Introduction

This Annex lists the data-fields available from the Airwave Study Tissue Bank APOE haplotypes export.

Document Configuration

Title	Data Dictionary, Annex H
Subject	APOE haplotypes
Version	1.0
Status	Final
Authors	Heard, Andy H, Database Manager at Imperial College London. Georg Otto, Lead Bioinformatician at Imperial College London.
Filename	annex-h.docx

Overview

This export shows the APOE haplotypes of Airwave samples where we or our collaborators have conducted array genotyping leading to determination of the APOE haplotype. You should read the notes below to understand basis on which the analyses have been determined, and its limitations.

GWAS in Airwave

Ideally, every Airwave sample would have had a full genome sequence determined at or soon after it was collected. This was of course not possible because of budget limitations, and instead samples were stored cryogenically until we had the funding to analyse them.

We have conducted four different array based genotyping assays, summarised below. Further technical information on each assay is available if required. The “Samples” column is the number of aliquots submitted for assay, and the number producing complete data is slightly smaller (but well within the expected failure rate).

Year	Laboratory	Samples	Chip / Array
2014	Oxford Wellcome	15024	Illumina Infinium HumanCoreExome-12v1-1 BeadChip
2016	Affymetrix (USA)	4545	Axiom BioBank Chip
2020	Human Genomics Facility, Rotterdam	1029	Infinium® HumanCoreExome BeadChip
2022	Clinical Research Facility, Edinburgh	1824	Illumina GSAMD24v3-0_A1_gb37

In addition, we also have:

- Exome data obtained from the Illumina Infinium HumanExome-12v1-1 BeadChip Array.
- Methylation data assay by IIGM, the Italian Institute for Genomic Medicine.

Neither of these assays provide a determination of APOE and are not considered further here.

Sample Selection and Preparation

To perform each analysis, samples were selected according to criteria that made sense to our funders and whatever analysis plans we had at the time. Once selected and picked, we extracted,

normalised and plated DNA from the sample, sent a primary aliquot for genotyping, and stored unused DNA for later assays.

Our preferred aliquot for obtaining DNA was buffy-coat (white-cells) that had been prepared and saved within 24-hours of venepuncture as part of the laboratory protocol. The buffycoat was stored in a -80 C freezer initially and may later have been transferred to liquid nitrogen. SOPs are available to researchers if required. Sample collected in an EDTA vacutainer were preferred, but when these were unavailable, we used lithium heparin, which seems to have resulted in equally good DNA.

Analysis of Genetic Datasets

The resulting datasets were analysed according to the usual methods at the time, and this resulted in the determination of the APOE haplotype. The 2014 dataset was analysed in 2016 by departmental colleagues. The 2016 dataset was analysed by members of the National BioResource project, which had funded the assay. The 2020 data was analysed in 2023 by departmental colleague, Georg W Otto, who also re-visited the 2014 and 2016 data. The 2022 data was analysed by researchers at Dementias Platform UK (DPUK) who funded the Edinburgh assay and are using it to conduct a separate sub-study of Airwave participants.

The resulting haplotypes are presented as two results: one for the analysis conducted in 2016, and one for the 2023 analysis. Most cases match (N = 16988, 97.9%), but differences exist (N = 356, 2.1%), despite attempts to reconcile the results. This has been explained by G. Otto as follows:

> *The APOE haplotype is determined by the state of two variants (rs429358 and rs7412). The coreExome ("Illumina") and the GSA arrays only determine the allele of rs7412, so one must calculate the full haplotype by imputation. We conducted imputation with the Phase 3v5 1000 Genomes Project data using the software Beagle, version 5.4. Due to the probabilistic nature of this method, discordance between different algorithms is expected, although it has been shown that a high degree of accuracy can be achieved, see Vuoksimaa et al (JAMA Netw Open. 2020, doi:10.1001/jamanetworkopen.2019.19960) and Lupton et al (J. Alzheimer's Dis. 2018, doi: 10.3233/JAD-171104).*

A fuller description of the Georg Otto's analysis is available as *analysis-method-otto.pdf*.

Frequency Distribution of Haplotypes

The table summarises the frequency of haplotypes for each analysis.

Haplotype	**2016**		**2023**	
("NA": Unable to determine)	**N**	**%**	**N**	**%**
E2_E2	97	0.5	118	0.6
E2_E3	2,255	12.2	2,558	12.3
E2_E4	124	0.7	535	2.6
E3_E3	11,144	60.3	12,145	58.3
E3_E4	4,408	23.9	4,966	23.8
E3_NA	0	0.0	2	0.0
E4_E4	441	2.4	524	2.5
	18,469		**20,848**	

Data Labels

The output file described below is tab-separated and includes a header record of column names.

Label	Data Type	Description
barcode	NUMBER (5)	Health-screening identifier.
part_id	NUMBER (7)	Participant identifier.
apoe_substudy	STRING	The sub-study during which GWAS was conducted.
apoe_analysis_2016	STRING	APOE haplotype according to the analysis conducted in 2016
apoe_analysis_2023	STRING	APOE haplotype according to the analysis conducted in 2023

Version History

VERSION	Filename	Date Exported	Total Rows	CRC-64
1	apoe-haplotype-v1.tsv	7th June 2023	21,974	F132CDBA1FC20C6C