Epigenome and phenome study reveals circulating markers pertinent to brain health

Danni A Gadd[1,], Robert F Hillary[1], Daniel L McCartney[1], Liu Shi[2], Robert I McGeachan[3], Aleks Stolicyn[4], Stewart W Morris[1], Rosie M Walker[5], Archie Campbell[1], Miruna C Barbu[4], Mathew A Harris[4], Ellen V Backhouse[5], Joanna M Wardlaw[5,6,7,8], J Douglas Steele[9], Diego A Oyarzún[10], Graciela Muniz-Terrera[12], Craig Ritchie[12], Alejo Nevado-Holgado[2], Caroline Hayward[1,13], Kathryn L Evans[1], David J Porteous[1], Simon R Cox[14,15], Heather C Whalley[4], Andrew M McIntosh[4], Riccardo E Marioni[1,†]

[1] Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU

[2] Department of Psychiatry, University of Oxford, UK, OX3 7JX

[3] Centre for Discovery Brain Sciences, University of Edinburgh, 1 George Square, Edinburgh EH8 9JZ, UK

[4] Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, EH10 5HF, UK

[5] Centre for Clinical Brain Sciences, Chancellor's Building, 49 Little France Crescent, Edinburgh BioQuarter, Edinburgh, EH16 4SB

[6] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

[7] Edinburgh Imaging, University of Edinburgh, Edinburgh, UK

[8] UK Dementia Research Institute, University of Edinburgh, Edinburgh, UK, EH8 9JZ

[9] Department of Mental Health, University of Aberdeen, Block A, Royal Cornhill Hospital, Aberdeen, UK

[10] School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB

[11] School of Biological Sciences, University of Edinburgh, Edinburgh, EH3 3JF

[12] Centre for Clinical Brain Sciences, Edinburgh Dementia Prevention, University of Edinburgh, Edinburgh, UK, EH4 2XU

[13] Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU

[14] Lothian Birth Cohorts, University of Edinburgh, Edinburgh, EH8 9JZ

[15] Department of Psychology, University of Edinburgh, Edinburgh, EH8 9JZ

[†] Corresponding author: Riccardo Marioni, riccardo.marioni@ed.ac.uk

Abstract

Characterising associations between the epigenome, proteome and phenome may provide insight into molecular regulation of biological pathways governing health. However, epigenetic signatures for many neurologically-associated plasma protein markers remain uncharacterised. Here, we report an epigenome and phenome-wide association study of the circulating proteome in relation to brain health. We perform epigenome-wide studies of 4,235 plasma proteins (n=778), identifying 2,895 CpG-protein associations (protein quantitative trait methylation loci; pQTMs) after stringent correction for multiple testing. These were independent of known genetic protein quantitative trait loci (pQTL) and common lifestyle effects, extending current knowledge by analysing a further 3,286 protein measurements with 2,854 novel pQTMs. We then perform a phenome-wide study of each protein in relation to neurological traits in 1,065 individuals, identifying 644 proteins related to cognitive, brain imaging phenotypes or *APOE* status. By integrating our pQTM dataset with our phenome association study, we uncovered 88 epigenetic associations for protein markers of neurological traits, 83 of which were previously unreported. These data are pertinent to understanding heterogeneity in brain health.

Introduction

Multiple layers of omics data indicate the biological pathways that underlie disease risk. Epigenetic modifications account for inter-individual variability in circulating protein levels [1–3] and have been linked to a range of neurological outcomes [4–6] and neurodegenerative conditions such as Alzheimer's disease [7]. Proteome-wide characterisation of blood protein signatures has also been facilitated at large-scale by SOMAscan® protein measurements, with studies that have begun to map plasma protein signatures of cognitive decline and dementia risk [8–10]. There is, however, a need to further integrate omics datasets at large-scale to characterise and predict disposition to complex disease.

Inferior neurological health during ageing has been associated with the increased incidence of many neurological diseases [11,12]. Relative brain age – a measure of brain health calculated using multiple volumetric brain imaging markers – has recently been shown to predict the development of dementia [13]. Well-characterised markers of neurological health, such as global grey matter volume and cognitive ability, are clearly associated with potentially modifiable traits such as body mass index (BMI), smoking and diabetes; [14–16] up to 40% of dementia cases may be preventable through modification of such factors [16]. The epigenome can record an individual's response to environmental exposures, stochastic biological effects, and genetic influences through changes to DNA methylation (DNAm). Profiling DNAm signatures for protein markers may therefore help to understand relationships between environmental factors and biological pathways occurring in the body that may impact brain health. Protein quantitative trait methylation loci (pQTMs) represent associations between DNAm at CpG sites and protein levels. As DNAm at CpG sites can affect gene expression in some instances [17], these relationships may carry information that informs on the regulation of proteins, the biological effector molecules of disease.

The largest epigenome wide association study (EWAS) of protein levels to date assessed 1,123 SOMAmer protein measurements in the German KORA cohort (n=944) [1]. In this study, Zaghlool *et al* reported 98 pQTMs that replicated in the QMDiab cohort (n=344), with significant associations between DNAm at the immune-associated locus *NLRC5* and seven immune-related proteins ($P < 2.5 \times 10^{-7}$). This suggested that DNAm not only reflects variability in the proteome but is also closely related to prolonged systemic inflammation. Hillary *et al* have also profiled 281 SOMAmer protein measurements that have been previously associated with Alzheimer's disease, in the Generation Scotland cohort that we utilise in this study [18]. However, proteome-wide assessment of pQTMs has not been tested against a comprehensive spectrum of neurological phenotypes. Blood-based methylation is unlikely to correlate highly with brain methylation in all cases [19]; however, proteomic and epigenetic signatures from the blood have been shown to be informative as predictors

of neurological outcomes [20,21] and brain health is also thought to be influenced by peripheral factors [12,22]. Brain morphology alterations and disease pathogenesis can occur many years prior to overt neurological, cognitive and behavioural symptom presentation. Risk stratification and intervention for secondary prevention in advance of clinical diagnoses may therefore be feasible, with blood sampling that supersedes more invasive and expensive tests. Therefore, while datasets that allow for integration of novel protein markers, phenotypic outcomes and epigenetic signatures of protein levels from the blood are rarely-available, they hold substantial potential to advance the development of these endeavours.

Here, we conduct an integrated epigenome-wide and phenome-wide study of the plasma proteome (**Fig. 1**). We characterise CpG-protein (pQTM) associations for 4,235 SOMAmer protein measurements in N=778 individuals from the Generation Scotland cohort (**Supplementary Table 1**). Epigenetic pQTM datasets for the plasma proteome are provided, with 3,286 SOMAmers previously untested in protein EWAS, enabling future studies to extract phenome-specific pQTM signatures for traits of interest. We then characterise proteome signatures for a range of neurological traits (structural brain imaging measures, cognitive ability and *APOE* carrier status) in N=1,065 individuals from the same cohort where the pQTM data are a nested subset. By integrating these datasets, we probe the epigenetic signatures of proteins that are related to brain health and dementia risk in an ageing population.

Results

Epigenome-wide studies of 4,235 plasma proteins

We conducted EWAS to test for pQTM associations between 772,619 CpG sites and 4,235 circulating protein levels (**Supplementary Table 2**) in N=778 individuals from Generation Scotland

(mean age 60.2, SD 8.8 years, 56% Female; **Supplementary Table 1).** A stringent threshold for multiple testing (0.05/(4,235x772,619) = $1.5x10^{-11}$) was applied across P-values for all EWAS results. In our basic model adjusting for age and sex, 131,691 (0.004% of tested associations) pQTMs passed this stringent threshold **(Supplementary Table 3)**. In our second model that further adjusted for Houseman-estimated white blood cell proportions[23] and available genetic pQTL effects from Sun *et al* [24], there were 2,974 associations (**Supplementary Table 4**). Smoking status and body mass index are known to have well-characterised DNAm signatures [25,26]; fully-adjusted models were therefore further adjusted for body mass index (BMI) and smoking status. There were 2,895 associations, with effect size ranging from -2.64 (SE 0.29) for PRG3 and cg16899419, to 2.62 (SE 0.19) for MDGA1 and cg12415337 (**Supplementary Table 5**). Across all models, 2,451 pQTM associations remained consistently significant. Lambda values for the 153 unique proteins that had pQTMs in the fully-adjusted EWAS are presented in **Supplementary Table 6**.

Of the 2,895 pQTMs from the fully-adjusted models, 1,141 involved Pappalysin-1 (PAPPA) protein and cg07839457 (*NLRC5* gene) was found to be the most frequently selected site, replicating the findings of a previous EWAS by Zaghlool *et al* [1]. Of the 98 pQTMs identified by Zaghlool *et al* (944 individuals, with 1,123 protein SOMAmers), 81 were comparable (both the protein and CpG sites from the 98 pQTMs were available across both EWAS) with our fully-adjusted results. Of these 81 pQTMs, 31 replicated at our stringent threshold ($1.5x10^{-11}$), a further 16 replicated at the epigenome-wide significance threshold of $P < 3.6x10^{-8}$ [27] and a further 31 replicated at nominal $P < 0.05$ (**Supplementary Table 7**). Three associations did not replicate at nominal $P < 0.05$; however, the direction of effect was consistent in all cases.

There were 2,854 novel SOMAmer pQTM associations in the protein EWAS, when accounting for 41 associations that were previously reported by EWAS from Zaghlool *et al* (31 associations) and Hillary *et al* (10 associations) [1,18]. In addition to the 1,141 pQTMs involving PAPPA, there were

1,157 pQTMs that involved the Proteoglycan 3 Precursor (PGR3) protein. The remaining 597 pQTMs involved 151 protein levels, with 342 *cis* (CpG <= 10Mb from the transcription start site of the protein gene) and 255 *trans* (CpG >= 10Mb from the transcription start site of the protein gene) associations **(Supplementary Table 5 and Fig. 2)**. A summary of CpGs selected in the fully-adjusted models, with EWAS catalogue [28] lookup of genome-wide significant phenotypic associations is presented in **Supplementary Table 8**. A summary of known pQTLs and whether these were available for adjustment in our models is provided in **Supplementary Table 9**.

Proteome associations with neurological phenotypes

We next conducted a proteome-wide association study on neurological phenotypes (protein PheWAS of brain imaging, cognitive and *APOE* e4 status, alongside age and sex; **Fig. 3).** A maximum sample of 1,065 individuals was available **(**mean age 59.9 years, SD 9.6 years, 59% Female; **Supplementary Table 1)**; all N=778 individuals from the pQTM study were included in these analyses.

At False Discovery Rate (FDR)-corrected P < 0.05, the levels of 798 plasma proteins were associated with age and 814 were associated with sex, with 394 proteins common to both phenotypes **(Supplementary Table 10)**. When compared with work by Lehallier *et al* using SOMAmer measurements [29], 340 of 476 age associations replicated and 372 of 450 sex associations replicated, with direction of change consistent across these associations.

There were 644 protein markers with FDR P < 0.05 that associated with neurological phenotypes. These consisted of 54 brain imaging (**Supplementary Table 11**), 579 cognitive test score (**Supplementary Table 12**) and 11 *APOE* (**Supplementary Table 13**) associations. **Supplementary Table 14** summarises markers identified by FDR adjustment. Of the seven brain morphology traits, Global Grey Matter Volume, General Fractional Anisotropy and Relative Brain Age had 10, 10 and 34 associations, respectively **(Supplementary Table 14)**. Of the cognitive score

traits, Processing Speed and General Fluid Cognitive Ability scores were associated with the highest numbers of protein markers (268 and 127, respectively; **Supplementary Table 14)**. Six of the 11 *APOE* associations replicated previous SOMAmer protein findings [10], and the five novel relationships were for NEFL, ING4, MENT, PAF and TMCC3. The strongest association for *APOE* was observed with levels of LRRN1 (Beta = 0.57, SE = 0.03, FDR P = 1.73×10^{-104}); a selection of associations for *APOE* are presented in **Fig. 3b**.

Many of the 644 protein marker associations were independent and did not cross neurological modalities. However, levels of APOB and ING4 were associated with *APOE* haplotype status and cognitive scores. Notably, 25 unique proteins were associated (FDR P < 0.05) with both cognitive and imaging traits (**Fig. 3a; Supplementary Table 15**). The direction of effect for all of the 25 common proteins was consistent with poorer brain health across both modalities. For example, with lower levels of Retinoblastoma-like protein 2 (RBL2) were associated with lower Global Grey Matter Volume Beta = -0.24, SE = 0.06, P = 9.18×10^{-5}), General Fluid Cognitive Ability (Beta = -0.09, SE = 0.02, P = 0.0008) and Processing Speed (Beta = -0.1, SE = 0.03, P = 0.002). A selection of associations across imaging and cognitive modalities are shown in **Fig. 3c**, for the RBL2 and TREM1 protein markers. Of the 25 common proteins, there were six independent signals, as determined by components with eigenvalues > 1 in principal components analyses **(Supplementary Fig.1).**

Integration of the neurological proteome with our pQTM dataset

pQTM signatures were explored for the 644 protein markers with FDR P < 0.05 in the protein PheWAS of neurological outcomes. We examined whether the 644 proteins identified in the PheWAS had significant pQTM associations in our EWAS. Twenty-five of the 644 proteins had significant pQTMs. 88 loci were shared across these 25 proteins. Of the 88 pQTMs, 42 were *trans*

(**Fig. 4**) and 46 were *cis* associations (**Supplementary Table 16**). Eighty-three of the 88 pQTMs were novel.

Several CpG sites were associated with multiple protein levels in the *trans* pQTMs (**Fig. 5**). DNAm at site cg06690548 in the *SLC7A11* locus was associated with PSAT1, SCUBE1, ACY1 and ALDOB levels. The cg11294350 site in the *CHPT1* genomic region was associated with HEXB and SMPD1 levels. pQTMs were reported between circulating RBL2 levels and sites cg01132052, cg0539861 and cg18404041, within the *NEK4/ITIH3/ITIH1* gene region of chromosome 3. Finally, a large proportion of *trans* pQTMs revolved around NLCR5 inflammatory CpGs, with five protein markers associated with either the cg07839457 or cg16411857 sites in seven pQTM associations (**Supplementary Fig. 2**). A further nine and 13 associations were identified for GBP1 and MX1, respectively. Two singular trans associations were found; one between cg02521229 and TREM2 and one between cg08228578 and PSAT1. Finally, the levels of CRTAM and SCUBE1 were also associated with DNAm at cg18544413 and cg06088069, respectively.

Discussion

We have conducted a large-scale study of epigenetic and phenotypic correlates of the circulating proteome in relation to brain health outcomes. We provide the first epigenome-wide characterisation of the full SOMAscan® panel V.4, for 4,235 protein measurements, adding data for 3,286 SOMAmer protein measurements and over 2,854 novel pQTM associations to current knowledge. By characterising 644 proteins associated with neurological phenotypes and integrating them with our pQTM dataset, we uncover 88 epigenome-proteome-phenome signatures pertinent to brain health. The directionality of pQTM relationships is unclear; however, these signatures may have predictive

value for risk stratification and identify biological effectors that may contribute to neurological phenotypes.

Though informative, protein biomarkers are only one component of phenotypic regulation. Integration of our protein EWAS dataset with protein PheWAS associations in this study revealed novel pQTMs associated with neurological markers (**Fig. 5**). While this study is focused on blood samples – limiting interpretation to tissues such as the brain – many of the 88 pQTMs involved CpG sites and proteins that have been previously implicated in studies of neurological phenotypes. The *NEK4/ITIH3/ITIH1* is a key locus implicated in schizophrenia and bipolar disorder by several large-scale, genome-wide association studies (GWAS) [30–33]. Similarly, the *RBL2* locus has been associated with intelligence, cognitive function and educational attainment in GWAS (n > 260,000 individuals) [34–36]. The HEXB and SMPD1 proteins associated with DNAm at cg11294350 (in the *CHPT1* gene) are involved in neuronal sphingolipid degradation pathways in the brain and have been associated with the pathology and onset of a range of neurodegenerative conditions across the lifespan [37–40]. We make our proteome-wide epigenetic dataset publicly available, such that proteome-specific pQTM signatures can be extracted by studies for any desired trait, as we have demonstrated here for traits related to brain health.

DNAm at cg06690548 has previously been identified as a blood-based marker for Parkinson's disease risk (N > 900 cases and N > 900 controls); in this study, Vallegra *et al* found evidence that DNAm at this site was consistent with environmental exposure and causally implicated in Parkinson's disease risk [41]. Xc- is the cystine-glutamate antiporter encoded by *SLC7A11*, which facilitates glutamatergic transmission, oxidative stress defence and microglial immune response in the brain [42,43]. Xc- is also a target for the environmental neurotoxin β-methylamino-L-alanine (BMAA), which has been linked to the onset of neurodegenerative process [41]. Given that the proteins implicated in the cg06690548 (*SLC7A11*) pQTMs were associated with declines in processing speed

and increased relative brain age, the CpG sites we identify in this study – such as cg06690548 – may be important plasma markers for brain health that merit further exploration.

In the fully-adjusted EWAS we corrected for smoking and BMI; the 88 identified pQTMs may therefore represent unmeasured biological or environmental factors associated with inter-individual variation in protein levels. The EWAS dataset we provide represents the largest set of pQTMs for circulating plasma proteins to date (4,235 protein levels, with 2,895 fully-adjusted pQTMs). Our replication of associations identified by Zaghlool *et al* [1] reinforces the presence of inflammation signalling as intrinsic to the epigenome signature of blood proteins. The presence of *NLRC5* in our neurological protein pQTMs (**Supplementary Fig. 2; Supplementary Note 1**) suggests that the epigenome may reflect an inflammatory component of brain health. Our results support previous work suggesting that sustained peripheral inflammation is related to mood disorders [44,45], influences general brain health [4647] and accelerates cognitive decline [8,20,34,48].

Several of the markers for cognitive function were identified in previous work relating Olink proteins to cognitive function (such as PVR, VWC2 and CPM) [49] and work that characterised SOMAmer signatures of cognitive decline and incident Alzheimer's disease (such as SVEP1 and TREM2) [8]. C5 and C3 are potent mediators of widespread peripheral inflammation [52] and were associated with cognitive, but not brain morphology traits. The *APOE* carrier-related proteins were also largely distinct from cognitive and brain imaging proteomic signatures. The strong association between lower BIRC2 levels with e4 carrier status replicates key findings of a previous study of *APOE* haplotype [10]. Each measurement strategy may therefore be reflective of divergent biological pathways. Continued integration of omics with these phenotypic modalities in increasing sample populations will help to clarify these signals.

Notably, there were 25 proteins that were associated with both cognitive and imaging phenotypes; our results suggest that these proteins are reflective of a common, biological signature of poorer

outcomes across modalities. Higher levels of three (IGLON5, NCAN and GLIPR2) of the 25 proteins had were associated with increased brain health, and previous studies have linked IGLON5 and NCAN to neural maintenance pathways [50,51]. Higher levels of the remaining 22 markers were associated with poorer neurological health; RBL2, HEXB and SMPD1 are examples of these and each had novel pQTMs. These proteins are therefore strong candidates for exploration of pathways pertaining to brain health.

Our study has several limitations. First, many of the proteins identified in our protein PheWAS did not have epigenetic pQTMs; this may be due to 1) the presence of pathways relating to neurological disease that are not reflected by the plasma epigenome, 2) underpowered analyses, or 3) the presence of indirect pathways that are not captured by our linear mixed model approach. Second, though full replication of our results was not possible, we did see overlap with published findings (across 1,123 SOMAmer protein measurements) from Zaghlool *et al* [53] suggesting that pQTMs generalise to individuals of other European ancestries. Third, we observed a degree of inflation for the PAPPA and PRG3 proteins in particular, which had large signals across the epigenome. This was despite the fact that mixed effects model structure was used to account for unknown confounding by modelling intercorrelations between CpGs. One reason for this inflation may be a lack of adjustment for eosinophil cell proportions [53], which were unavailable. Fourth, the extent of non-specific and cross-aptamer binding with SOMAmer technology has not been fully resolved [54]. Fifth, although we regressed known pQTLs effects for 3,622 of the 4,235 proteins [55], there are likely unknown genetic pQTLs that influenced pQTM associations. Within our neurological *trans* pQTMs, however, only one association for CD163 had a pQTL that was not possible to adjust for in our sample. There were overall 184 of 1,021 possible pQTLs [24] that were unavailable in the sample. Further characterisation of pQTLs and advances in multi-omic modelling techniques [2] will aid in the separation of genetic and environmental influences on epigenetic signatures. Finally, though we have incorporated a wide portfolio of neurological phenotypes, we recognise that the 14 we include in this study are not

extensive and results may vary depending on the type of cognitive or imaging measure studied. This is why increasing triangulation across modalities, as we have shown here, will be useful in identifying candidate markers of brain health.

In conclusion, by integrating epigenetic, proteomic and phenotypic data we have uncovered 2,854 novel pQTMs, 83 of which involve the levels of plasma proteins that were associated with neurological phenotypes. We also define proteomic signatures of cognitive ability, brain morphology and *APOE* carrier status. As these traits are pertinent to cognitive ageing and neurodegeneration, these data are likely to inform preventative approaches and risk stratification.

Methods

The Generation Scotland sample population

The Stratifying Resilience and Depression Longitudinally (STRADL) cohort used in this study is a subset of N=1,188 individuals from Generation Scotland: The Scottish Family Health Study (GS). Generation Scotland constitutes a large, family-structured, population-based cohort of >24,000 individuals from Scotland [56]. Individuals were recruited to GS between 2006 and 2011. During a clinical visit detailed health, cognitive, and lifestyle information was collected in addition to biological samples. Of the 21,525 individuals contacted for participation, N=1,188 completed additional health assessments and biological sampling approximately five years after GS baseline [57]. There were N=1,065 individuals with proteomic data available and N=778 of these had DNAm data available. **Supplementary Table 1** summarises the demographic characteristics across the two groups, with descriptive statistics for phenotypes.

Proteomic measurement

SOMAscan® V.4 technology was used to quantify plasma protein levels. This aptamer-based assay facilitates the simultaneous measurements of multiple SOMAmers (Slow Off-rate Modified Aptamers) [58]. SOMAmers were processed for 1,095 individuals from the STRADL subset of GS. Briefly, binding between plasma samples and target SOMAmers was achieved during incubation and quantification was recorded using a fluorescent signal on microarrays. Quality control steps included hybridization normalization, signal calibration and median signal normalization to control for inter-plate variation. In the final dataset, SOMAmer epitope measures for the 4,235 proteins were available in 1,095 individuals. **Supplementary Table 2** provides annotation information for the 4,235 SOMAmers.

DNAm measurement

Measurements of blood DNAm in the STRADL subset of GS subset were processed using the same methodology as those collected in the wider Generation Scotland cohort. Two sets were generated using the Illumina EPIC array. Quality control details have been reported previously [59–61]. Briefly, outlying probes were removed based on visual inspection, bead count and detection P value. Samples were removed if there was a mismatch between predicted and actual sex, as well as those with outlying detection P values. All non-specific CpG and SNP probes (with allele frequency > 5%) were removed from the methylation file. After quality control, 793,706 and 773,860 CpG were available in sets 1 and 2, respectively. These sets were truncated to include a total of 772,619 common probes and were joint together for use in the EWAS, with 479 individuals included in set 1 and 299 individuals in set 2.

Phenotypes in STRADL

An epigenetic score for smoking exposure, EpiSmokEr [62] and Houseman-estimates for white blood cell (WBC) proportions [23] were calculated for all individuals with DNAm. Imputed WBC measures were available for Monocytes, Natural Killer cells, Bcells, Granulocytes, $CD4^+$T cells and $CD8^+$T cells. Body mass index (body weight in kg, divided by squared height in m^2) and smoking status (1 = smoker, 2 = ex-smoker, 3 = non-smoker – history unknown , 4 = never smoked) were available alongside depression status (defined using either SCID or linked diagnosis information was coded as a binary variable).

APOE haplotype status was available for all individuals. *APOE* risk was coded as a numeric variable (e2e2 = 0, e2e3 = 0, e3e3 = 1, e3e4 = 2, e4e4 = 2). Fifteen e2e4 individuals were excluded due to small sample size. Scores from five cognitive tests measured at the clinic visit for the STRADL subset of GS were considered. The sum of immediate and delayed recall of one oral story from the Wechsler Logical Memory Test was taken as the logical memory phenotype [63]. The Wechsler Digit Symbol Substitution Test, which requires individuals to recode digits to symbols was used to measure processing speed phenotype [64]. The verbal fluency phenotype was based on the phonemic fluency test with letters C, F and L [65]. The vocabulary score was calculated by collating the Mill Hill Vocabulary scale, junior and senior synonyms [66]. The Matrix Reasoning test was used as a measure of the non-verbal reasoning phenotype [66]. The first unrotated principal component combining logical memory, verbal fluency, vocabulary and digit symbol tests (sum of squares loadings = 1.76 and proportion of variance = 0.44) was calculated as a measure of general cognitive ability (‘*g*’). General fluid cognitive ability (‘*gf*’) was extracted using the same approach, but with the vocabulary test (a crystallised measure of intelligence) excluded from the composite (sum of squares loadings = 1.52 and proportion of variance = 0.51).. While highly similar to *g*, the *gf* score

is exclusive to measures such as memory and processing capability that are considered fluid. *gf* may therefore be of pertinence to delineate when assessing cognitive performance in ageing individuals.

The acquisition protocols for the brain imaging phenotypes are documented in a previous publications [574867]. The derived brain volume measures were recorded at two sites (Aberdeen and Edinburgh) and site was included as a covariate in all imaging analyses. Brain volume data included whole brain volume, global grey matter volume, cerebral total volume and total intracranial volume. Intracranial volume was used as a nuisance covariate to adjust for head size in all tests of brain volume associations. The derived global white matter integrity measures included global fractional anisotropy (gFA) and global mean diffusivity (gMD). Please refer to the previous publications for details of the protocols applied to derive the brain volume measures from T1-weighted scans, and white matter integrity measures from diffusion tensor imaging (DTI) scans [4867]. White-matter hyperintensity ratings for each participant were derived through visual inspection of FLAIR scans, using the Fazekas scale [68]. White matter hyperintensities (WMH) were defined as punctuate, focal or diffuse lesions in the deep or periventricular white matter, basal ganglia or brainstem, visible as areas of hyperintensity on FLAIR images. Total score was on the scale of 0-6 and was obtained by summing deep and periventricular hyperintensity scores (each on the scale 0-3). Brain Age was estimated using the software package 'brainageR' (Version 2.1; DOI: 10.5281/zenodo.3476365, available at https://github.com/james-cole/brainageR), which uses machine learning and a large training set to predict age from whole-brain voxel-wise volumetric data derived from structural T1 images [13]. This estimate was residualised over chronological age to produce a measure of Relative Brain Age. Sex and scan site were also included as predictors in the model used to residualise to ensure that any bias related to either was excluded from the final measure.

Phenome-wide association analyses

Prior to running protein PheWAS analyses, protein levels were transformed by rank-based inverse normalisation and scaled to have a mean of zero and standard deviation of 1. Models were run using the lmekin function in the coxme R package [69]. This modelling strategy allows for mixed-effects linear model structure with adjustment for relatedness between individuals. Models were run in the maximum sample of 1,095 individuals, with the 4,235 protein levels as dependent variables and phenotypes as independent variables. A random intercept was fitted for each individual and a kinship matrix was included as a random effect to adjust for relatedness. Age, sex (male = 1, reference female = 0), numerical *APOE* variable (e2 = 0, e3 = 1, e4 = 2), cognitive and brain imaging phenotypes were included as outcomes. Diagnosis of depression (case = 1, reference control = 0) at the STRADL clinic visit in GS was included as a covariate in all models, due to known selection bias for depression phenotypes in STRADL [57]. Missing data were excluded from lmekin models.

Coefficients were extracted from the following model for age and sex:

Protein level ~ α + β1 age + β2 sex + depression + (1|individual) + (1|kinship) + ε

Coefficients were extracted from the following model for cognitive scoring phenotypes:

Protein level ~ α + β1 age + β2 sex + phenotype + depression + (1|individual) + (1|kinship) + ε

Coefficients were extracted from the following model for brain imaging phenotypes:

Protein level ~ α + β1 age + β2 sex + phenotype + depression + imaging site + technical variables + (1|individual) + (1|kinship) + ε

All analyses of brain volume measures included covariates to adjust for intracranial volume, imaging data processing batch, and presence or absence of manual intervention during quality control. Sensitivity analyses were run for cognitive and imaging phenotypes, with additional adjustments for

smoking status and BMI. P-values were calculated in R using effect size estimates (beta) and standard errors (SE) as follows: pchisq((beta/SE)^2, 1, lower.tail=F). False discovery rate (FDR) P-value adjustment was applied for each proteome assessment (4,235 associations corrected for each phenotype).

Epigenome-wide association study of protein levels

There were 778 individuals with protein level and DNAm measurements. Prior to running the EWAS, protein levels were log transformed and regressed onto age, sex, study site, lag group and 20 genetic principal components (generated from multidimensional scaling of genotype data from the Illumina 610-Quadv1 array). Residuals from these models were then rank-based inverse normalised. Methylation data were in M-value format and were pre-adjusted for age, sex, processing batch, methylation set, and depression status in the basic model [62]. A second model further adjusted the DNAm for estimated white blood cell proportions ($CD4^+T$, $CD8^+T$, BCell, Natural Killer cells and Granulocytes), in addition to regressing known pQTL effects (from a previous genome-wide association study of 4,034 SOMAmers targeting 3,622 proteins from Sun *et al*) onto protein levels [55]. Finally, the fully-adjusted model regressed DNAm onto an epigenetic score for smoking, EpiSmokEr[62] and body mass index (BMI).

Omics-data-based complex trait analysis (OSCA) [70] Version 0.41 was used to run EWAS analyses. Within OSCA, a genetic relatedness matrix (GRM) was constructed for the STRADL population and a threshold of 0.05 was used to identify 120 individuals likely to be related based on their genetic similarity. For this reason, the MOA method was used to calculate associations between individual CpG sites and protein levels, with the addition of the GRM as a fixed covariate to adjust for relatedness between individuals. MOA is a mixed-linear-model-based method that is designed to account for unobserved confounders and the correlation between distal probes that are likely to be introduced by such confounders, by modelling each CpG as a random effect [70]. CpG sites were the

independent variables and the 4,235 proteins were the dependent variables. A threshold for multiple testing correction was applied based on the significance level of $P < 0.05 / 4{,}235$ proteins / 772,619 CpG sites, which equalled $P < 1.5 \times 10^{-11}$. Five fully-adjusted models did not converge (NAGLU, CFHR2, DAP, MST1, PILRA) and were excluded. Circos plots were created with the circlize package (Version 0.4.12) [71]. BioRender.com was used to create Figs. 1 and 5. All analyses were performed in R (Version 4.0) [72].

Ethics declarations

All components of GS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). GS has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20/ES/0021), providing generic ethical approval for a wide range of uses within medical research.

Data availability

Datasets generated in this study are made available in Supplementary Tables.

The source datasets from the cohorts that were analysed during the current study are not publicly available due to them containing information that could compromise participant consent and confidentiality. Data can be obtained from the data owners. Instructions for accessing Generation Scotland data can be found here: https://www.ed.ac.uk/generation-scotland/for-researchers/access; the 'GS Access Request Form' can be downloaded from this site. Completed request forms must be sent to access@generationscotland.org to be approved by the Generation Scotland Access Committee.

Code availability

All R code used in this study is available with open access at the following Gitlab repository: https://gitlab.com/dannigadd/epigenome-and-phenome-wide-study-of-brain-health-outcomes/-/tree/main

Acknowledgements

This research was funded in whole, or in part, by the Wellcome Trust [104036/Z/14/Z, 108890/Z/15/Z, 203771/Z/16/Z, 221890/Z/20/Z**]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.** D.A.G. and R.F.H. are supported by funding from the Wellcome Trust 4-year PhD in Translational Neuroscience–training the next generation of basic neuroscientists to embrace clinical research [108890/Z/15/Z]. R.I.M is supported by funding from the Wellcome Trust PhD for clinicians, Edinburgh Clinical Academic Track for Veterinary Surgeons. D.L.Mc.C. and R.E.M. are supported by Alzheimer's Research UK major project grant ARUK-PG2017B−10. 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 GS samples was carried out by the Genetics Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" ([STRADL; Reference 104036/Z/14/Z]). 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]. C.H. is supported by an MRC University Unit Programme Grant MC_UU_00007/10 (QTL in Health and Disease). L.S. is funded by DPUK through MRC (grant no. MR/L023784/2) and the UK Medical Research

Council Award to the University of Oxford (grant no. MC_PC_17215). L.S. also received support from the NIHR Biomedical Research Centre at Oxford Health NHS Foundation Trust.

S.R.C is supported by the Medical Research Council (MR/R024065/1), a National Institutes of Health research grant (R01AG054628) and a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 221890/Z/20/Z).

JDS is funded by MRC grants MR/S010351/1, MR/W002566/1 and MR/W002388/1. JMW is supported by the UK Dementia Research Institute which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK. EB is supported by Stroke Association/BHF/Alzheimer's Society 'Rates Risks and Routes to Reduce Vascular Dementia' (R4VaD) Priority Programme Award in Vascular Dementia (16 VAD 07).

The authors acknowledge the work of Rebecca Madden and Claire Green, of the University of Edinburgh, who aided in the quality control of volumetric brain imaging data.

Author contributions

D.A.G., and R.E.M., were responsible for the conception and design of the study. D.A.G. carried out the data analyses. D.A.G., and R.E.M., drafted the article. S.R.C., and H.W., advised on methodology. R.F.H., and D.L.Mc., contributed to methodology and data analyses. R.I.McG., S.M., R.M.W., L.S., D.L.Mc., R.M.W., A.C., A.N.H., C.H., K.L.E., D.J.P., H.W., A.M.M., and S.R.C., contributed to data methylation and proteomic data collection and preparation. A.S., M.C.B., M.A.H., E.V.B., J.D.S., and J.M.W processed the brain imaging data. D.A.O., G.M.T., and C.R., provided scientific counsel. R.E.M., supervised the project. All authors read and approved the final manuscript.

Competing interests

R.E.M has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation. A.M.M has previously received speaker's fees from Illumina and Janssen and research grant funding from The Sackler Trust. All other authors declare no competing interest.

Materials and correspondence

All correspondence and material requests should be sent to Dr Riccardo Marioni at riccardo.marioni@ed.ac.uk.

References

1. Zaghlool, S. B. *et al.* Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nat. Commun.* **11**, 15 (2020).

2. Hillary, R. F. *et al.* Multi-method genome- And epigenome-wide studies of inflammatory protein levels in healthy older adults. *Genome Med.* **12**, 60 (2020).

3. Hillary, R. F. *et al.* Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. *Nat. Commun.* **10**, 3160 (2019).

4. Marioni, R. E. *et al.* Meta-analysis of epigenome-wide association studies of cognitive abilities. *Mol. Psychiatry 2017 2311* **23**, 2133–2144 (2018).

5. Hillary, R. F. *et al.* An epigenetic predictor of death captures multi-modal measures of brain health. *Mol. Psychiatry* 1–11 (2019) doi:10.1038/s41380-019-0616-9.

6. Freytag, V. *et al.* A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and memory. *Nat. Commun. 2017 81* **8**, 1–12 (2017).

7. Liu, X., Jiao, B. & Shen, L. The Epigenetics of Alzheimer's Disease: Factors and Therapeutic Implications. *Front. Genet.* **9**, 579 (2018).

8. Lindbohm, J. V *et al.* Association of plasma proteins with rate of cognitive decline and dementia: 20-year follow-up of the Whitehall II and ARIC cohort studies. *medRxiv* 2020.11.18.20234070 (2020) doi:10.1101/2020.11.18.20234070.

9. Walker, K. A. *et al.* Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nat. Aging* **1**, 473–489 (2021).

10. Sebastiani, P. *et al.* A serum protein signature of APOE genotypes in centenarians. *Aging Cell*

18, e13023 (2019).

11. Ly, M. *et al.* Late-life depression and increased risk of dementia: a longitudinal cohort study. *Transl. Psychiatry 2021 111* **11**, 1–10 (2021).

12. Shi, Y. & Wardlaw, J. M. Update on cerebral small vessel disease: a dynamic whole-brain disease. *Stroke Vasc. Neurol.* **1**, 83–92 (2016).

13. Biondo, F. *et al.* Brain-age predicts subsequent dementia in memory clinic patients. *medRxiv* 2021.04.03.21254781 (2021) doi:10.1101/2021.04.03.21254781.

14. Corley, J. *et al.* Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. *Transl. Psychiatry* **9**, (2019).

15. Stillman, C. M., Weinstein, A. M., Marsland, A. L., Gianaros, P. J. & Erickson, K. I. Body–Brain Connections: The Effects of Obesity and Behavioral Interventions on Neurocognitive Aging. *Front. Aging Neurosci.* **9**, 115 (2017).

16. Livingston, G. *et al.* Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet* vol. 396 413–446 (2020).

17. Lea, A. J. *et al.* Genome-wide quantification of the effects of DNA methylation on human gene regulation. *Elife* eLife 2018;7:e37513 (2018) doi:10.7554/eLife.37513.

18. Hillary, R. F. *et al.* Genome and epigenome wide studies of plasma protein biomarkers for Alzheimer's disease implicate TBCA and TREM2 in disease risk. *medRxiv* 2021.06.07.21258457 (2021) doi:10.1101/2021.06.07.21258457.

19. Hannon, E., Lunnon, K., Schalkwyk, L. & Mill, J. Interindividual methylomic variation across

blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* **10**, 1024–1032 (2015).

20. Conole, E. L. S. *et al.* An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. medRxiv. (2020) doi:https://doi.org/10.1101/2020.10.08.20205245.

21. Gadd, D. A. *et al.* Epigenetic scores for the circulating proteome replicate protein-disease predictions as tools for biomarker discovery. *Melanie Waldenberger* **8**, 15 (2021).

22. Cox, S. R. *et al.* Associations between vascular risk factors and brain MRI indices in UK Biobank. *Eur. Heart J.* **40**, 2290–2299 (2019).

23. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

24. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

25. McCartney, D. L. *et al.* Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**, 214–220 (2018).

26. McCartney, D. L. *et al.* Epigenetic prediction of complex traits and death. *Genome Biol.* **19**, 136 (2018).

27. Saffari, A. *et al.* Estimation of a significance threshold for epigenome-wide association studies. **42**, 22–23 (2017).

28. MRC-IEU. The MRC-IEU catalog of epigenome-wide association studies. Avaialable at: http://www.ewascatalog.org. Accessed April 2021. (2021).

29. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the lifespan.

Nat. Med. **25**, 1843–1850 (2019).

30. Hamshere, M. L. *et al.* Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Mol. Psychiatry* **18**, 708–712 (2013).

31. Witt, S. H. *et al.* Investigation of manic and euthymic episodes identifies state-and trait-specific gene expression and stab1 as a new candidate gene for bipolar disorder. *Transl. Psychiatry* **4**, 426 (2014).

32. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–978 (2011).

33. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43**, 977–985 (2011).

34. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

35. Davies, G. *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun.* **9**, 1–16 (2018).

36. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).

37. Park, M. H., Jin, H. K. & Bae, J. sung. Potential therapeutic target for aging and age-related neurodegenerative diseases: the role of acid sphingomyelinase. *Exp. Mol. Med.* **52**, 380–389 (2020).

38. Lee, J. K. *et al.* Acid sphingomyelinase modulates the autophagic process by controlling

lysosomal biogenesis in Alzheimer's disease. *J. Exp. Med.* **211**, 1551–1570 (2014).

39. Kyrkanides, S. *et al.* Conditional expression of human β-hexosaminidase in the neurons of Sandhoff disease rescues mice from neurodegeneration but not neuroinflammation. *J. Neuroinflammation* **9**, 186 (2012).

40. Bley, A. E. *et al.* Natural history of infantile G M2 gangliosidosis. *Pediatrics* **128**, e1233 (2011).

41. Vallerga, C. L. *et al.* Analysis of DNA methylation associates the cystine–glutamate antiporter SLC7A11 with risk of Parkinson's disease. *Nat. Commun.* **11**, 1–10 (2020).

42. Fournier, M. *et al.* Implication of the glutamate-cystine antiporter xCT in schizophrenia cases linked to impaired GSH synthesis. *npj Schizophr.* **3**, 1–7 (2017).

43. Mesci, P. *et al.* System xC- is a mediator of microglial function and its deletion slows symptoms in amyotrophic lateral sclerosis mice. *Brain* **138**, 53–68 (2015).

44. Yuan, N., Chen, Y., Xia, Y., Dai, J. & Liu, C. Inflammation-related biomarkers in major psychiatric disorders: a cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. *Translational Psychiatry* vol. 9 1–13 (2019).

45. Felger, J. C. *et al.* Inflammation is associated with decreased functional connectivity within corticostriatal reward circuitry in depression. *Mol. Psychiatry* **21**, 1358–1365 (2016).

46. Jefferson, A. L. *et al.* Inflammatory biomarkers are associated with total brain volume: The Framingham Heart Study. *Neurology* **68**, 1032–1038 (2007).

47. Janowitz, D. *et al.* Inflammatory markers and imaging patterns of advanced brain aging in the general population. *Brain Imaging Behav.* **14**, 1108–1117 (2020).

48. C, G. *et al*. Structural brain correlates of serum and epigenetic markers of inflammation in major depressive disorder. *Brain. Behav. Immun.* **92**, 39–48 (2021).

49. Harris, S. E. *et al*. Neurology-related protein biomarkers are associated with cognitive ability and brain volume in older age. *Nat. Commun.* **11**, 1–12 (2020).

50. Montagna, M. *et al.* IgLON5-Associated Encephalitis With Atypical Brain Magnetic Resonance Imaging and Cerebrospinal Fluid Changes. *Front. Neurol.* **0**, 329 (2018).

51. Schmidt, S., Arendt, T., Morawski, M. & Sonntag, M. Neurocan Contributes to Perineuronal Net Development. *Neuroscience* **442**, 69–86 (2020).

52. Merle, N. S., Noe, R., Halbwachs-Mecarelli, L., Fremeaux-Bacchi, V. & Roumenina, L. T. Complement system part II: Role in immunity. *Front. Immunol.* **6**, 257 (2015).

53. Zaghlool, S. B. *et al.* Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nat. Commun.* **11**, 15 (2020).

54. Pietzner, M. *et al*. Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).

55. Sun, B. B. *et al*. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

56. Smith, B. H. *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. *Int. J. Epidemiol*. **42**, 689–700 (2013).

57. Habota, T. *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. *Wellcome Open Res.* **4**, 185 (2019).

58. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).

59. Seeboth, A. *et al.* DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936. *Clin. Epigenetics* **12**, 49 (2020).

60. McCartney, D. L. *et al.* Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* **10**, 429–437 (2018).

61. Amador, C. *et al.* Recent genomic heritage in Scotland. *BMC Genomics* **16**, 437 (2015).

62. Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. *Epigenomics* **11**, 1469–1486 (2019).

63. Wechsler, D. WMS-III Wechsler memory Scale. Psychological Corporation. 1997.

64. Wechsler, D. WAIS--III: Wechsler Adult Intelligence Scale (3rd ed.) Administration and scoring manual. The Psychological Corporation, San Antonio, TX. 1997.

65. Lezak, M. Neuropsychological assessment (3rd ed.). 1995.

66. Raven, J., Raven, J. & Court, J. The Mill Hill vocabulary scale. Manual for Raven's progressive matrices and vocabulary scales. 1988.

67. A, S. *et al.* Automated classification of depression from structural brain measures across two independent community-based cohorts. *Hum. Brain Mapp.* **41**, 3922–3937 (2020).

68. F, F., JB, C., A, A., HI, H. & RA, Z. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *AJR. Am. J. Roentgenol.* **149**, 351–356 (1987).

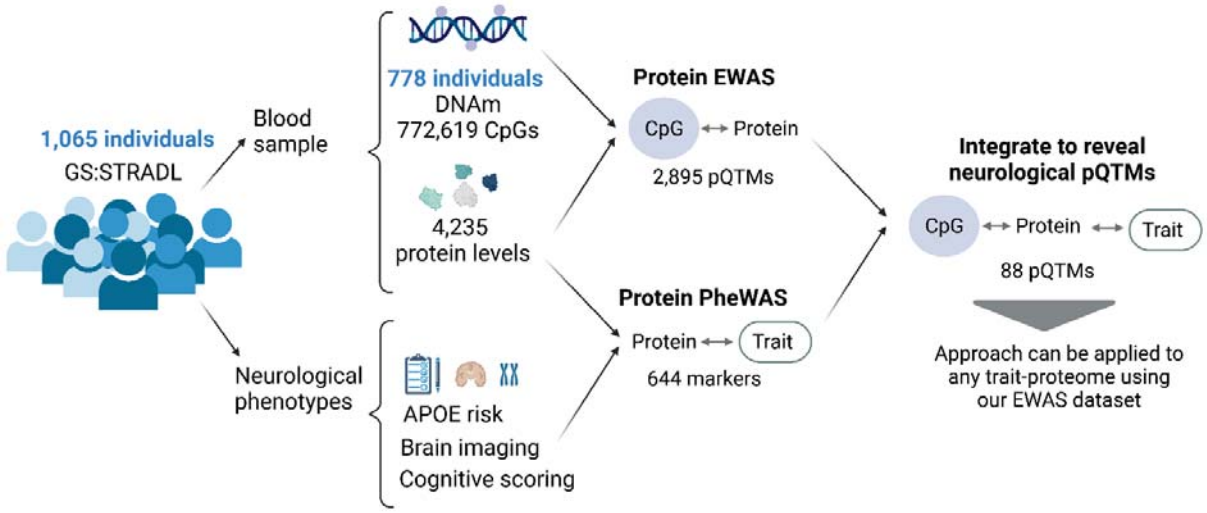
69. Therneau, T. M. coxme: Mixed Effects Cox Models. R package version 2.2-16. https://CRAN.R-project.org/package=coxme. Accessed April 2021. (2020).

70. Zhang, F. *et al.* OSCA: A tool for omic-data-based complex trait analysis. *Genome Biol.* **20**, (2019).

71. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

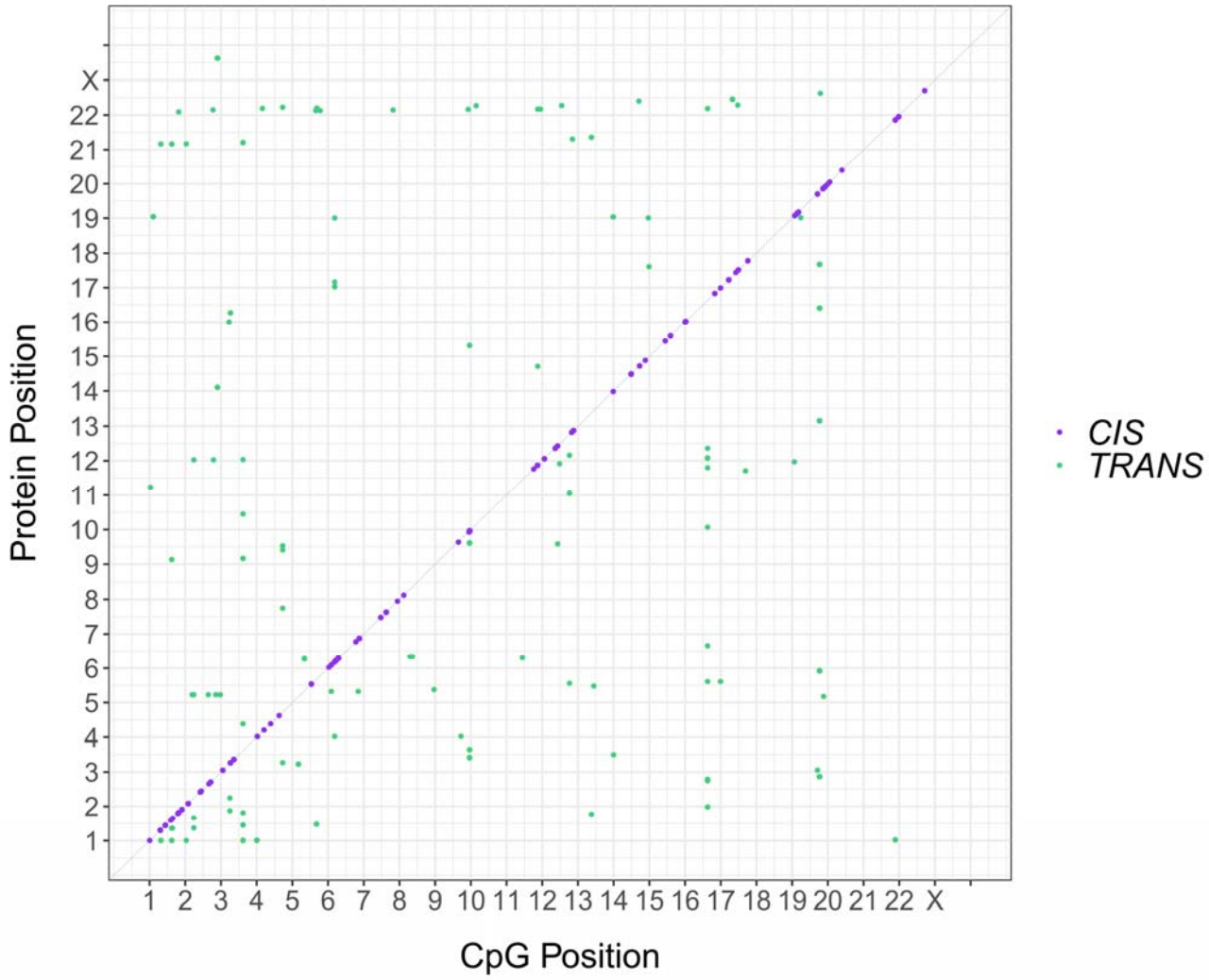
72. (2017), R. C. T. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Fig. 1. Integrated epigenome and phenome study of the plasma proteome



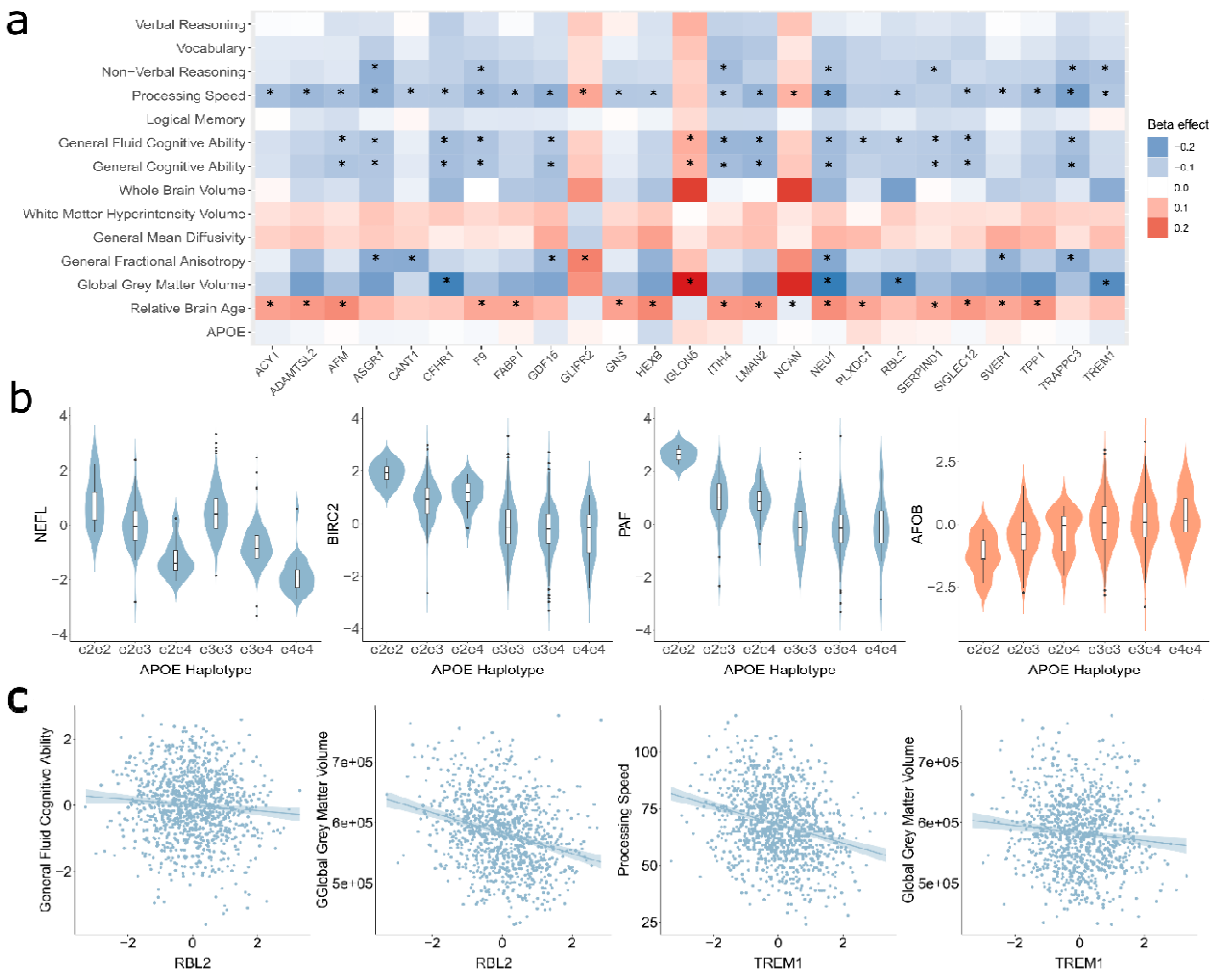
Study design and key results are presented in this flow diagram. An epigenome-wide association study (EWAS) of 4,235 plasma protein levels in 1,065 individuals from Generation Scotland identified 2,895 CpG-protein (pQTM) associations. A phenome-wide protein association assessment (Protein PheWAS) in 778 of the individuals identified 644 protein levels that were associated with a minimum of one neurological phenotype. Integration of the protein EWAS and PheWAS results identified 88 neurologically-relevant pQTMs that involved 25 unique protein measurements. The protein EWAS dataset we provide for 4,235 circulating proteins can be applied to any trait-proteome of interest to extract pQTM signatures.

Fig. 2. Epigenome-wide association study of 4,235 plasma proteins



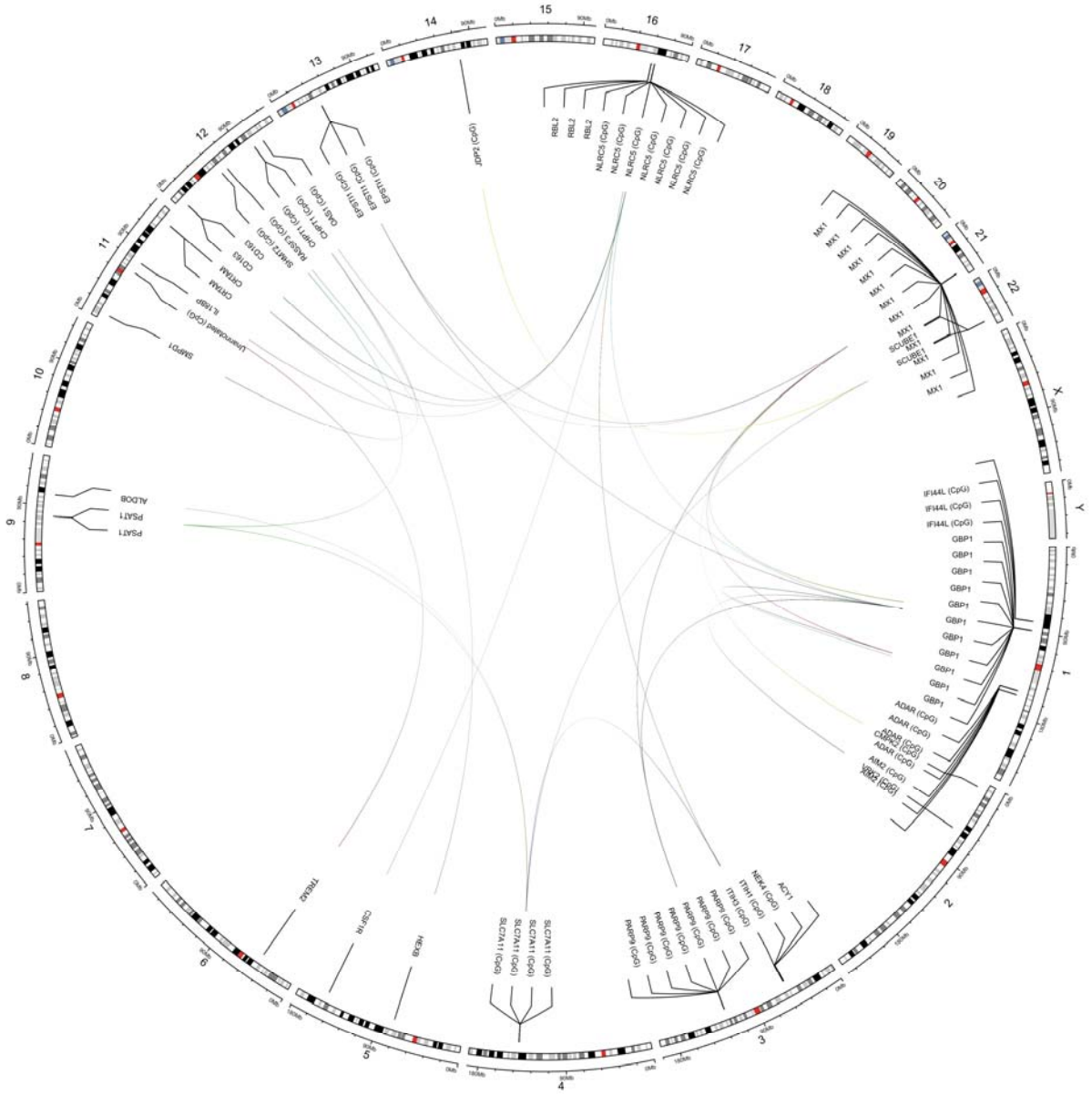
a, Genomic locations of CpG sites associated with differential levels of circulating SOMAscan® proteins. The 597 pQTMs that were significant ($P = 1.5x10^{-11}$) for 151 of the 153 proteins are shown here. The x-axis represents the chromosomal location of CpG sites. The y-axis represents the position of the gene encoding the associated protein. The 342 *cis* CpG sites (purple) identified by our EWAS on protein levels lay within 10Mb of the protein-coding gene, whereas the 255 *trans* CpG sites (green) lay outside of the protein gene. The PGR3 and PAPPA proteins had a combined pQTM total of 2,298 and are not shown in this plot.

Fig. 3. Plasma protein associations with neurological phenotypes



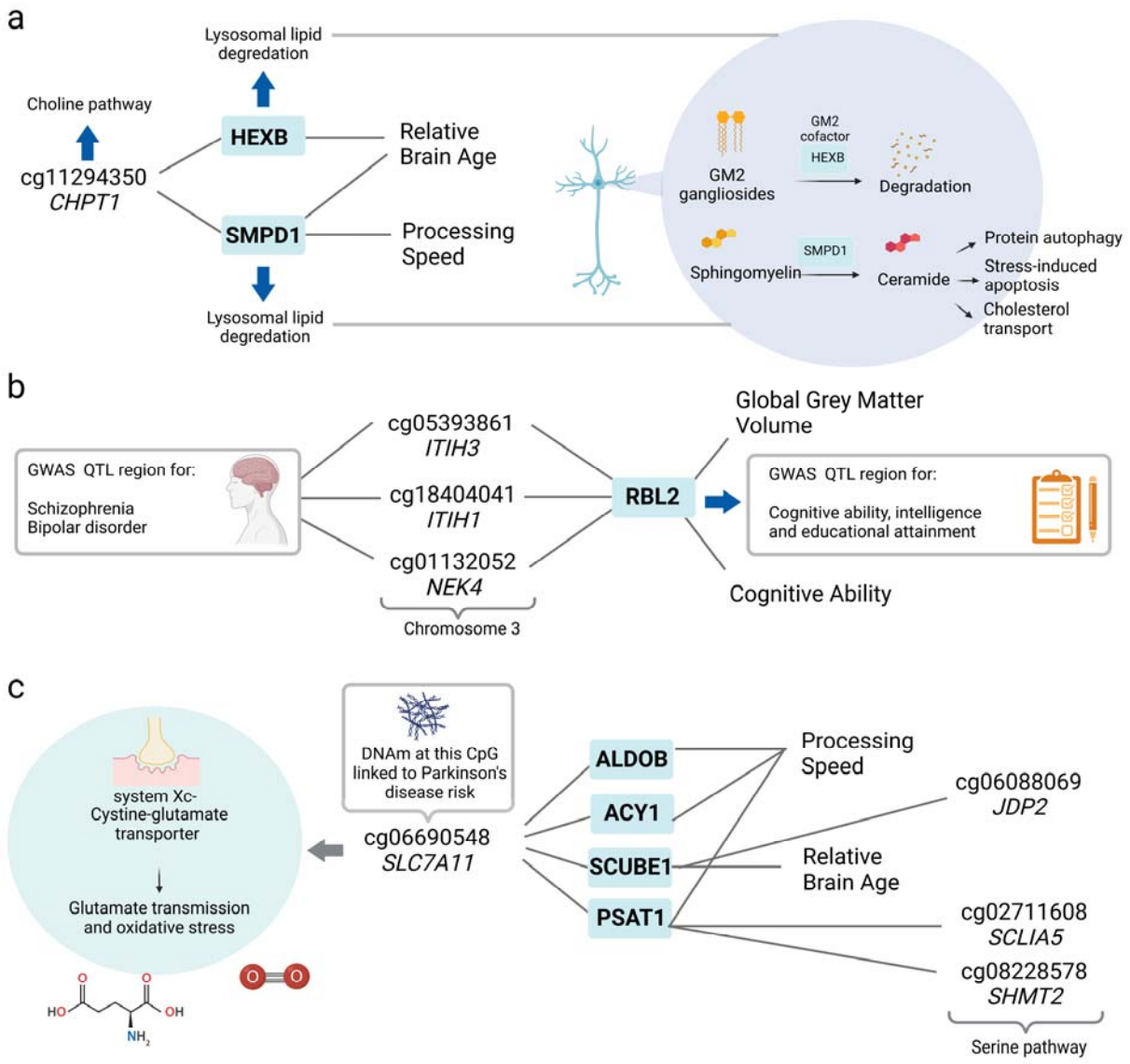
a, Standardised beta coefficients plotted from phenome-wide protein association studies (PheWAS) between 4,235 protein levels and 14 neurologically-relevant phenotypes in Generation Scotland (maximum n=1,095). Negative and positive direction of effects are shown in blue and red, respectively. Associations for 25 unique proteins that were associated (FDR P < 0.05) with both a cognitive and imaging modality and are indicated by an asterisk**. b,** Four of the 11 protein markers associated with *APOE* status with FDR P < 0.05. Positive associations (orange) indicate a relationship between APOB levels and the presence of one or more *APOE* e4 alleles. Negative associations (blue) indicate inverse relationships between protein levels and *APOE* e4 status. **c,** A selection of associations (FDR P < 0.05) for the RBL2 and TREM1 protein markers in relation to both cognitive and imaging modalities.

Fig. 4. pQTMs for protein markers of neurological phenotypes



Circular plot showing 42 *trans* pQTM associations between DNAm at CpG sites and the levels of 14 proteins that were associated with one of more of the neurological phenotypes (FDR P < 0.05). Chromosomal positions are given on the outermost circle. Cis pQTM associations are available in **Supplementary Fig. 1**. Full details of the 88 pQTMs are reported in **Supplementary Table 16**.

Fig. 5. Candidate markers revealed through neurological pQTM associations



a, Two *trans* associations between DNAm at cg11294350 in the *CHPT1* gene and two proteins with lysosomal-associated function (HEXB and SMPD1) that were associated with relative brain age and processing speed. **b,** Three *trans* associations between the region *ITIH3/ITIH1/NEK4* on chromosome 3 and the levels of RBL2, which was associated with reductions in both Global Grey Matter Volume and Cognitive Ability (*g*-score). **c,** Four *trans* associations revolving around the CpG site cg06690548 in the *SLC7A11* gene, which encodes for a synaptic protein involved in glutamate transmission and oxidative stress. The seven *trans* associations for *NLCR5* inflammatory-associated CpG sites are not shown, but can be viewed in **Supplementary Fig. 2**.