Supplementary Material

Assessing Genetic Overlap and Causality Between Blood Plasma Proteins and Alzheimer's Disease

Supplementary Notes: Additional comments on materials and methods

Plasma protein shortlist extended literature review

Studies were sourced based on expert recommendation (Proitsi and Hye) and searches on PubMed and Google Scholar for "Alzheimer blood protein discovery". The search term "Alzheimer blood protein discovery" provided no additional studies passing Kiddle's screening criteria (non-targeted discovery studies, except for panel based studies with >100 candidates). Therefore, search terms were expanded to include "Alzheimer's Disease blood plasma proteins", "Alzheimer's Disease blood biomarkers" and "Alzheimer's Disease proteomics". Proteins were added to the baseline list if individually associated with AD status (at author defined significance threshold, of at least p-value <0.05) or an AD-related phenotype (as defined by list of "outcome variables" in Kiddle et al.). AD status associations were prioritized in studies where association was tested with multiple AD-related phenotypes.

Plasma protein data preparation: Sun et al. GWAS

Blood samples were collected in 6-ml EDTA tubes using standard venepuncture protocols and stored at −80°C before analysis (see [1] for full protocol). Plasma proteins were measured using SomaLogic's multiplexed, aptamer-based assay (SOMAscan) with 4,034 modified single-stranded DNA SOMAmers that bind to specific protein targets which are then quantified using a DNA microarray (see [2] for further detail). QC procedures resulted in 3,283 SOMAmers (mapping to 2,994 unique proteins using UniProt identifiers) for GWAS. Individuals were genotyped for 830,000 variants on the Affymetrix Axiom UK Biobank array and imputation was conducted using a combined 1000 Genomes Phase 3-

UK10K reference panel via the Sanger Imputation Server. After QC to exclude sex mismatches, low call rates, duplicates, extreme heterozygosity, relatedness, and population stratification (see [2] for full protocol), 10,572,788 variants aligned to Genome Reference Consortium genome build 37 (GRCh37) remained for GWAS. Association analysis was performed on the rank-inverse normalized residuals from the linear regression of natural log-transformed protein levels adjusted for age, sex, duration between blood draw and processing (binary, ≤1 day/>1day) and the first three principal components of ancestry from multi-dimensional scaling.

AD data preparation: GERAD1, ADNI, and ANM

GERAD1 participants were genotyped at the Sanger Institute on the Illumina 610-quad chip. These samples were recruited by the Medical Research Council (MRC) Genetic Resource for AD (Cardiff University; Kings College London; Cambridge University; Trinity College Dublin), the Alzheimer's Research UK (ARUK) Collaboration (University of Nottingham; University of Manchester; University of Southampton; University of Bristol; Queen's University Belfast; the Oxford Project to Investigate Memory and Ageing (OPTIMA), Oxford University); Washington University, St Louis, United States; MRC PRION Unit, University College London; London and the South East Region AD project (LASER-AD), University College London; Competence Network of Dementia (CND) and Department of Psychiatry, University of Bonn, Germany and the National Institute of Mental Health (NIMH) AD Genetics Initiative. All AD cases met criteria for either probable (NINCDS-ADRDA, DSM-IV) or definite (CERAD) AD. All elderly controls were screened for dementia using the MMSE or ADAS-cog, were determined to be free from dementia at neuropathological examination or had a Braak score of 2.5 or lower.

Genotype data is available for ADNI1 and ADN2, typed across three separate genotype chips: ADNI1 (Illumina Human610-Quad BeadChip), ADNI2 (Illumina HumanOmniExpress BeadChip), and OMNI (a combination of phase 1 and 2 participants typed on a high coverage Illumina chip - Omni 2.5M). ADNI2 and OMNI are aligned to GRCh37, while ADNI1 is aligned to GRCh36. For ANM, like ADNI1, samples were typed using the Illumina Human610-Quad BeadChip, but data had been moved to GRCh37 prior to acquisition.

MR methodology

Inverse weighted variance (IVW) regresses exposure SNP-instrument associations with outcome SNP-instrument associations, weighted by the inversed variance of outcome SNP-instrument associations [3]. In IVW, the intercept is constrained to zero under the assumption that there is no horizontal pleiotropy. Odds ratios (OR) per 1 standard deviation were calculated to enable comparison to other exposures. Two robust methods, MR-Egger and weighted median [3], were used to generate alternative causal estimates. MR-Egger removes IVW's intercept constraint with large deviations from zero at the intercept and between IVW and egger causal estimates providing evidence of horizontal pleiotropy [4]. Weighted median MR controls for bias in its causal estimate, even if up to 50% of the instruments are invalid, by ordering and weighting estimates by association strength and taking the estimate at the 50^{th} percentile [3]. Lastly, leave one out analysis was conducted to estimate the impact of individual SNPs and Cochran's Q was calculated to test for between SNP heterogeneity.

Protein heritability analysis

Average h^2 across the proteins was 0.10 (see full results at https://alexhandy1.shinyapps.io/ad-genetic-overlap-web-results/); however, results were treated as indicative given the average standard error was 0.16 (including 8 proteins with h^2

less than 0). There are also known issues of applying LDSR to samples less than 5,000 (see https://github.com/bulik/ldsc/wiki/FAQ). Genetic correlation analysis was limited by the same lack of power with 63% of pairwise correlations not able to be calculated ("NA"), an average standard error of 2.3 and only 2 pairwise correlations with a nominal p-value less than 0.05.

AD genetic data preparation

Following alignment of genetic QC with procedures applied to all participant GERAD1, ADNI and ANM data, 3,691,311 variants were available for analysis from the 5,218,413 overlapping variants across cohorts. PRSice removed a further 668,813 variants that did not overlap with the AD base data.

REFERENCES

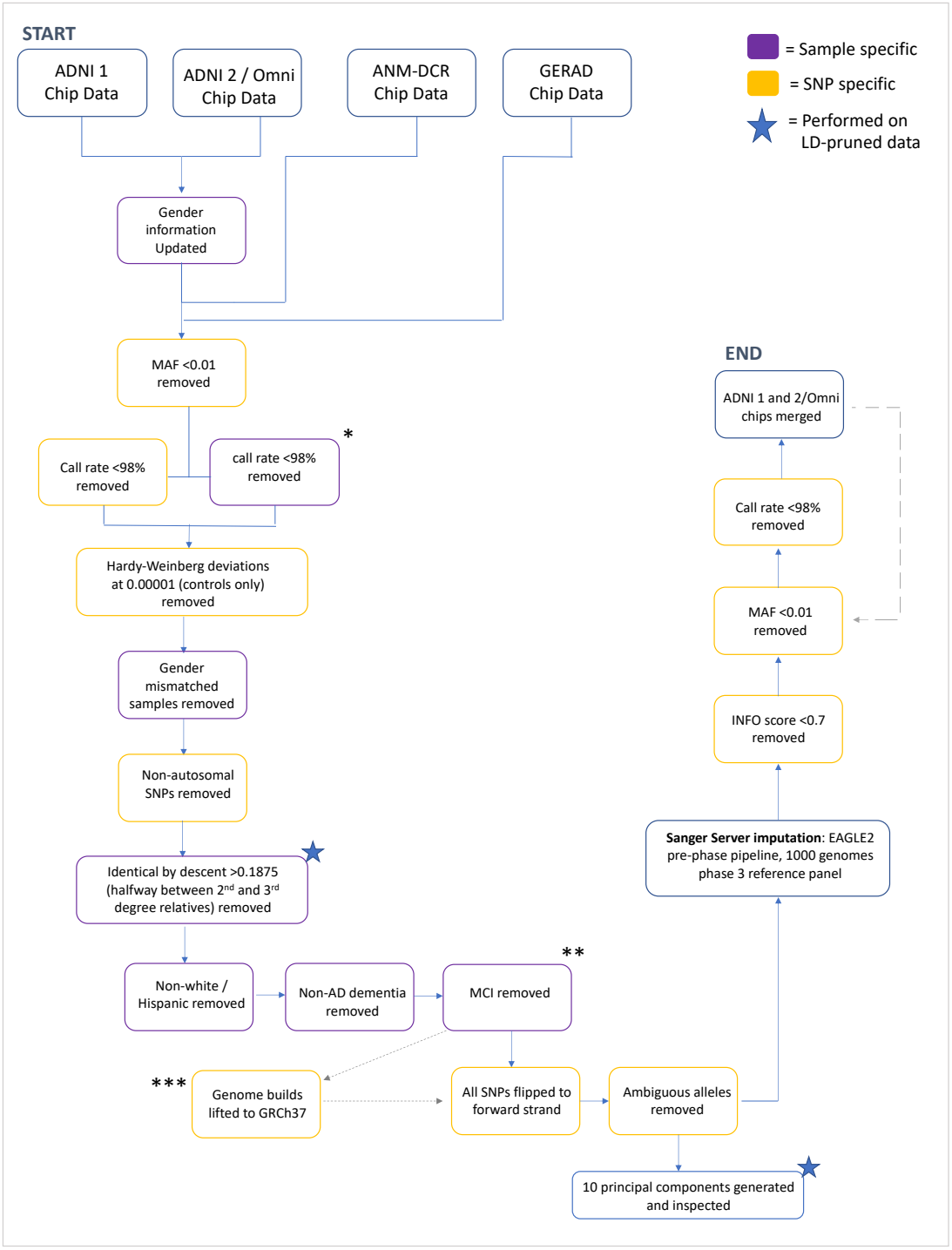
[1] Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, Mehenny S, Mant J, Angelantonio ED, Thompson SG, Ouwehand W, Roberts DJ, Danesh J (2014) The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: Study protocol for a randomised controlled trial. *Trials* **15**, 363.

[2] Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR, Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS, Fox CS, Plenge RM, Danesh J, Runz H, Butterworth AS (2018) Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79.

[3] Bowden J, Davey Smith G, Haycock PC, Burgess S (2016) Consistent estimation in

Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol* **40**, 304–314.

[4] Bowden J, Smith GD, Burgess S (2015) Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–525.



Flow chart illustrating data preparation steps applied to all raw datasets (separately) prior to PRS analyses.
** Call rate missingness for both SNPs and samples was applied iteratively, from 90-98%, iterating between SNPs and samples in steps of 1%.*
*** Latest diagnosis was used to classify samples into cases and controls. Late stage MCI, with MCI due to probable AD, and clinician confidence score of 3-4 (indicating high confidence) remained in analyses as cases.*
**** Required for ADNI1 and GERAD only.*

Supplementary Figure 1. Illustrative overview of genetic QC procedures applied to GERAD, ADNI, and ANM individual level genetic data.

Supplementary Table 1. Summary characteristics of Sun et al participants selected randomly in two subcohorts from the INTERVAL study.

	Subcohort 1 (n=2,481)	**Subcohort 2 (n=820)**
Age (y, SD)	43.6 (14.3)	44.1 (14.2)
Sex (% male)	51.6%	49.5%
BMI (and SD)	26.3 (4.8)	26.6 (4.9)
Current smokers (%)	8.6%	8.5%
Current alcohol use (%)	92.6%	91.0%

Supplementary Table 2. Total SNPs remaining for each protein in shortlist after QC.

Total SNPs pre QC	**10,572,809**	
QC step	**SNPs removed**	**SNPs cumulative drop-out**
Remove duplicate variants	21	21
Remove non bi-allelic variants	1,154,361	1,154,382
Remove variants not in target data	4,208,324	5,362,706
Total SNPs post QC	**5,210,103***	

*PRSice removed an additional 770,215 strand ambiguous SNPs and 42 SNPs (1 GERAD1, 1 ADNI, 39 ANM) with mismatching variants across target and base data.