MIDUS Genomics Project

APOE Data Documentation

For File:

M2M3MR1_GEN_APOE_N1535_20250512

May 2025

MIDUS Apolipoprotein E (APOE) Data Documentation

Overview

APOE genotyping was conducted on 1,535 Biomarker project participants from MIDUS 2 (M2), MIDUS 3 (M3), and MIDUS Refresher 1 (MR1). Specifically, 567 from M2, 148 from M3, and 820 from MR1 Biomarker participants who consented to genomic study and provided blood sample were included in the APOE dataset. The 148 Core cases in M3 listed above either did not initially consent to genomic testing or did not provide a blood sample for DNA isolation during M2 but did consent and provide a sample during M3 and therefore were included in the APOE sequencing data. The data file only includes participants whose APOE data is available.

MIDUS follows variable naming and coding conventions (seen Naming and Coding Conventions included with the Survey documentation), thus the variable names for the APOE data begin with the unique 6 characterset BCRA6D based on the source of the blood DNA sample (i.e., B for M2, C for M3, RA for Refresher 1, 6 for Genomics project, D for data derived from DNA). The SPSS data file "M2M3MR1_GEN_APOE_N1535_20250507.sav" contains 6 variables as follows:

- Administrative:
  - M2MRID – contains the public identifiers for the MIDUS core sample and the MIDUS Refresher (for details, see README of APOE data)
  - SAMPLMAJ – the identifier created by the MIDUS Administrative Core to indicate the participants' sample of origin (e.g. MIDUS Refresher, Twin, etc.),
  - M2M3MRCASE – indicates source of blood DNA sample (e.g., MIDUS 2, MIDUS 3 or MIDUS Refresher 1)
- APOE
  - BCRA6DRS429358 – Rs429358
  - BCRA6DRS7412 – Rs7412
  - BCRA6DAPOE – APOE genotype

DNA Samples and Isolation

Whole-EDTA blood was collected from each participant and aliquoted into cryovials that were later stored at –80°C. QIAamp 96 DNA Blood Isolation kits (Qiagen, Germantown, MD, USA; cat. no. 51162) were used for extraction of genomic DNA according to the manufacturer's protocol. DNA concentrations were determined using the QuantiFluor ONE dsDNA System (Promega, Madison, WI, USA; cat. no. E4870) according to manufacturer's protocol and quantified using the Agilent Biotek Synergy H1 fluorometer. Following determination of DNA concentrations, 20ul samples in 96-well PCR plates were submitted in batches to the University of Wisconsin-Madison Biotechnology Center for APOE haplotype sequencing.

Construction and Sequencing of APOE libraries

Concentration of genomic DNA submitted to the University of Wisconsin-Madison Biotechnology Center was further verified fluorometrically using either the Qubit® dsDNA HS Assay Kit or Quant-iT™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA). Samples were then prepared in a similar process to the one described in Illumina's 16s Metagenomic Sequencing Library Preparation

Protocol, Part #15044223 Rev. B (Illumina Inc., San Diego, California, USA) with the following modifications:

The APOE gene was amplified with fusion primers:
Forward primer:
5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGAATTCGCCCCGGCCTGGTACAC-3'

Reverse primer:
5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCTTGGCACGGCTGTCCAAGGA-3'

Region specific primers, previously described in Hixson and Vernier 1990 (underlined sequences above) were modified to add Illumina adapter overhang nucleotide sequences 5' of the gene-specific sequences. Following initial amplification, reactions were cleaned using a 1x volume of AxyPrep Mag PCR clean-up beads (Axygen Biosciences, Union City, CA). In a subsequent PCR, Illumina dual indexes and sequencing adapters were added using the following primers

Forward primer:
5'-AATGATACGGCGACCACCGAGATCTACAC[55555555]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

Reverse Primer:
5'-CAAGCAGAAGACGGCATACGAGAT[77777777]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT -3', Bracketed sequences are equivalent to the Illumina Dual Index adapters D501-D508 and D701-D712,N716,N718-N724,N726-N729).

Following PCR, reactions were cleaned using a 0.8x volume of AxyPrep Mag PCR clean-up beads (Axygen Biosciences). Quality and quantity of the finished libraries were assessed using an Agilent DNA 1000 kit (Agilent Technologies, Santa Clara, CA) and Qubit® dsDNA HS Assay Kit (ThermoFisher Scientific), respectively. Libraries were pooled in an equimolar fashion and appropriately diluted prior to sequencing. Paired end, 150 bp sequencing was performed using the Illumina NovaSeq X Plus (Illumina, San Diego, CA).

APOE Variant Calling and Haplotype Inference

Apolipoprotein E (APOE) genomic reference data were obtained from Ensembl (gene ID: ENSG00000130203), corresponding to a region on chromosome 19 from positions 44,905,754 to 44,909,393 on the forward strand of the GRCh38 human genome assembly. Two key single nucleotide polymorphisms (SNPs), rs429358 and rs7412 - referred to in earlier literature as SNV-122 and SNV-158 (Hixson and Vernier, 1990) - are located at offsets 2931 and 3069, respectively, within this sequence.

Raw FASTQ files were preprocessed using fastp (v0.23.2), which performed adapter trimming, quality filtering, and read merging. Processed reads were aligned to the GRCh38 human genome assembly (GCA_000001405.15) using BWA-MEM (v0.7.17-r1188) with default parameters. To ensure mapping specificity, only uniquely aligned, properly paired reads were retained. Alignments exhibiting soft clipping or indels were excluded to minimize mapping artifacts.

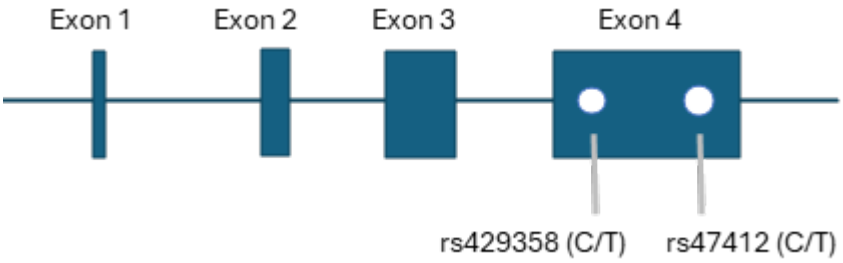
Variant calling was performed using freebayes (v1.3.6) with the following parameters: --min-alternate-fraction 0.1 --report-monomorphic. The resulting variant calls were stored in VCF format. Variant annotation was conducted using snpEff (v5.1d), and phasing of the variants was performed using

Whatshap (v1.7). To assign dbSNP identifiers, the annotate function from bcftools (v1.9) was applied using the dbSNP release 146 database.

APOE genotyping was performed using a custom R script developed at the Bioinformatics Resource Center, University of Wisconsin–Madison. Phased variants at rs429358 and rs7412 were extracted and interpreted according to established haplotype nomenclature (Hixson and Vernier, 1990) to identify APOE protein isoforms (ε2, ε3, ε4). Low-frequency haplotypes (defined as <10% of the maximum observed haplotype count) were excluded to reduce the influence of sequencing artifacts.

APOE Haplotype

APOE haplotypes are determined by two single nucleotide polymorphisms (SNPs), rs7412 and rs429358. Combinations of these SNPs result in the three major APOE isoforms ε2, ε3, and ε4. APOE genotype column report summarized data from Rs429358 and Rs7412 sequencing.
The summarized data results from combining the values in the SNP columns (Rs429358 and Rs7412).



Allele	**ε2**	**ε3**	**ε4**
Haplotype	Rs429358 (T) Rs7412 (T)	Rs429358 (T) Rs7412 (C)	Rs429358 (C) Rs7412 (C)

Top: Schematic representation of the APOE gene and positions of rs429358 and rs7412 in exon 4.
Bottom: The haplotype for **ε2, ε3, ε4** is dependent upon the combinations of C/T at rs429358 and rs7412.

References:

Hixson, J E and Vernier, D T. 1990. Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with HhaI. Journal of Lipid Research. 31:(3), 545-549.