Supplementary Materials 2 - Supplementary Methods: Establishment of Cohorts, Identification of NAFLD and Genotyping

## Establishment of Cohorts

### UK Biobank

The UKB is a prospective cohort study with 502,616 participants recruited between 2006 and 2010 from across the UK. Participants were men and women aged between 37 and 73 years at recruitment. Initial assessment involved self-completed touch-screen questionnaire, computer-assisted interview, physical and functional measures and collection of blood samples. Additional data has been generated from health record linkage to national registries and hospital discharges. Over 2.5 million hospital admissions were available for analysis. Additional primary care records were made available for 270,000 participants.

The recruitment process, consent process and data collection has been described extensively elsewhere(1–3). Data is available to researchers after a two-stage online application process. The UKB received ethical approval (research ethics committee reference 11/NW/0382). UKB data access was approved under projects 30439 (phenotype data) and 19655 (genotype data).

### Generation Scotland: Scottish Family Health Study

Generation Scotland: Scottish Family Health Study (GS-SFHS) is a family-based genetic epidemiology study with DNA and socio-demographic and clinical data covering 24,096 volunteers across Scotland aged 18-98 years, from 2006 to 2011. Participants were identified within general practices in Scotland and were aged between 35 and 55 years with at least one first degree relative aged 18 years or over. Participants underwent baseline interview with a blood sample collected and stored for genotyping. Records for individuals were also linked to hospital- and community-based records using their community hospital index number provided to all citizens in Scotland registered with a general practice. A full description of the GS-SFHS protocol has been published previously(4,5). Ethical approval was granted by NHS Tayside Research Ethics Committee (REC reference number 05/S1401/89).

## Identification of Non-Alcoholic Fatty Liver Disease

### UK Biobank

NAFLD was defined as any hospital admission with an ICD-9 or ICD-10 code relating to NAFLD or any primary care encounter with a Read code relating to NAFLD. Controls were defined as those without NAFLD diagnoses. Whilst a diagnostic code should represent genuine NAFLD diagnoses, it is possible that the diagnostic code was incorrectly generated despite the presence of alternative hepatic pathology such as alcohol-related liver disease or viral hepatitis. A second cohort was therefore derived in which alternative hepatic pathology was excluded, remaining NAFLD cases were considered as cases, and all remaining healthy participants without any hepatic pathology served as controls. Diagnostic codes for identification of NAFLD and exclusion of differential diagnoses followed recent expert consensus guidelines (reference in main paper). A full list of the relevant ICD and Read codes, in line with these recommendations, are available in Supplementary Materials 1 (Supplementary Tables 3 and 4).

### Generation Scotland: Scottish Family Health Study

NAFLD was defined as any hospital admission or primary care consultation resulting in generation of an ICD code or Read code relating to NAFLD. The codes and pathologies that were excluded were the same as in the UKB.

## Genotyping and Blood Sampling

### UK Biobank

For the UKB, collection of blood for genotyping was performed at study enrolment for approximately 470,000 individuals, the remainder were excluded due to insufficient sample volume. The cardiorespiratory-focused Affymetrix UK BiLEVE Axiom array was used in 50,000 participants and the Affymetrix UKB Axiom array was used for the remaining participants; the arrays were over 95% similar. Genotyping was performed in 106 batches (4000-5000 individuals per batch). Approximately 900,000 single nucleotide polymorphisms (SNPs) were directly genotyped and subsequent imputation resulted in 93 million SNPs for assessment. The full Affymetrix protocol and description of quality control performed prior to release of genotyped and imputed data to researchers have been published previously(1,6,7).

### Generation Scotland: Scottish Family Health Study

For GS-SFHS, collection of blood samples and genotyping was performed at study enrolment with 20,032 having DNA extracted via blood or saliva. Genotyping was performed on an Illumina HiScan platform and genotypes were called using GenomeStudio Analysis software v2011.1. Genotypes were imputed using the Haplotype Reference Consortium reference panel (HRC.r1-1) via the Sanger Imputation Server pipeline (<https://imputation.sanger.ac.uk>). A total of 24,161,581 SNPs were available for analysis. Genotype data covering each of the identified loci in the UKB GWAS were extracted from GS-SFHS.

## Apolipoprotein E Analysis

Rs429358 is a common variant within apolipoprotein E (*APOE*) which in combination with rs7412 governs determination of the three main alleles of *APOE* (, and )(8). As rs429358 was identified as a risk factor for non-alcoholic fatty liver disease (NAFLD) within the main GWAS, an analysis of the genotype status and risk of NAFLD was undertaken.

### Derivation of Apolipoprotein E Allelic Status

Both rs429358 and rs7412 were directly called within the genotyping arrays meaning that allelic dosages were available in integer format. Genotype dosage was retrieved from the files using QCTOOLS version 2.0.6 (reference in main text).

Genotype was determined based on number of alternate alleles at each SNP (see Supplementary Materials 1: Supplementary Table 6). As the genotype data was unphased it was not possible to determine whether rs7412 TC and rs429358 TC corresponded to a TT and CC haplotype or two CT haplotypes and as such it is not certain whether the patient was of a or a genotype. However, given the extreme rarity of the allele it was determined that patients who were heterozygotes at both SNPs should be considered as genotypes. The other allelic carriers were excluded from the analysis based on rarity of the genotype (see Results section).

Unadjusted association between APOE genotype and NAFLD was tested with test. The association was then evaluated in a logistic regression with the 6 included APOE genotypes. Each of these logistic regressions was adjusted for age and sex and the genetic principal components within the main GWAS model.

## References

1. UK Biobank. UK Biobank: Protocol for a large-scale prospective epidemiological resource. Stockport, UK: UK Biobank; 2007.

2. Elliott P, Peakman TC, UK Biobank. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. International Journal of Epidemiology. 2008;37:234–244.

3. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine. 2015;12:e1001779.

4. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, et al. Generation Scotland: The Scottish Family Health Study; a new resource for researching genes and heritability. BMC Medical Genetics. 2006;7:74.

5. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. International Journal of Epidemiology. 2013;42:689–700.

6. Thermo Fisher Scientific. Axiom Genotyping Solution Data Analysis. 2018. <https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf> (accessed 2018-06-20).

7. Genetic data | UK Biobank. <https://www.ukbiobank.ac.uk/scientists-3/genetic-data/> (accessed 2020-11-04).

8. Kern S, Mehlig K, Kern J, Zetterberg H, Thelle D, Skoog I, et al. The Distribution of Apolipoprotein E Genotype Over The Adult Lifespan and in Relation to Country of Birth. American Journal of Epidemiology. 2015;181:214–217.