


## RESEARCH ARTICLE

# Genome- and epigenome-wide studies of plasma protein biomarkers for Alzheimer's disease implicate TBCA and TREM2 in disease risk

Robert F. Hillary<sup>1</sup>  | Danni A. Gadd<sup>1</sup> | Daniel L. McCartney<sup>1</sup> | Liu Shi<sup>2</sup> | Archie Campbell<sup>1</sup> | Rosie M. Walker<sup>1,3</sup> | Craig W. Ritchie<sup>4</sup> | Ian J. Deary<sup>5</sup> | Kathryn L. Evans<sup>1</sup> | Alejo J. Nevado-Holgado<sup>2</sup> | Caroline Hayward<sup>6</sup> | David J. Porteous<sup>1</sup> | Andrew M. McIntosh<sup>1,7</sup> | Simon Lovestone<sup>2,8</sup> | Matthew R. Robinson<sup>9</sup> | Riccardo E. Marioni<sup>1</sup>

<sup>1</sup> Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

<sup>2</sup> Department of Psychiatry, University of Oxford, Oxford, UK

<sup>3</sup> Centre for Clinical Brain Sciences, Chancellor's Building, 49 Little France Crescent, University of Edinburgh, Edinburgh, UK

<sup>4</sup> Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

<sup>5</sup> Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK

<sup>6</sup> MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

<sup>7</sup> Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

<sup>8</sup> Neurodegeneration, Johnson and Johnson Medical Ltd, Wokingham, UK

<sup>9</sup> Medical Genomics Group, Institute of Science and Technology Austria, Klosterneuburg, Austria

## Correspondence

Riccardo Marioni, Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, Crewe Road South, University of Edinburgh, EH4 2XU, Edinburgh, UK.  
 Email: [riccardo.marioni@ed.ac.uk](mailto:riccardo.marioni@ed.ac.uk)

## Funding information

Chief Scientist Office of the Scottish Government Health Directorates, Grant/Award Number: CZD/16/6; Scottish Funding Council, Grant/Award Number: HR03006; Medical Research Council UK; Wellcome Trust, Grant/Award Number: 104036/Z/14/Z; UKRI MRC, Grant/Award Numbers: MC\_PC\_17209, MR/S035818/1; European Union H2020, Grant/Award Number: SEP-210574971; Alzheimer's Research UK, Grant/Award Number: ARUK-PG2017B-10; Alzheimer's Society, Grant/Award Number: AS-PG-19b-010; University of Edinburgh; Horizon 2020 Virtual

## Abstract

**Introduction:** The levels of many blood proteins are associated with Alzheimer's disease (AD) or its pathological hallmarks. Elucidating the molecular factors that control circulating levels of these proteins may help to identify proteins associated with disease risk mechanisms.

**Methods:** Genome-wide and epigenome-wide studies ( $n_{\text{individuals}} \leq 1064$ ) were performed on plasma levels of 282 AD-associated proteins, identified by a structured literature review. Bayesian penalized regression estimated contributions of genetic and epigenetic variation toward inter-individual differences in plasma protein levels. Mendelian randomization (MR) and co-localization tested associations between proteins and disease-related phenotypes.

**Results:** Sixty-four independent genetic and 26 epigenetic loci were associated with 45 proteins. Novel findings included an association between plasma triggering receptor expressed on myeloid cells 2 (TREM2) levels and a polymorphism and

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* published by Wiley Periodicals, LLC on behalf of Alzheimer's Association

Brain Cloud, Grant/Award Number: H2020-SC1-DTH-2018-1; MRC, UK Rosetrees; King Abdullah University of Science and Technology, Saudi Arabia; MRC University Unit Programme, Grant/Award Number: MC\_UU\_00007/10; DPUK - core grant support from the Medical Research Council, Grant/Award Number: MR/L023784/2; UK Medical Research Council Award to the University of Oxford, Grant/Award Number: MC\_PC\_17215; NIHR Biomedical Research Centre at Oxford Health NHS Foundation Trust; Swiss National Science Foundation Eccellenza, Grant/Award Number: PCEGP3-181181; Wellcome, Grant/Award Numbers: 108890/Z/15/Z, 104036/Z/14/Z, 104036/Z/14/Z, 216767/Z/19/Z, 220857/Z/20/Z

cytosine-phosphate-guanine (CpG) site within the *MS4A4A* locus. Higher plasma tubulin-specific chaperone A (TBCA) and *TREM2* levels were significantly associated with lower AD risk.

**Discussion:** Our data inform the regulation of biomarker levels and their relationships with AD.

## 1 | INTRODUCTION

Alzheimer's disease (AD) is one of the leading causes of disease burden and death globally.<sup>1,2</sup> Blood-based methods for assessing disease risk are potentially more cost-effective and less invasive than neuroimaging methods or lumbar punctures for collecting cerebrospinal fluid (CSF). Approaches that use genomics and untargeted proteomics have suggested that there are signals in blood that can supplement targeted assays, and contribute to the understanding and prediction of AD.<sup>3,4</sup> However, the relevance of many candidate protein markers identified by untargeted approaches to AD remains unclear.<sup>5,6</sup> Understanding the molecular factors that regulate the levels of AD-associated proteins may identify proteins that bear relevance to disease risk mechanisms.

Unlike genetic factors, which remain largely stable over the life-course, differential DNA methylation (DNAm) profiles are influenced by genetic and non-genetic factors.<sup>7</sup> DNAm is characterized by the addition of methyl groups to DNA, typically in the context of cytosine-phosphate-guanine (CpG) nucleotide base pairings. Clusters of CpG sites termed CpG islands are located near 70% of gene promoters. CpG island methylation is typically associated with reduced gene expression. However, it is important to note that DNAm is dynamic, tissue-specific, and cell-specific.<sup>8</sup> DNAm data may capture independent information beyond genetic factors in explaining inter-individual variation in circulating protein levels. Several genome-wide association studies (GWAS) have catalogued polymorphisms associated with plasma protein levels and identified proteins that correlate with risk scores for various disease states including AD.<sup>9-11</sup> Zaghlool et al. (2020) performed the only large-scale epigenome-wide association study (EWAS) to date on plasma protein levels (>1000 proteins).<sup>12</sup> Few studies have combined GWAS and EWAS data to quantify the independent and combined contributions of genetic and epigenetic factors toward differential protein biomarker levels.<sup>13-15</sup>

We performed a structured literature review of studies that report associations between plasma proteins and AD diagnosis or related traits such as amyloid beta (A $\beta$ ) burden and cortical atrophy.<sup>16-27</sup> We focused on studies that measured plasma protein levels using

the SOMAscan affinity proteomics platform (SomaLogic Inc.), as this matches the protocol used in our study, Generation Scotland. We identified 282 proteins that were also measured in our sample ( $n \leq 1064$ ). Our first aim was to conduct GWAS and EWAS on plasma levels of 282 AD-associated proteins. Using Bayesian penalized regression, we estimated the proportion of inter-individual variability in plasma protein levels that can be accounted for by variation in genetic and DNAm factors. BayesR+ implicitly adjusts for probe intercorrelations and data structure, including relatedness.<sup>28</sup> For our second aim, we used Mendelian randomization (MR) and co-localization analyses to test for relationships between plasma protein levels and AD phenotypes.

## 2 | METHODS

### 2.1 | Study cohort

Analyses were performed using blood samples from participants of the **ST**ratifying **R**esilience and **D**epression **L**ongitudinally (STRADL) cohort, comprising 1188 individuals from the larger, family-structured Generation Scotland: the Scottish Family Health Study (GS). GS consists of 24084 individuals from across Scotland. Recruitment for GS took place between 2006 and 2011. STRADL participants partook in follow-up data collection 4 to 13 years after baseline.<sup>29,30</sup>

### 2.2 | Search strategy

We searched MEDLINE (Ovid interface, Ovid MEDLINE in-process and other non-indexed citations and Ovid MEDLINE 1946 onwards), Embase (Ovid interface, 1980 onwards), Web of Science (core collection, Thomson Reuters), and medRxiv/bioRxiv to identify relevant articles indexed as of May 28, 2021. Search terms are outlined under supplementary information. Twenty-five articles were identified and one further article was identified through a supplemental manual literature search. After removal of duplicates, 23 articles were assessed for inclusion criteria: (1) original research article, (2) proteins were measured in

plasma, (3) proteins were measured using SOMAscan technology, and (4) proteins were associated with AD or related phenotypes. Twelve articles met inclusion criteria.

### 2.3 | Protein quantification

The 5k SOMAscan v4 array was used to quantify the levels of plasma proteins in GS participants ( $n = 1065$ ). This highly multiplexed platform uses chemically modified aptamers termed SOMAmers (Slow Off-rate Modified Aptamers) that recognize epitopes on their cognate protein targets with high specificity and high affinity in the nanomolar-to-picomolar range. The recognition signal is measured as relative fluorescence units (RFUs) on microarrays.

Plasma samples were collected in 150  $\mu\text{L}$  aliquots and stored at  $-80^{\circ}\text{C}$ . Samples were run in 96-well plates and reagents were spread across three dilution factors (0.005%, 0.5%, and 20%) to create distinct sets for high, medium, and low abundance proteins, respectively. Raw microarray data were normalized through a number of quality control steps, which are detailed in the supplementary information. After quality control and the exclusion of non-human proteins, deprecated markers and spuriomers, 4235 SOMAmers were retained for proteomic analyses.

Normalized RFUs (from SomaLogic) were log-transformed and regressed onto the following covariates: age, sex, study site (Aberdeen/Dundee), time between sample being collected and processed for proteomics (factor, 4 levels), and 20 genetic principal components (PCs) of ancestry from multidimensional scaling (to control for population structure). Relationships between covariates and SOMAmers are shown in Table S1. Residualized RFUs were transformed by rank-based inverse normalization. We refer to these as protein levels; however, they reflect RFUs that have undergone a number of quality control, transformation and pre-correction steps.

### 2.4 | GWAS

Generation Scotland samples were genotyped using the Illumina Human OmniExpressExome-8v1.0 Bead Chip and processed using the Illumina Genome Studio software v2011 (Illumina, San Diego, CA, USA). Quality control steps are outlined under supplementary information. After quality control, 561125 single nucleotide polymorphisms (SNPs) remained for 1064 individuals. In total, 1064 individuals had both genotype and proteomic data available for analyses.

Bayesian penalized regression GWAS were performed using BayesR+ software in C++.<sup>28</sup> BayesR+ utilizes a mixture of prior Gaussian distributions to allow for markers with effect sizes of different magnitudes. It also includes a discrete spike at zero that enables the exclusion of markers with non-identifiable effects on the trait of interest. Guided by data from our previous studies, mixture variances for the stand-alone GWAS were set to 0.01 and 0.1 to allow for markers that account for 1% or 10% of variation in circulating protein levels, respectively.<sup>14,15</sup> In the combined GWAS/EWAS analysis, genotype

### RESEARCH IN CONTEXT

1. Systematic Review: The authors performed a structured literature review of studies that reported associations between SOMAscan-measured plasma proteins and Alzheimer's disease (AD). Twelve studies were included following a search of MEDLINE, Embase, Web of Science, and preprint servers. The goal of the study was to combine genome-wide and epigenome-wide association studies (GWAS and EWAS) with causal modeling methods to investigate associations between plasma proteins and AD risk. The study used data from the Scottish population-based cohort, Generation Scotland.
2. Interpretation: Two hundred eighty-two proteins across the included studies were available for testing in Generation Scotland. Seven novel genetic and 19 novel cytosine-phosphate-guanine (CpG) sites were associated with plasma levels of 18 proteins. Higher plasma levels of tubulin-specific chaperone A (TBCA) and triggering receptor expressed on myeloid cells 2 (TREM2) associated with lower risk of AD.
3. Future Directions: Triangulation of evidence across other experimental and epidemiological approaches will be necessary to determine if blood proteins influence AD risk.

and DNAm data must have had the same number of prior variances ( $n = 3$  each). Mixture variances for SNP data were set to 0.01, 0.1, and 0.2 in combined analyses. Input data were scaled to mean zero and unit variance. Gibbs sampling was used to sample over the posterior distribution conditional on input data and 10000 samples were used. The first 5000 samples of burn-in were removed and a thinning of five samples was applied to reduce autocorrelation. SNPs that exhibited a posterior inclusion probability (PIP)  $\geq 95\%$  were deemed significant.

### 2.5 | EWAS

Blood DNAm in Generation Scotland participants was quantified using the Illumina HumanMethylationEPIC BeadChip Array. Blood DNAm was assessed in two separate sets. After quality control, 793706 and 773860 CpG remained in sets 1 and 2, respectively. In total, 772619 CpG sites were shared across sets. Each set was truncated to these overlapping probes.

In the stand-alone EWAS and combined GWAS/EWAS, mixture variances were set to 0.001, 0.01, and 0.1 ( $n = 778$ ). Missing DNAm data were mean imputed separately within each set as BayesR+ cannot accept missing values. Both sets were combined and adjusted for DNAm batch, set, age, and sex. Each CpG site was scaled to mean zero and unit variance. Houseman-estimated white blood cell proportions were included as fixed-effect covariates.<sup>31</sup> CpG sites that had a PIP  $\geq 95\%$  were deemed significant.

Sensitivity EWAS analyses were performed using linear mixed-effects models and the *lme4* function from the R *coxme* package (version 2.2-16).<sup>32</sup> DNAm data were pre-corrected for age, sex, batch, and set. Houseman-estimated white blood cell proportions were incorporated as fixed-effect covariates and a kinship matrix was fitted to account for relatedness among individuals in STRADL.

## 2.6 | Co-localization analyses

Formal Bayesian tests of co-localization were used to determine whether a shared causal variant likely underpinned two traits of interest.<sup>33</sup> A 200 kilobase (kb) region (upstream and downstream, recommended default setting) surrounding the variant was extracted from our GWAS summary statistics.

Expression quantitative trait loci (eQTL) data were extracted from eQTLGen summary statistics. Methylation QTL (mQTL) summary statistics were extracted from phenoscanner, GoDMC, or our own mQTL analyses. Methylation QTL analyses were performed using additive linear regression models and by regressing CpG sites (beta values) on SNPs (0, 1, 2) while adjusting for age, sex, DNAm batch, set, Houseman-estimated white blood cell proportions, and 20 genetic PCs ( $n = 778$ ). In instances where an mQTL effect was identified in more than one database, summary statistics from the study with the largest sample size were used in *coloc*.<sup>34-36</sup> For AD-related traits, summary statistics were extracted from the relevant GWAS.<sup>3,37-39</sup> Default priors were applied. Summary statistics for all SNPs ( $\pm 200$  kb from the queried SNP) were used to estimate the posterior probability for five separate hypotheses: a single variant underlying both traits, separate variants for both traits, a causal variant for one trait (encompassing two hypotheses), or no causal variant for either trait. Posterior probabilities  $\geq 95\%$  provided strong evidence for a given hypothesis.

## 2.7 | Mendelian randomization (MR)

Bidirectional Mendelian randomization (MR) was used to test for associations between (1) gene expression and plasma protein levels, (2) DNAm and plasma protein levels, and (3) plasma protein levels and AD risk or related biomarkers. Pruned variants ( $r^2 < 0.1$ ) were used as instrumental variables (IVs) in MR analyses. Analyses were conducted using MR-base.<sup>40</sup> Two-sample MR was applied and relationships were assessed using the Wald ratio test. Further information on IVs used are provided in supplementary information.

# 3 | RESULTS

## 3.1 | Identification of plasma proteins associated with AD

Twelve studies were identified that reported associations between SOMAscan plasma proteins and AD or related traits (Figure 1). Three hundred fifty-nine unique proteins were identified and 22 (6.1%) were reported in more than one study (Table S2-S4). In the Generation

Scotland dataset, there were 308 SOMAmers (Slow Off-rate Modified Aptamers) that targeted 282 of 359 proteins of interest (Table S5 and Figure S1). The 282 unique proteins were brought forward for analyses (UniProt IDs and Seq-ids are shown in Table S6).

## 3.2 | GWAS on AD-associated proteins

There were 1064 individuals with genotype and proteomic data in Generation Scotland. The mean age of the sample was 59.9 years (standard deviation [SD] = 5.9) and 59.1% of the sample was female. In the BayesR+ GWAS, 64 independent variants (or protein quantitative trait loci, pQTLs) were associated with 41 SOMAmers that mapped to 39 unique protein targets (PIP  $\geq 95\%$ ; Table S7). The phenotypic correlation structure of these 41 SOMAmers is presented in Figure S2. The median correlation coefficient between SOMAmer levels was 0.18. Thirty-six pQTLs represented *cis* associations (pQTLs within 10 megabases [Mb] of transcription start site [TSS] for a given gene) and 28 pQTLs were *trans*-chromosomal effects (Figure 2). The majority of variants were located in intronic regions using annotations from the ENSEMBL variant effect predictor (46.9%, Figure S3).

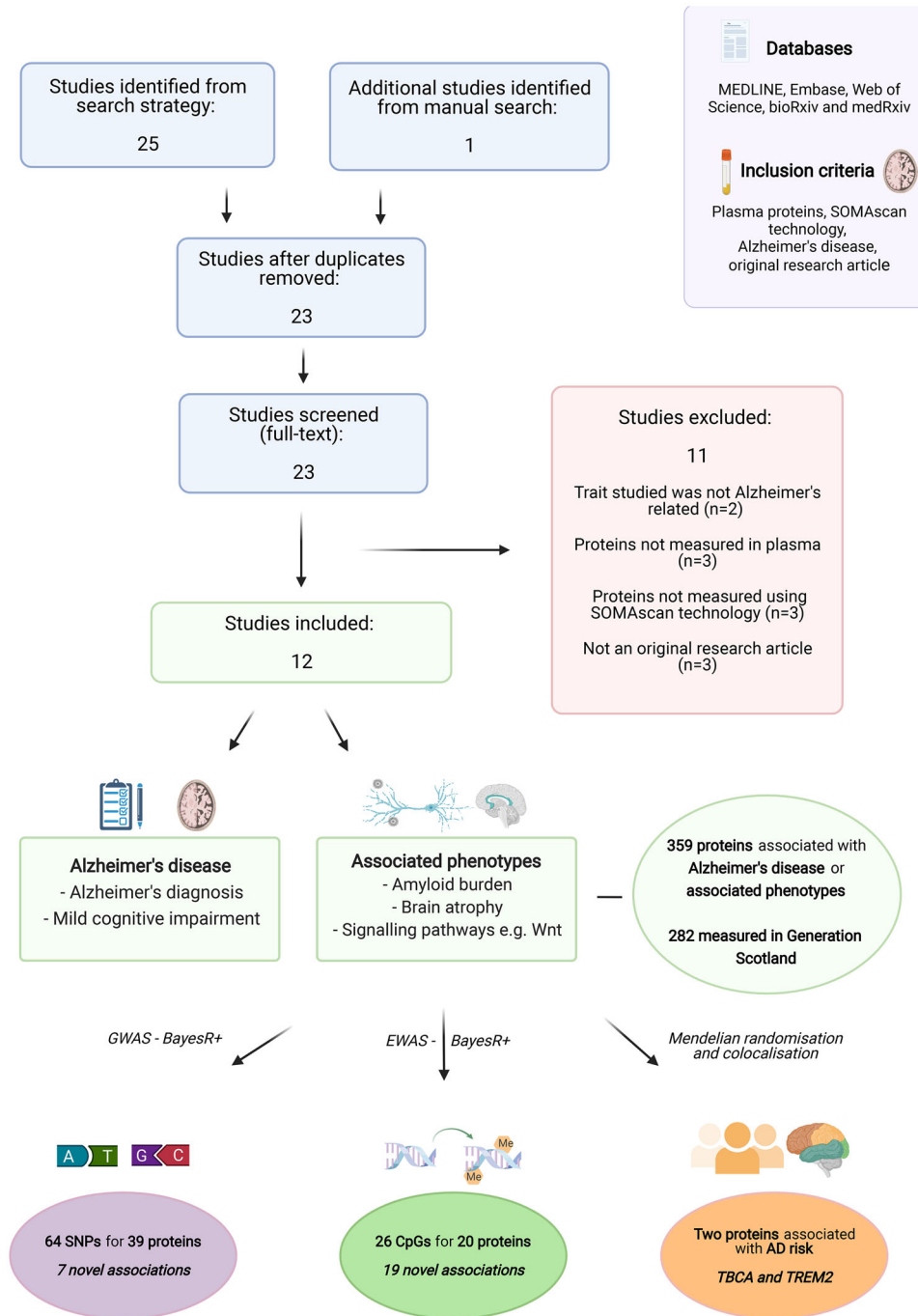
Fifty-seven pQTLs were previously reported in GWAS of blood protein levels (Table S8). Variants either directly replicated known associations or showed high linkage disequilibrium (LD,  $r^2 > 0.75$ ) with known pQTLs for queried proteins. Relative effect sizes reported in the literature correlated strongly with those in our study ( $r = 0.77$ , 95% confidence interval [CI] = 0.66, 0.84). We identified seven novel pQTLs associated with seven unique proteins. Three pQTLs were in *cis* (for GM2A, MATN3 and IL1RAP). Four pQTLs represented *trans*-chromosomal effects: rs1126680 (*BCHE* for *KLK6*), rs7867739 (near *ABO* for *ALPI*), rs3820897 (*COLEC11* for *ALPL*), and rs1530914 (*MS4A4A* for triggering receptor expressed on myeloid cells 2 [*TREM2*]).

Thirty-three pQTLs were associated with at least one trait in the GWAS Catalog at  $P < 5 \times 10^{-8}$  (range = 1 to 96 associations; Table S9).<sup>41</sup> In relation to AD traits, the *trans* pQTL in *MS4A4A* (rs1530914) for *TREM2* levels is in high LD with a *TREM2* variant (rs1582763,  $r^2 \sim 0.9$ ) associated with AD in apolipoprotein E (*APOE*)  $\epsilon 4$  carriers and family history of AD.<sup>3,42</sup> In addition, the *trans* pQTL in *APOE* (rs769449) for tubulin-specific chaperone A (*TBCA*) levels was associated with 15 AD-related traits including genetic predisposition to AD and CSF biomarkers of the disease.

BayesR+ was used to estimate the proportions of inter-individual variation in plasma protein levels that were attributable to common SNPs (minor allele frequency  $> 1\%$ ). Estimates ranged from 5.3% (PRL; 95% credible interval [CrI] = 0%, 24.4%) to 73.0% (IL1RAP; 95% CrI = 56.0%, 83.0%), with a median estimate of 13.0% across all 308 SOMAmers (Table S10).

## 3.3 | Co-localization of protein QTLs with expression QTLs

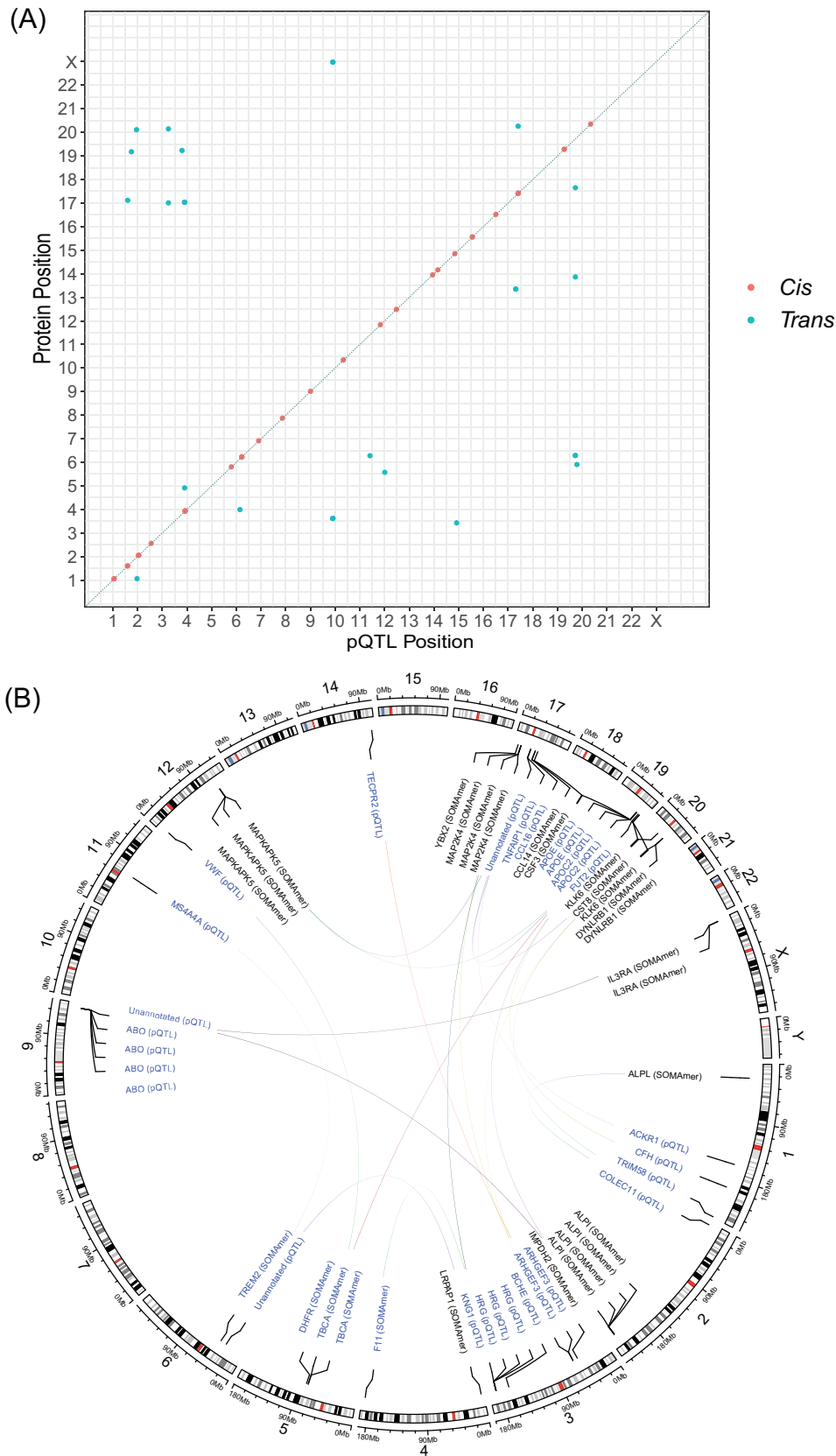
The 36 *cis* pQTLs identified in BayesR+ were annotated to 23 unique proteins. For 12 of 23 proteins, at least one pQTL was previously



**FIGURE 1** Structured literature review of SOMAscan plasma proteins that were associated with AD in the literature, and assessment of their molecular architectures and relationships with AD in the present study. The MEDLINE, Embase, Web of Science databases, and preprint servers were queried to identify studies that reported associations between SOMAscan-measured plasma proteins and AD. GWAS, EWAS, and causal inference analyses were performed to identify molecular correlates of 282 AD-associated plasma protein levels and to probe their associations with AD and related traits. AD, Alzheimer's disease; EWAS, epigenome-wide association studies; GWAS, genome-wide association studies. Figure created using Biorender.com

reported to be an expression QTL for the respective gene in blood tissue (eQTL consortium database).<sup>34</sup> The R package *coloc*<sup>33</sup> was used to test the hypothesis that a single variant associates with differences in gene expression (eQTL) and protein levels (pQTL) for each gene of interest. For two proteins (PCSK7 and F7), there was strong evidence (posterior probability [PP] >95%) for a shared variant underlying

gene expression and protein levels (Table S11). MR analyses provided evidence for reciprocal associations between changes in gene expression and circulating levels of these proteins (Table S12). Three proteins had weaker evidence for co-localization ( $PP \geq 75\%$  for GM2A, LYZ, PDCD1LG2) and seven proteins had strong evidence for separate variants underlying gene expression and protein levels.



**FIGURE 2** GWAS on plasma protein levels previously associated with AD and disease-related phenotypes. (A) Chromosomal locations of pQTLs identified through Bayesian penalized regression GWAS. The x-axis shows the chromosomal location of pQTLs associated with the levels of SOMAers that correlate with AD status or related pathways. The y-axis represents the position of the gene encoding the target protein. *Cis* (red circles); *trans* (blue circles). (B) A circos plot for the 28 *trans*-associated pQTLs from (A). Lines indicate an association between a pQTL and SOMAer. AD, Alzheimer's disease; GWAS, genome-wide association studies; pQTL, protein quantitative trait locus

### 3.4 | EWAS on AD-associated proteins

There were 778 individuals with DNAm and proteomic data in the Generation Scotland sample. The mean age of the sample was 60.2 (SD = 8.8) years and 56.4% of the sample were female. Twenty-six CpGs were associated with the levels of 20 unique proteins (PIP >95%, Table S13 and Figure S4). The median correlation coefficient between measured protein levels was 0.16. The associations consisted of 10 *cis* CpG sites and 16 *trans* CpG loci (Figure 3). The cg07839457 probe in the *NLR5* locus was associated with IL18BP and CSF1R levels, and the smoking-associated probe cg05575921 in *AHRR* was associated with GHR, PIGR, and WFDC2 levels.

We used linear mixed-effects models that accounted for relatedness to perform sensitivity analyses for the 26 CpG associations identified in BayesR+ (Table S14).<sup>32</sup> Effect sizes were highly correlated with those from BayesR+ and showed full directional concordance ( $r = 0.95$ , 95% CI = 0.90, 0.98; Figure S5). Twenty-one associations were replicated at a genome-wide significance threshold of  $P < 3.6 \times 10^{-8}$ , and the remaining five associations were replicated at  $P < 2.0 \times 10^{-3}$ . Furthermore, 7 of 26 CpG associations were previously reported in the literature and relative effect sizes correlated strongly with those in our study ( $r = 0.98$ , 95% CI = 0.87, 1.0). The 19 novel CpG sites were associated with levels of 14 unique proteins.

In BayesR+, estimates for the proportions of variability in SOMAmer levels that could be accounted for by DNAm measured on the EPIC BeadChip array ranged from 7.1% (EEA1; 95% CrI = 0%, 27.7%) to 33.8% (MAP kinase-activated protein kinase, MAPKAPK5; 95% CrI = 22.6%, 47.0%), with a median estimate of 10.0% (Table S15).

Estimates for variance in SOMAmer levels accounted for by genetic and methylation data together, while conditioned on each other, ranged from 21.8% for *ENTPD1* (95% CrI = 0.0%, 59.1%) to 93.3% for *GHR* (95% CrI = 80.1%, 100%) (Table S16). The mean and median estimates were 48.7% and 46.8%, respectively.

### 3.5 | Co-localization of protein QTLs with methylation QTLs

Fourteen proteins had both genome-wide significant pQTL and CpG associations in our study. There were 39 possible SNP-CpG pairs across these proteins. For each pair, we used linear regression to test if the SNP was associated with CpG methylation at  $P < 5 \times 10^{-8}$ , thereby representing an mQTL effect (Table S17). We also performed look-up analyses of mQTL databases including GoDMC and phenoscanner.<sup>35,36</sup> In instances where an mQTL effect was identified in more than one database, coefficients from the study with the largest sample size were brought forward for co-localization analyses. In addition, in instances where two or more mQTLs were associated with the same CpG site in a given locus, only the most significant mQTL was brought forward for co-localization analyses ( $n = 19$  mQTLs, 13 proteins; Table S18).

For six proteins, we observed strong evidence in *coloc* that a single *cis*-acting variant might underpin differential DNAm levels and protein abundances (PP >95%, Table S19). The six proteins were ANXA2, F7,

MATN3, PCSK7, PLA2G2A, and SERPINA3. MR analyses provided evidence that relationships between methylation and protein levels were bidirectional (Table S20).

### 3.6 | MR analyses between plasma proteins and AD risk

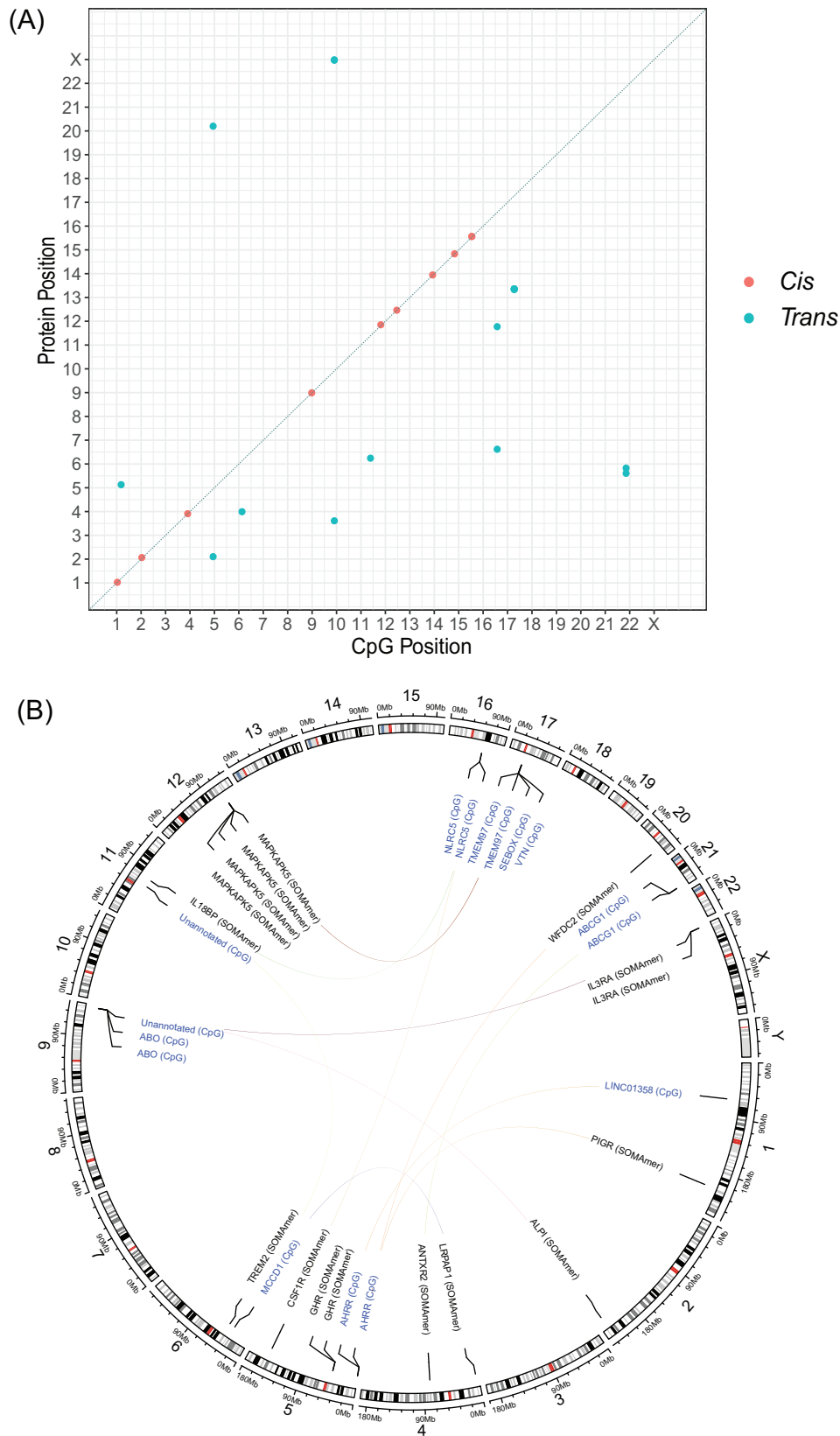
Bidirectional MR was applied to test for associations between the 41 SOMAmers with pQTL associations in BayesR+ and 20 AD-related traits (Table S21). A Bonferroni-corrected threshold of  $P < 6.10 \times 10^{-5}$  ( $< 0.05/41 \times 20$ ) was set. Plasma levels of three proteins had a unidirectional association with AD risk: TREM2 (Table 1, Wald ratio test,  $\beta = -0.13$ , SE = 0.05,  $P = 8.4 \times 10^{-17}$ ), colony stimulating factor 3 (CSF3) (Wald ratio test,  $\beta = 0.10$ , SE = 0.02,  $P = 5.9 \times 10^{-6}$ ), and TBCA (inverse variance-weighted method,  $\beta = -0.50$ , SE = 0.12,  $P = 1.2 \times 10^{-5}$ ). Conversely, AD risk was not associated with plasma levels of these proteins. Co-localization analyses suggested that one variant was associated with TREM2 or TBCA levels and AD risk, and two separate variants were associated with CSF3 levels and AD risk (Table S22).

## 4 | DISCUSSION

We identified seven novel protein QTLs and 19 novel CpG sites that associated with plasma levels of 18 AD-related proteins. Using BayesR+, we provided estimates for associations between common genetic and DNAm variation and inter-individual differences in plasma levels of 282 AD-related proteins. We integrated our data with publicly available gene expression and methylation QTL databases and highlighted molecular mechanisms that might link pQTLs to differential levels of six proteins. We observed strong associations between plasma levels of TREM2 or TBCA and AD risk. These associations were driven by *trans* pQTLs in *MS4A4A* and *APOE*, respectively.

We show that a *trans* pQTL (rs1530914) in the *MS4A4A* locus associates with higher plasma TREM2 levels. It is in strong LD ( $r^2 \sim 0.9$ ) with the variant rs1582763, which has been associated with higher CSF TREM2 levels and lower AD risk.<sup>3,43</sup> It is also in moderate LD ( $r^2 = 0.6$ ) with a variant in the 3'UTR region of *MS4A6A* (rs610932), which was associated with plasma TREM2 levels in a sample of 35,559 Icelanders.<sup>11</sup> Polymorphisms in *MS4A4A* were shown to alter *MS4A4A* expression and subsequently modulate TREM2 concentration in human macrophages.<sup>44</sup> We supplement existing data by identifying a novel blood CpG correlate of plasma TREM2 levels (cg02521229) located near *MS4A4A* that previously associated with dementia risk in Generation Scotland participants.<sup>45</sup> Our data suggest that risk mechanisms arising from *MS4A4A* polymorphisms and TREM2 levels can be captured in plasma assays and that these mechanisms involve differential methylation in the *MS4A4A* locus.

We observed associations between plasma levels of three proteins (CSF3, MAPKAPK5, and TBCA) and *trans* pQTLs in the *TOMM40-APOE-APOC2* locus. Furthermore, we identified two pQTLs and three CpG



**FIGURE 3** EWAS on plasma protein levels previously associated with AD and disease-related phenotypes. (A) Chromosomal locations of CpGs identified through Bayesian penalized regression EWAS. The x-axis shows the chromosomal location of CpG sites and the y-axis represents the position of the gene encoding the target protein. *Cis* (red circles); *trans* (blue circles). (B) A circos plot for the 16 *trans*-associated CpGs from (A). Lines indicate an association between a CpG site and SOMAmer. AD, Alzheimer's disease; CpG, cytosine-phosphate-guanin; EWAS, epigenome-wide association studies



**TABLE 1** MR analyses of plasma protein levels and AD-associated traits (Bonferroni-corrected  $P < 6.10 \times 10^{-5}$ )

Protein	Trait	Method	Beta	SE	P	Reference
<i>Protein levels affecting AD-associated traits</i>						
TBCA	Log-transformed CSF A $\beta$ 42	IVW	-0.09	0.01	$2.5 \times 10^{-17}$	38
TREM2	AD risk	Wald ratio	-0.13	0.02	$8.4 \times 10^{-17}$	3
TBCA	CSF APOE	Wald ratio	0.75	0.10	$7.3 \times 10^{-14}$	39
TBCA	CSF A $\beta$ (Z-scores)	IVW	-0.45	0.06	$2.1 \times 10^{-13}$	38
TBCA	Log-transformed CSF A $\beta$ 42/A $\beta$ 40	IVW	-0.08	0.01	$6.9 \times 10^{-10}$	38
CSF3	AD risk	Wald ratio	0.10	0.02	$5.9 \times 10^{-6}$	3
TBCA	AD risk	IVW	-0.50	0.12	$1.2 \times 10^{-5}$	3
<i>AD-associated traits affecting protein levels</i>						
TBCA	Log-transformed CSF A $\beta$ 42	Wald ratio	-11.14	0.53	$4.4 \times 10^{-98}$	38
TBCA	CSF A $\beta$ (Z-scores)	Wald ratio	-2.13	0.10	$5.7 \times 10^{-98}$	38
TBCA	Log-transformed CSF A $\beta$ 42/A $\beta$ 40	Wald ratio	-11.13	0.53	$5.7 \times 10^{-98}$	38
TBCA	CSF A $\beta$	Wald ratio	12.21	0.63	$3.7 \times 10^{-84}$	37

Abbreviations: CSF, cerebrospinal fluid; IVW, inverse variance-weighted method; MR, mendelian randomization; SE, standard error.

correlates of plasma MAPKAPK5 levels near the transmembrane protein 9 (*TMEM97*) locus. MAPKAPK5 correlated with cognitive decline in the Twins UK cohort; however, its relationship with neuropathology is unknown.<sup>22</sup> *TMEM97* acts a synaptic receptor for A $\beta$  and mediates its cellular uptake via APOE-dependent and APOE-independent mechanisms.<sup>46</sup> Given that *TMEM97* polymorphisms may influence MAPKAPK5 levels, our data prioritize MAPKAPK5 for follow-up studies as a potential downstream effector or correlate of *TMEM97* in A $\beta$  clearance. TBCA correlates with A $\beta$  burden.<sup>16</sup> TBCA levels are higher in individuals with the protective APOE  $\epsilon$ 2/ $\epsilon$ 2 genotype and lower in carriers of the risk  $\epsilon$ 4 polymorphism.<sup>47</sup> These data are consistent with our GWAS and MR analyses. Future studies should examine whether TBCA dysregulation is a cause or consequence of disease risk mechanisms in carriers of APOE  $\epsilon$ 4 polymorphisms.

Our study has a number of limitations. First, our review does not reflect an exhaustive list of potential AD-associated traits. Furthermore, there is heterogeneity across studies in terms of diagnostic criteria and phenotype definitions. Second, by focusing on the SOMAscan platform alone, we do not capture all blood protein correlates of AD that are reported in the literature. Third, an insufficient number of variants were available to test for horizontal pleiotropy in MR analyses. Fourth, it is important to note that variants may alter SOMAscan reactivity with protein targets, or reflect technical artifacts such as sample handling and cross-reactive events. Fifth, our sample consisted of Scottish individuals with a relatively homogenous genetic background thereby limiting generalizability of findings.

## 5 | CONCLUSIONS

Our strategy of integrating multiple omics measures determined the degree to which molecular factors can explain inter-individual differ-

ences in blood levels of possible biomarkers for AD, and advanced understanding of mechanisms underlying AD risk.

## ACKNOWLEDGMENTS

This research was funded in whole, or in part, by Wellcome [108890/Z/15/Z, 104036/Z/14/Z]. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. The authors are grateful to the families who took part in this study, the general practitioners, and the Scottish School of Primary Care for their help in recruiting them and the wider Generation Scotland team. Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping and DNA methylation profiling of the Generation Scotland samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland, and was funded by the Medical Research Council (MRC) UK and the Wellcome Trust (Wellcome Trust Strategic Award "Stratifying Resilience and Depression Longitudinally" ([STRADL] Reference [104036/Z/14/Z]). Andrew M. McIntosh is supported by Wellcome [104036/Z/14/Z, 216767/Z/19/Z, 220857/Z/20/Z], United Kingdom Research and Innovation (UKRI) MRC [MC\_PC\_17209, MR/S035818/1] and the European Union H2020 [SEP-210574971]. Ian J. Deary received support from Age UK, Wellcome, and the Medical Research Council. David J. Porteous is supported by Wellcome as principal investigator (PI), and MRC and National Institute for Health Research (NIHR) grants as co-PI, made to the University of Edinburgh. Robert F. Hillary and Danni A. Gadd are supported by funding from the Wellcome 4-year PhD in Translational Neuroscience—training the next generation of basic neuroscientists to embrace clinical research [108890/Z/15/Z]. Daniel L. McCartney and Riccardo E. Marioni are supported by Alzheimer's

Research UK major project grant ARUK-PG2017B-10. Riccardo E. Marioni is supported by Alzheimer's Society major project grant AS-PG-19b-010. Proteomic analyses in STRADL were supported by Dementias Platform UK (DPUK). DPUK funded this work through core grant support from the Medical Research Council [MR/L023784/2]. Kathryn L. Evans was supported by a grant from Alzheimer's Research UK, paid to the University of Edinburgh. Alejo J. Nevado-Holgado was funded by a Horizon 2020 Virtual Brain Cloud project (H2020-SC1-DTH-2018-1), in addition to funding from the MRC, UK Rosetrees, and King Abdullah University of Science and Technology, Saudi Arabia. Caroline Hayward is supported by an MRC University Unit Programme Grant MC\_UU\_00007/10 (QTL in Health and Disease). Liu Shi is funded by DPUK through MRC [MR/L023784/2] and the UK Medical Research Council Award to the University of Oxford [MC\_PC\_17215]. Liu Shi received support from the NIHR Biomedical Research Centre at Oxford Health NHS Foundation Trust. Matthew R. Robinson is funded by a Swiss National Science Foundation Eccellenza Grant [PCEGP3-181181].

### CONFLICT OF INTEREST

Andrew M. McIntosh has received research support from Eli Lilly, Janssen, and the Sackler Foundation. Andrew M. McIntosh has also received speaker fees from Illumina and Janssen and consulting fees. Simon Lovestone is currently an employee of Johnson & Johnson Medical Ltd and previously received grant support from multiple pharmaceutical companies and the EU through the Innovation Medicines Initiative programmes European Medical Information Framework and European Prevention of Alzheimer's Dementia (EPAD). Simon Lovestone is also co-founder of Akriveria Health. Alejo J. Nevado-Holgado has received funding from GlaxoSmithKline, UK, Ono Pharma, Japan, and Johnson & Johnson, UK. David J. Porteous is PI on a grant from Wellcome made to the University of Edinburgh with travel allowance. Craig W. Ritchie has been a paid consultant for several companies developing treatments for Alzheimer's disease over the last 5 years including Biogen, Eli Lilly, Merck, Roche, Janssen, AbbVie, Kyowa Kirin, Actinogen, and Eisai. Craig W. Ritchie was the UK Chief Investigator for the ENGAGE Trial and Academic Lead on the EPAD Programme, which was a public-private partnership between the EU and several companies with an interest in developing treatments for AD ([www.ep-ad.org](http://www.ep-ad.org)). Craig W. Ritchie's unit at the University of Edinburgh (Edinburgh Dementia Prevention) has received grant funding from Biogen, Janssen, AC Immune, and Actinogen; he is the unpaid chairperson of the Brain Health Clinic Consortium established in the UK by Biogen; was a member of a Data Safety Monitoring Board (DSMB) for a University College London (UCL) sponsored study (no payment); and serves as Director of Brain Health Scotland and Chair of the Scottish Dementia Research Consortium. Ian J. Deary received royalties from Oxford University Press and Cambridge University Press. Archie Campbell was a member of the EMREC Edinburgh Medical Research Ethics Committee (no payment involved). Riccardo E. Marioni has received speaker fees from Illumina and is an advisor to the Epigenetic Clock Development Founda-

tion. The remaining authors declare that they have no competing interests.

### ETHICS STATEMENT

All components of the Generation Scotland study received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Numbers: 05/S1401/89 and 10/S1402/20). All participants provided broad and enduring written informed consent for biomedical research. Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20-ES-0021). This study was performed in accordance with the Declaration of Helsinki.

### DATA AVAILABILITY STATEMENT

According to the terms of consent for Generation Scotland participants, access to data must be reviewed by the Generation Scotland Access Committee. Applications should be made to [access@generationscotland.org](mailto:access@generationscotland.org). Full and openly accessible summary statistics from GWAS and EWAS on SOMAmers levels are available on the University of Edinburgh Datashare site.<sup>48,49</sup>

### CODE AVAILABILITY

All code is available at the following Github repository.<sup>50</sup>

### ORCID

Robert F. Hillary  <https://orcid.org/0000-0002-2595-552X>

### REFERENCES

- 2020 GHE. *Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019*. World Health Organization; 2020.
- DALYs GBD, Collaborators H. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet (London, England)*. 2017;390(10100):1260-1344.
- Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51(3):404-413.
- Wightman DP, Jansen IE, Savage JE, et al. Largest GWAS (N = 1,126,563) of Alzheimer's disease implicates microglia and immune cells. *medRxiv*. 2020;53(9):1276-1282.
- Olsson B, Lautner R, Andreasson U, et al. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *Lancet Neurol*. 2016;15(7):673-684.
- Lista S, Faltraco F, Prvulovic D, Hampel H. Blood and plasma-based proteomic biomarker research in Alzheimer's disease. *Prog Neurobiol*. 2013;101-102:1-17.
- Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*. 2005;102(30):10604-10609.
- Bestor TH, Edwards JR, Boulard M. Notes on the role of dynamic DNA methylation in mammalian development. *Proc Natl Acad Sci U S A*. 2015;112(22):6796-6799.
- Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun*. 2017;8:14357.

10. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018;558(7708):73-79.
11. Ferkingstad E, Sulem P, Atlason BA, et al. Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics*. 2021;53(12):1712-1721.
12. Zaghlool SB, Kühnel B, Elhadad MA, et al. Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nat Commun*. 2020;11(1):1-12.
13. Ahsan M, Ek WE, Rask-Andersen M, et al. The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. *PLoS Genet*. 2017;13(9):e1007005.
14. Hillary RF, McCartney DL, Harris SE, et al. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. *Nat Commun*. 2019;10(1):3160.
15. Hillary RF, Trejo-Banos D, Kousathanas A, et al. Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults. *Genome Med*. 2020;12(1):60.
16. Shi L, Westwood S, Baird AL, et al. Discovery and validation of plasma proteomic biomarkers relating to brain amyloid burden by SOMAscan assay. *Alzheimers Dement*. 2019;15(11):1478-1488.
17. Shi L, Winchester LM, Liu BY, et al. Dickkopf-1 overexpression in vitro nominates candidate blood biomarkers relating to Alzheimer's disease pathology. *J Alzheimers Dis*. 2020;77:1353-1368.
18. Tanaka T, Lavery R, Varma V, et al. Plasma proteomic signatures predict dementia and cognitive impairment. *Alzheimers Dement (N Y)*. 2020;6(1):e12018.
19. Sattlecker M, Kiddle SJ, Newhouse S, et al. Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimers Dement*. 2014;10(6):724-734.
20. Sattlecker M, Khondoker M, Proitsi P, et al. Longitudinal protein changes in blood plasma associated with the rate of cognitive decline in Alzheimer's disease. *J Alzheimers Dis*. 2016;49(4):1105-1114.
21. Begic E, Hadzidedic S, Kulagic A, Ramic-Brkic B, Begic Z, Causevic M. SOMAscan-based proteomic measurements of plasma brain natriuretic peptide are decreased in mild cognitive impairment and in Alzheimer's dementia patients. *PLoS one*. 2019;14(2):e0212261.
22. Kiddle SJ, Steves CJ, Mehta M, et al. Plasma protein biomarkers of Alzheimer's disease endophenotypes in asymptomatic older twins: early cognitive decline and regional brain volumes. *Transl Psychiatry*. 2015;5(6):e584.
23. Begic E, Hadzidedic S, Obradovic S, Begic Z, Causevic M. Increased levels of coagulation factor XI in plasma are related to Alzheimer's disease diagnosis. *J Alzheimers Dis*. 2020;77(1):375-386.
24. Shi L, Winchester LM, Westwood S, et al. Replication study of plasma proteins relating to Alzheimer's pathology. *Alzheimers Dement*. 2021;17(9):1452-1464.
25. Walker KA, Chen J, Zhang J, et al. Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nature Aging*. 2021;1(5):473-489.
26. Kiddle SJ, Sattlecker M, Proitsi P, et al. Candidate blood proteome markers of Alzheimer's disease onset and progression: a systematic review and replication study. *J Alzheimers Dis*. 2014;38(3):515-531.
27. Lindbohm JV, Mars N, Walker KA, et al. Plasma proteins, cognitive decline, and 20-year risk of dementia in the Whitehall II and Atherosclerosis Risk in Communities studies. *Alzheimers Dement*. 2021.
28. Trejo Banos D, McCartney DL, Patxot M, et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nat Commun*. 2020;11(1):2865.
29. Habota T, Sandu A, Waiter G, et al. Cohort profile for the STRatifying Resilience and Depression Longitudinally (STRADL) study: a depression-focused investigation of Generation Scotland, using detailed clinical, cognitive, and neuroimaging assessments [version 1; peer review: 1 approved, 1 not approved]. *Wellcome Open Res*. 2019;4(185).
30. Navrady LB, Wolters MK, MacIntyre DJ, et al. Cohort profile: stratifying resilience and depression longitudinally (STRADL): a questionnaire follow-up of generation Scotland: Scottish Family Health Study (GS:sFHS). *Int J Epidemiol* 2017;47(1):13-14g.
31. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform*. 2012;13:86.
32. Therneau TM, coxme: Mixed Effects Cox Models. R package version 2.2-16. 2020.
33. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. 2014;10(5):e1004383.
34. Vösa U, Claringbould A, Westra H-J, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta analysis. *bioRxiv*. 2018:447367.
35. Staley JR, Blackshaw J, Kamat MA, et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*. 2016;32(20):3207-3209.
36. Min JL, Hemani G, Hannon E, et al. Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. *medRxiv*. 2020;53(9):1311-1321.
37. Deming Y, Li Z, Kapoor M, et al. Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers. *Acta Neuropathol*. 2017;133(5):839-856.
38. Hong S, Prokopenko D, Dobricic V, et al. Genome-wide association study of Alzheimer's disease CSF biomarkers in the EMIF-AD Multimodal Biomarker Discovery dataset. *Transl Psychiatry*. 2020;10(1):403.
39. Kauwe JS, Bailey MH, Ridge PG, et al. Genome-wide association study of CSF levels of 59 Alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. *PLoS Genet*. 2014;10(10):e1004758.
40. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife*. 2018;7:e34408.
41. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012.
42. Jun G, Ibrahim-Verbaas CA, Vronskaya M, et al. A novel Alzheimer disease locus located near the gene encoding tau protein. *Mol Psychiatry*. 2016;21(1):108-117.
43. Liu C, Yu J. Genome-wide association studies for cerebrospinal fluid soluble TREM2 in Alzheimer's disease. *Front Aging Neurosci*. 2019;11:297.
44. Deming Y, Filipello F, Cignarella F, et al. The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Sci Transl Med*. 2019;11(505):eaau2291.
45. Walker RM, Bermingham ML, Vaher K, et al. Epigenome-wide analyses identify DNA methylation signatures of dementia risk. *Alzheimers Dement (Amst)*. 2020;12(1):e12078.
46. Colom-Cadena M, Tulloch J, Rose J, Smith C, Spires-Jones T. TMEM97 is a potential amyloid beta receptor in human Alzheimer's disease synapses. *AlzheimersDement*. 2020;16(S2):e041782.
47. Sebastiani P, Monti S, Morris M, et al. A serum protein signature of APOE genotypes in centenarians. *Aging Cell*. 2019;18(6):e13023.
48. Hillary RF, Gadd DA, McCartney DL, et al. BayesR+ GWAS on Alzheimer's disease-associated SOMAmers. *Edinburgh Datashare*. 2021. Accessed: December 3, 2021. <https://datashare.ed.ac.uk/handle/10283/4095>
49. Hillary RF, Gadd DA, McCartney DL, et al. BayesR+ EWAS on Alzheimer's disease-associated SOMAmers. *Edinburgh Datashare*. 2021. Accessed: December 3, 2021. <https://datashare.ed.ac.uk/handle/10283/4096>

50. Hillary RF, GitHub repository. 2021. GitHub. Accessed: December 3, 2021. [https://github.com/robertfhillary/gwas\\_ewas\\_AD\\_plasma\\_biomarkers](https://github.com/robertfhillary/gwas_ewas_AD_plasma_biomarkers)

#### SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Hillary RF, Gadd DA, McCartney DL, et al. Genome- and epigenome-wide studies of plasma protein biomarkers for Alzheimer's disease implicate TBCA and TREM2 in disease risk. *Alzheimer's Dement*. 2022;14:e12280. <https://doi.org/10.1002/dad2.12280>