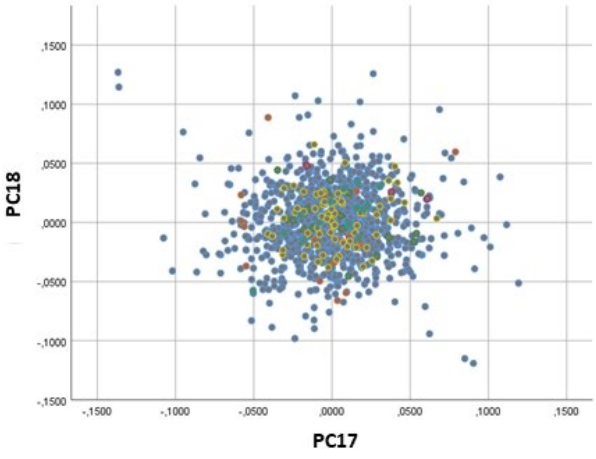


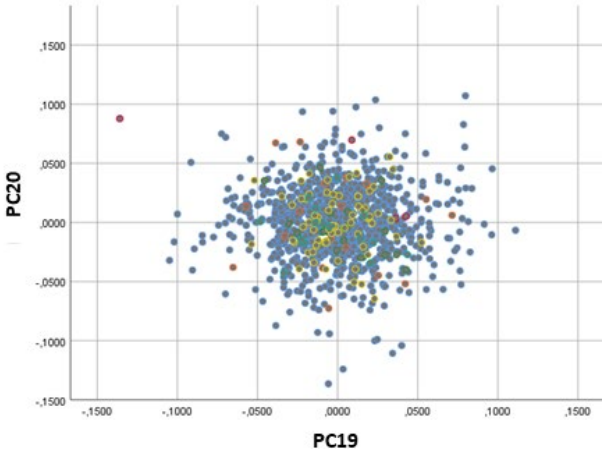




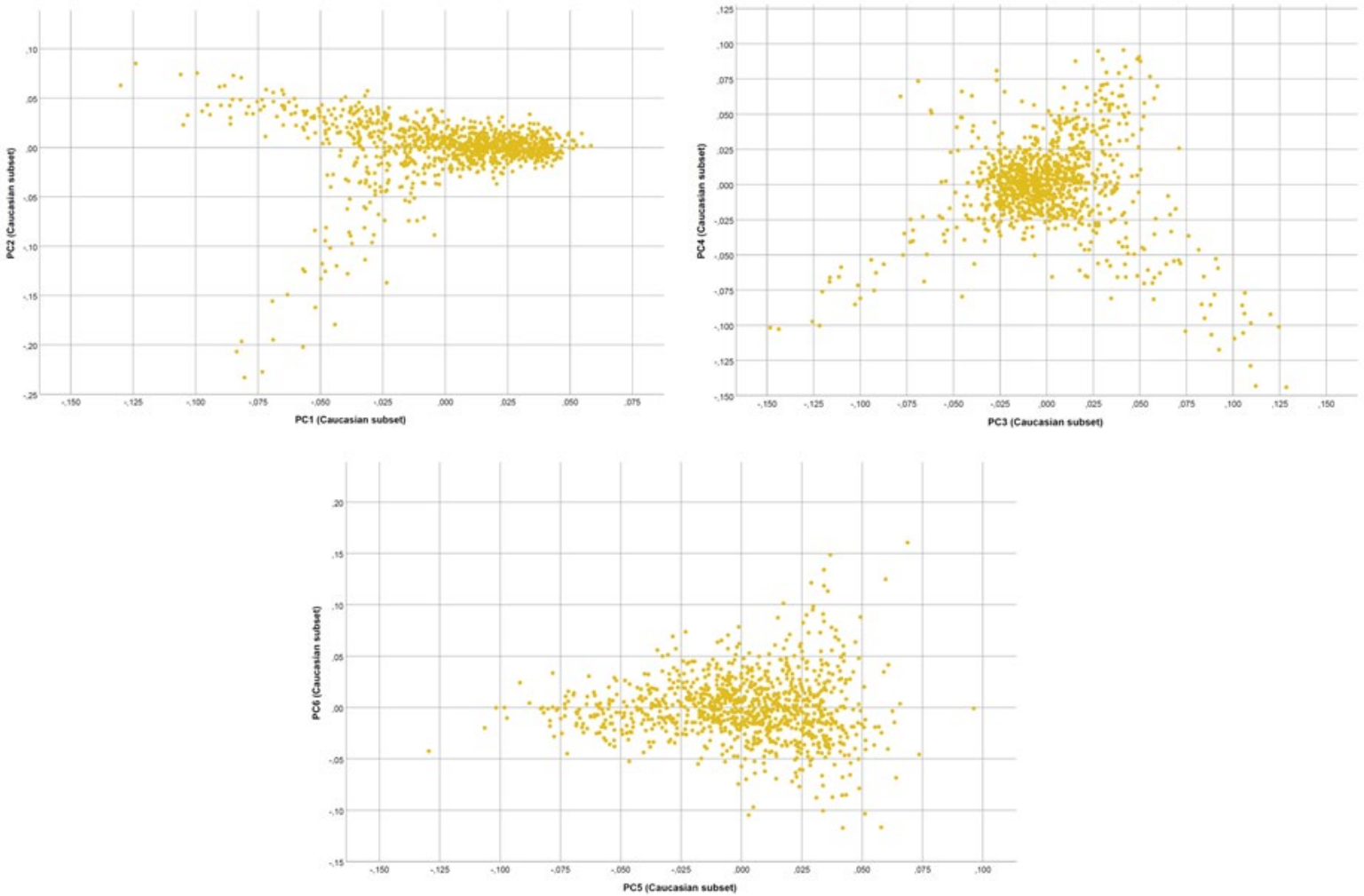








ATTACHMENT E – PAIRWISE PLOTS OF THE 6 FIRST PRINCIPAL COMPONENTS OBTAINED FOR THE 985 UNRELATED NUAGE PARTICIPANTS INCLUDED IN THE CAUCASIAN SUBSET



ATTACHMENT F – APOE TAQMAN RT-PCR GENOTYPING DATASET FORMAT

This file contains information about the QC of the NuAge participant samples. It is a plaintext file separated by tabs:

- APOE_NuAge_1303ID_finalv1.txt

It contains the following columns:
Column order Column header name — Comments if applicable [Datatype]

1 IID — NuAge participant identifier (same as in the PLINK .fam file) and specific for each secondary project (double coding) [string]

2 rs429358 — Biallelic genotype [A/C/G/T; missing='.']

3 rs7412 — Biallelic genotype [A/C/G/T; missing='.']

4 Haplo_APOE_A1 — Putative haplotype allele 1 (unphased) based on rs429358 and rs7412 alleles (section 3.6) [A/C/G/T; missing='.']

5 Haplo_APOE_A2 — Putative haplotype allele 2 (unphased) based on rs429358 and rs7412 alleles (section 3.6) [A/C/G/T; missing='.']

6 APOE_Allele1 — Assigned APOE allele 1 based on Haplo_APOE_A1 [E1/E2/E3/E4; missing='.']

7 APOE_Allele2 — Assigned APOE allele 2 based on Haplo_APOE_A2 [E1/E2/E3/E4; missing='.']

Note:
Ambiguous APOE alleles were observed when rs429358 genotypes were CT or TC and rs7412 genotype was TC (n = 22). As done in Lumsden et al. (2020), APOE alleles were assigned as E4/E2 (haplotypes = CC/TT) since the E1/E3 (haplotypes = CT/TC) are extremely rare.