medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health

Danni A Gadd[1,], Robert F Hillary[1], Daniel L McCartney[1], Liu Shi[2], Aleks Stolicyn[3], Neil Robertson[4], Rosie M Walker[5], Robert I McGeachan[6], Archie Campbell[1], Shen Xueyi[3], Miruna C Barbu[3], Claire Green[3], Stewart W Morris[1], Mathew A Harris[3], Ellen V Backhouse[5], Joanna M Wardlaw[5,7,8,9], J Douglas Steele[10], Diego A Oyarzún[11,12,13], Graciela Muniz-Terrera[14], Craig Ritchie[14], Alejo Nevado-Holgado[2], Tamir Chandra[4], Caroline Hayward[1,15], Kathryn L Evans[1], David J Porteous[1], Simon R Cox[16,17], Heather C Whalley[3], Andrew M McIntosh[3], 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] Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, EH10 5HF, UK

[4] MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU

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

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

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

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

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

[10] Division of Imaging Science and Technology, Medical School, University of Dundee, Dundee, UK, DD1 9SY

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

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

[13] The Alan Turing Institute, 96 Euston Road, London, NW1 2DB

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

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

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Abstract

Characterising associations between the methylome, proteome and phenome may provide insight into biological pathways governing brain health. Here, we report an integrated DNA methylation and phenotypic study of the circulating proteome in relation to brain health. Methylome-wide association studies of 4,058 plasma proteins are performed (N=774), identifying 2,928 CpG-protein associations after adjustment for multiple testing. These were independent of known genetic protein quantitative trait loci (pQTLs) and common lifestyle effects. Phenome-wide association studies of each protein are then performed in relation to 15 neurological traits (N=1,065), identifying 405 associations between the levels of 191 proteins and cognitive scores, brain imaging measures or *APOE* e4 status. We uncover 35 previously unreported DNA methylation signatures for 17 protein markers of brain health. The epigenetic and proteomic markers we identify are pertinent to understanding and stratifying brain health.

Introduction

The health of the ageing brain is associated with risk of neurodegenerative disease [1,2]. Relative brain age – a measure of brain health calculated using multiple volumetric brain imaging measures – has recently been shown to predict the development of dementia [3]. Structural brain imaging and performance in cognitive tests are well-characterised markers of brain health [4], which clearly associate with potentially modifiable traits such as body mass index (BMI), smoking and diabetes [5–7]. Understanding the interplay between environment, biology and brain health may therefore inform preventative strategies.

Multiple layers of omics data indicate the biological pathways that underlie phenotypes. Proteomic blood sampling can track peripheral pathways that may impact brain health, or record proteins secreted from the brain into the circulatory system. Although proteome-wide characterisation of cognitive decline and dementia risk [8–10] have been facilitated at large-scale by SOMAscan® protein

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

measurements, there is a need to further integrate omics to characterise brain health phenotypes. Epigenetic modifications to the genome record an individual's response to environmental exposures, stochastic biological effects, and genetic influences. Epigenetic changes include histone modifications, non-coding RNA, chromatin remodelling, and DNA methylation (DNAm) at cytosine bases, such as 5-hydroxymethylcytosine. These are implicated in changes to chromatin structure and the regulation of pathways associated with neurological diseases [11,12]. However, DNAm at cytosine-guanine (CpG) dinucleotides is the most widely profiled blood-based epigenetic modification at large scale.

Modifications to DNAm at CpG sites play differential roles in influencing gene expression at the transcriptional level [13]. Additionally, DNAm accounts for inter-individual variability in circulating protein levels [14–16]. Recently, through integration of DNAm and protein data, we have shown that epigenetic scores for plasma protein levels – known as 'EpiScores' – associate with brain morphology and cognitive ageing markers [17] and predict the onset of neurological diseases [18]. These studies highlight that while datasets that allow for integration of proteomic, epigenetic and phenotypic information are rarely-available, they hold potential to advance risk stratification. Integration may also uncover candidate biological pathways that may underlie brain health.

Associations between protein levels and DNAm at CpGs are known as protein quantitative trait methylation loci (pQTMs) and can be quantified by methylome-wide association studies (MWAS) of protein levels. The largest MWAS of protein levels to date assessed 1,123 SOMAmer protein measurements in the German KORA cohort (n=944) [14]. In that 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 closely related to chronic systemic inflammation. Hillary *et al* have also assessed epigenetic signatures for 281

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

SOMAmer protein measurements that were previously associated with Alzheimer's disease, in the Generation Scotland cohort that we utilised in this study [19]. However, proteome-wide assessment of pQTMs has not been tested against a comprehensive spectrum of brain health traits.

Here, we conduct an integrated methylome- and phenome-wide assessment of the circulating proteome in relation to brain health (Fig. 1), using 4,058 protein level measurements (Supplementary Table 1). We characterise CpG-protein associations (pQTMs) for these proteins in 774 individuals from the Generation Scotland cohort using EPIC array DNAm at 772,619 CpG sites. We then identify which of the 4,058 protein levels associate with one or more of 15 neurological traits (seven structural brain imaging measures, seven cognitive scores and *APOE* e4 status) in 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. For these signatures, we map potential underlying genetic components and chromatin interactions that may play a role in protein level regulation. A YouTube video summarising the study and detailing access to the datasets can be viewed at https://www.youtube.com/channel/UCxQrFFTIItF25YKfJTXuumQ.

Results

Methylome-wide studies of 4,058 plasma proteins

We conducted MWAS to test for pQTM associations between 772,619 CpG sites and 4,058 circulating protein levels (corresponding to 4,235 SOMAmer measurements; Supplementary Table 1). The MWAS population included 774 individuals from Generation Scotland **(**mean age 60 years [SD 8.8], 56% Female; Supplementary Table 2**).** 143 principal components explained 80% of the cumulative variance in the 4,235 measurements (Supplementary Fig. 1 and Supplementary Table 3). A threshold for multiple testing based on these components was applied across all MWAS ($P < 0.05/(143 \times 772{,}619) = 4.5 \times 10^{-10}$).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

In our basic model adjusting for age, sex and available genetic pQTL effects from Sun *et al* [20] 238,245 pQTMs (2,107 *cis* and 236,138 *trans*, representing 0.005% of tested associations) had $P < 4.5x10^{-10}$ (Supplementary Table 4). In our second model that further adjusted for Houseman-estimated white blood cell proportions[21], there were 3,213 associations (453 *cis* and 2,760 *trans*) that had $P < 4.5x10^{-10}$ (Supplementary Table 5). Smoking status and BMI are known to have well-characterised DNAm signatures [22,23]; fully-adjusted models were therefore further adjusted for these factors. There were 2,928 associations (451 *cis* and 2,477 *trans*) in the fully-adjusted models (Supplementary Table 6). 2,847 pQTM associations were significant in all models**.** Figure 2 summarises these findings.

There were 191 unique proteins with associations in the fully-adjusted models, corresponding to 195 SOMAmer measurements (two SOMAmers were present for CLEC11A, GOLM1, ICAM5 and LRP11). Genomic inflation statistics for these 195 SOMAmer measurements (fully-adjusted MWAS) are presented in Supplementary Table 7. In a sensitivity analysis, restriction of the threshold for *cis* pQTMs from 10Mb to 1Mb from the transcription start site of the gene encoding the protein yielded 409 cis pQTMs (a drop of 42 pQTMs) in the fully-adjusted MWAS. A summary of known pQTLs [24] and a record of whether these were available for adjustment is provided in Supplementary Table 8. Characterising the genomic location of the findings, 46% of *cis* and 29% of *trans* pQTMs in the fully-adjusted MWAS involved CpGs positioned in either a CpG Island, Shore or Shelf (Supplementary Table 6).

Pleiotropic pQTM associations in the fully-adjusted MWAS

Pleiotropy was observed for both CpG sites and protein levels (Fig. 3). Nineteen proteins had 10 or more pQTMs in the fully-adjusted MWAS (Supplementary Table 9). Of the 2,928 pQTMs in the fully-adjusted MWAS, 987 involved Pappalysin-1 (PAPPA) and there were a further 1,116 pQTMs that involved the Proteoglycan 3 Precursor (PGR3) protein. The remaining 825 pQTMs involved 189

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

unique protein levels, with 434 *cis* and 391 *trans* associations (Fig. 2**)**. Principal components analyses indicated high correlations between CpGs associated with the pleiotropic proteins PAPPA and PRG3, whereas the CpGs involved in the remaining 825 pQTMs were largely uncorrelated (Supplementary Fig. 2). pQTM frequencies for the 1,837 unique CpGs selected in the fully-adjusted models, with their respective genes and EWAS catalogue [25] lookup of epigenome-wide significant ($P < 3.6x10^{-8}$) phenotypic associations is presented in Supplementary Table 10. Of these CpGs, sites within the *NLRC5*, *SLC7A11* and *PARP9* gene regions exhibited the highest levels of pleiotropy (Fig. 3).

The pleiotropic findings for PAPPA and cg07839457 (*NLRC5* gene) replicated previous MWAS results from Zaghlool *et al* [14] (944 individuals, with 1,123 protein SOMAmers). Of the 98 pQTMs identified by Zaghlool *et al*, 81 were comparable (both the protein and CpG sites from the 98 pQTMs were available across both MWAS). Of these 81 pQTMs, 26 replicated at our significance threshold ($P < 4.5x10^{-10}$) with the same direction of effect, a further 16 replicated at the epigenome-wide significance threshold ($P < 3.6x10^{-8}$) [26] and a further 39 replicated at nominal $P < 0.05$ (Supplementary Table 11 and Supplementary Fig. 3). When accounting for 26 pQTMs that were previously reported by Zaghlool et al and 10 pQTMs that were previously reported by Hillary et al [14,19], 2,892 of the 2,928 fully-adjusted pQTMs were novel. Of these 2,892 novel pQTMs, 1,109 involved the levels of 41 proteins that were measured by Zaghlool et al (973 pQTMs for PAPPA and 136 additional pQTMs for the levels of 40 proteins), whereas 1,783 pQTMs involved the levels of proteins that were previously unmeasured (1,116 pQTMs for PRG3 and 667 further pQTMs for 148 proteins).

Proteome associations with brain health phenotypes

We next conducted a proteome-wide association study of brain health characteristics (protein PheWAS of brain imaging, cognitive scoring and *APOE* e4 status, alongside age and sex; Fig. 4).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Distribution plots for the seven cognitive scores and seven brain imaging phenotypes are presented in Supplementary Figs. 4-5. A maximum sample of 1,065 individuals was available **(**mean age 59.9 years [SD 9.6], 59% Female; Supplementary Table 2**)**; all 774 individuals from the pQTM study were included in these analyses. A threshold for multiple testing adjustment was calculated based on 143 independent components that explained >80% of the 4,235 SOMAmer levels (Supplementary Table 3 and Supplementary Fig. 1). This equated to $P < 0.05/(143) = 3.5x10^{-4}$. The levels of 587 plasma proteins were associated with age and 545 were associated with sex, with 222 proteins common to both phenotypes **(**Supplementary Table 12**)**. When comparable associations from three studies (with N>1000) were tested [20,27,28], 97% of age and 98% of sex associations replicated in one or more of studies (Supplementary Table 12).

There were 191 unique protein markers that had a total of 405 associations with brain health characteristics (Supplementary Fig. 6 and Fig. 4a). These consisted of 95 brain imaging (Supplementary Table 13), 296 cognitive test score (Supplementary Table 14) and 14 *APOE* e4 status (Supplementary Table 15) associations. Supplementary Table 16 stratifies these associations by direction of effect and Supplementary Table 17 provides full summary statistics for all 405 associations. Of the seven brain morphology traits, Relative Brain Age and General Fractional Anisotropy (gFA) had the largest number of associations, with 24 and 22 protein markers identified, respectively. Of the cognitive score traits, Processing Speed and General Cognitive Ability scores were associated with the highest number of protein markers (102 and 73, respectively). The 14 *APOE* e4 status associations are plotted in Supplementary Fig. 7.

Stratifying the 405 associations by direction of effect revealed that the majority (89%) of associations indicated that higher levels of the proteins were associated with less favourable brain health (Supplementary Table 16). Eighty-seven of the 405 associations involved protein leves that were associated with more favourable brain health; this signature included the levels of SLITRK1,

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

NCAN and COL11A2. Higher levels of ASB9, RBL2, HEXB and SMPD1 associated with poorer brain health. Protein interaction network analyses for the genes corresponding to the 191 protein markers (Supplementary Fig. 8) indicated that many of the proteins in these signatures clustered together, implying shared underlying functions. An inflammatory cluster including CRP, ITIH4, C3, C5, COL11A2 and SIGLEC2 was present and higher levels of these markers were associated with poorer brain health outcomes. Gene set enrichment analyses on the 191 genes corresponding to the protein markers (Supplementary Fig. 9) supported the link between many of the proteins associated with poorer brain health and the innate immune system, while also implicating extracellular matrix, lysosomal, metabolic and additional inflammatory pathways. Tissue expression profiles of the 191 genes (Supplementary Fig. 10**)** indicated that many of the markers were expressed non-neurological tissues; however, some proteins were expressed in nervous tissues. Markers such as ASB9 and NCAN were found to be consistently identified across multiple brain imaging traits as markers of poorer and better brain health, respectively (Supplementary Table 16). While many of the associations for brain imaging measures identified proteins that were distinct from those found for cognitive scores and *APOE* e4 status, 22 protein markers were associated with both a cognitive score and a brain imaging trait (Fig. 4b and Supplementary Table 18). Of these 22 proteins, there were 10 principal components that had a cumulative variance of >80% and five components had eigenvalues > 1 (Supplementary Fig. 11). Three *APOE* e4 status markers (ING4, APOB and CRP) were also associated with cognitive scores (Fig. 4b).

Replication of protein PheWAS associations

Six of the 14 *APOE* e4 status associations replicated previous SOMAmer protein findings (N SOMAmers= 4,785 and N participants=227) [10], and eight novel relationships involved NEFL, ING4, PAF, MENT, TMCC3, CRP, FAM20A and PEF1. Several of the markers for cognitive function were identified in previous work relating Olink proteins to cognitive function (such as CPM) [29] and

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

work that characterised SOMAmer signatures of cognitive decline and incident Alzheimer's disease (such as SVEP1) [8]. No studies have performed SOMAmer-based, whole proteome PheWAS studies of the brain imaging and cognitive score traits we have profiled in a heathy ageing population that were not enriched for neurodegenerative diseases. However, replication of associations from several studies [9,29,30] was found for a small subset of associations (Supplementary Table 19).

Integration of the brain health proteome with our pQTM dataset

Differential DNAm signatures were explored for the 191 protein markers that had $P < 3.5x10^{-4}$ in associations with either cognitive scores, brain imaging measures or *APOE* e4 status in the protein PheWAS. Of the 191 proteins, 17 had pQTMs in the fully-adjusted MWAS. Higher levels of 15 of these proteins were associated with poorer brain health, while AMY2A and CST5 were associated with more favourable brain health. There were a total of 35 pQTMs involving 31 unique CpGs that were located within 20 distinct genes (Supplementary Table 20), with 15 *trans* (Fig. 5) and 20 *cis* associations. All pQTMs were previously unreported. The 20 *cis* pQTMs involved the levels of CHI3L1, IL18R1, SIGLEC5, OLFM2, UGDH, CRHBP, AMY2A and CFHR1 proteins. The *trans* pQTMs involved the levels of SCUBE1, RBL2, TNFRSF1B, CST5, HEXB, ACY1, CRTAM, SMPD1 and RBP5 proteins.

Of the 20 cis pQTMs, 11 involved CpGs in different genes to the protein-coding gene on the same chromosome, whereas the remaining 9 pQTMs involved CpGs located within the protein-coding gene. Several CpG sites were associated with multiple protein levels in the *trans* pQTMs (Fig. 5). DNAm at site cg06690548 in the *SLC7A11* gene was associated with RBP5, ACY1 and SCUBE1 levels. The cg11294350 site in the *CHPT1* gene was associated with HEXB and SMPD1 levels. The cg07839457 site in the NLRC5 gene was associated with the levels of CRTAM and TNFRSF1B. There was also a protein that had several *trans* associations with multiple CpG sites; pQTMs were

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

identified between circulating RBL2 levels and cg01132052, cg0539861, cg18487916, cg27294008 and cg18404041, within the *NEK4/ITIH3/ITIH1* gene region of chromosome 3.

Functional mapping of neurological pQTMs

A lookup that integrated information from the GoDMC and eQTLGen databases assessed whether pQTMs were partially driven by an underlying genetic component. This identified methylation quantitative trait loci (mQTLs) for CpGs that were associated with CHI3L1, IL18R1 and SIGLEC5 and were also expression quantitative trait loci (eQTLs) for the respective protein levels (Supplementary Table 20). Further visual inspection of the distributions for the 35 pQTMs indicated that trimodal distributions – suggestive of unaccounted SNP effects – were present for CpGs involved in seven of the pQTMs (Supplementary Fig. 12).

Tissue expression profiles for the 33 genes that were linked to either CpGs or proteins in the 35 neurological pQTMs are summarised in Supplementary Fig. 13. Gene set enrichment for these 33 genes identified enrichment for immune effector pathways in a subset of 11 genes, whereas a cluster of four genes (SMPD1, HEXB, AMY2A and AMY2B) were enriched for amylase and hydrolase activity (Supplementary Fig. 14).

Of the 35 pQTMs, seven had CpGs that were located in either a CpG Shore or Shelf position and there were 13 that were located either 1500 bp or 200 bp from the TSS of the protein-coding gene (Supplementary Table 20). Fifteen pQTMs involved CpGs that were located in the gene body and 7 were located in either the first exon or UTR regions (Supplementary Table 20).

Promoter-capture Hi-C and ChIP-sequencing integration was used to assess the interactions and chromatin states of our pQTMs and associated CpG loci. This analysis focused on 11 of the 20 *cis* pQTMs that involved CpGs on the same chromosome as the protein-coding gene, but were located in a different gene. Mapping information is presented for the seven proteins involved in these pQTMs

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

in Supplementary Figs. 15-21. In all instances, we found evidence of spatial co-localisation of these genes using promoter-capture Hi-C data from brain hippocampal tissue. We attempted to contextualise these sites further with ChIP-seq (ENCODE project) analyses of active chromatin marks H3K27ac and H3K4me1 and repressive chromatin H3K4me3 and H3K27me3 in both peripheral blood mononuclear cells (PBMCs) and brain hippocampus. ChIP-seq data suggested that in many instances there were shared regulatory regions that existed across both blood and hippocampal samples that were hubs for local promoter interactions. For example, promoter loops were found linking the S100Z and CRHBP genes, with a signature of activating (H3Kme1 and H3K27ac) and silencing (H3k27me and H3K4me3) marks (normally considered bivalent chromatin) that may form the basis for shared regulation of this gene locus.

Discussion

We have conducted a large-scale integration of the circulating proteome with indicators of brain health and blood-based DNA methylation. We characterised 191 protein markers that were associated with either brain imaging measures, cognitive scores or *APOE* e4 status in an ageing population. We also report methylome-wide characterisations for the SOMAscan® panel V.4 (4,058 protein measurements) in a nested subset of this population. By integrating these datasets, we uncovered 35 methylation signatures for 17 protein markers of brain health. We delineated pQTM CpGs that had evidence of underlying genetic influence and characterised the potential for chromatin interactions for genes involved in *cis* pQTMs. As this population consists of older individuals that were not enriched for neurodegenerative diseases, the markers we identify are likely indicators of healthy brain ageing.

Many of the 191 proteins identified in the protein PheWAS were part of inflammatory clusters with shared functions in acute phase response, complement cascade activity, innate immune activity and cytokine pathways. Tissue expression analyses suggested that a large proportion of the 191 protein

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

markers were not expressed in the brain; this supports work suggesting that sustained peripheral inflammation influences general brain health [3132] and accelerates cognitive decline [8,33–35]. However, a subset of proteins were expressed in the central nervous system. Given that leakage at the blood-brain-barrier interface has been hallmarked as a part of healthy brain ageing [36,37], there is a possibility that brain-derived proteins may enter the bloodstream as biomarkers. SLIT and NTRK Like Family Member 1 (SLITRK1), Neurocan (NCAN) and IgLON family member 5 (IGLON5) were examples of proteins expressed in brain for which higher levels associated with either larger grey matter volume, larger whole brain volume, or higher general fractional anisotropy. SLITRK1 localises at excitatory synapses and regulates synapse formation in hippocampal neurons [38]. Neurocan (NCAN) is a component of neuronal extracellular matrix and is linked to neurite growth [39]. IGLON5 has been implicated in maintenance of blood-brain-barrier integrity and an anti-IGLON5 antibody disease involves the deterioration of cognitive health [40]. Taken together, the protein markers identified in the PheWAS may, therefore, reflect pathways that could be targeted to improve brain health.

Integration of our fully-adjusted protein MWAS dataset revealed 35 associations between DNAm and 17 protein markers of brain health (Fig. 6; Supplementary Table 20). All 35 associations were novel. While this study is focused on blood DNAm – limiting generalisation to brain DNAm – many of the 35 pQTMs involved CpGs and proteins that have been previously implicated in neurological processes. DNAm at site cg06690548 (located in the *SLC7A11* gene) was of particular interest; differential DNAm at this CpG in blood has been identified as a causal candidate for Parkinson's disease (N > 900 cases and N > 900 controls) [41]. Xc- is the cystine-glutamate antiporter encoded by *SLC7A11*, which facilitates glutamatergic transmission, oxidative stress defence and microglial response in the brain [4243] and is a target for the neurodegeneration-associated environmental neurotoxin β-methylamino-L-alanine [41]. Analyses in the wider Generation Scotland cohort suggests that cg06690548 is a site associated with alcohol consumption [44]. The proteins associated with

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

cg06690548 in the subset of this cohort we assessed (ACY1, SCUBE1 and RBP5) have known links to liver function [45–47]. DNAm at cg06690548 in blood has also been recently implicated in the largest MWAS of amyotrophic lateral sclerosis (ALS) to date (6,763 cases, 2,943 controls) [48]. Given that ACY1, SCUBE1 and RBP5 were markers for either lower processing speed and higher relative brain age, the CpG sites we identify in this study – such as cg06690548 – may be important plasma markers for mediation of environmental risk on brain health that merit further exploration. cg06690548 lies within the first intron of *SLC7A11* [41], indicating that this site is of potential functional significance.

The presence of *NLRC5* and various other inflammatory proteins in our neurological protein pQTMs suggests that the methylome may capture an inflammatory component of brain health. Many of the genes corresponding to CpGs and proteins involved in the 35 pQTMs were enriched for immune effector processes and were not expressed in brain. However, some markers did show evidence for brain-specific expression, such as acid sphingomyelinase (SMPD1) and Hexosaminidase Subunit Beta (HEXB). The HEXB and SMPD1 proteins associated with DNAm at cg11294350 (in the *CHPT1* gene), are involved in neuronal lipid degradation in the brain and have been associated with the onset of a range of neurodegenerative conditions [49–52]. RBL2 is another protein that had partial expression signals across brain regions; the *NEK4/ITIH3/ITIH1* region was the location for five CpGs with differential DNAm linked to RBL2 levels. This region is implicated in schizophrenia and bipolar disorder by several large-scale, genome-wide association studies (GWAS) [53–56]. Similarly, the *RBL2* locus has been associated with intelligence, cognitive function and educational attainment in GWAS (n > 260,000 individuals) [33,57,58].

Given that this study utilised CpGs from the Illumina EPIC array, 15 of the 31 unique CpGs did not have mQTL characterisations in public databases, which primarily comprise results from the earlier 450K array. However, our plots showing pQTM associations suggested that for several CpGs (such

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

as cg11294350 that associated with SMPD1 and HEXB), there may be a partial genetic component influencing DNAm. As mQTLs tend to explain 15-17% of the additive genetic variance of DNAm [59], it is possible that the signals we isolate in these instances are partially driven by genetic loci, but are also likely driven by unmeasured environmental and biological influences. In the case of SIGLEC5, IL18R1 and CHI3L, mQTLs were identified that were also eQTLs, providing evidence that mQTLs for these CpG sites were possible regulators of protein expression.

Integration of promoter-capture Hi-C chromatin interaction and ChIP-seq databases [60] provided evidence for long-range interaction relationships for *cis* pQTMs with CpGs in different gene regions that are proximal to the protein-coding gene of interest. This suggests that in such instances, the pQTMs may reflect regulatory relationships in the 3-dimensional genomic neighbourhood. The pQTMs therefore direct us towards pathways that can be tested in experimental constructs. Positional information suggested that many CpGs involved in neurological pQTMs lay within 1500 bp of the TSS of the respective protein-coding gene. While positional information of CpGs is thought to infer whether DNAm is likely to play a role in the expression regulation of nearby genes, this is still somewhat disputed. Some studies suggest that transcription factors regulate DNAm [61] and differential methylation at gene body locations predicts dosage of functional genes [62]. Additionally, the DNAm signatures of proteins we quantify represent widespread differences across blood cells that are related to circulating protein levels and are therefore not derived from the same cell-types as proteins. Despite this limitation, previous work supports DNAm scores for proteins as useful markers of brain health, suggesting there is merit in integrating DNAm signatures of protein levels in disease stratification [18].

Our study has several limitations. First, though full replication of our results was not possible, our replication of pQTMs identified by Zaghlool *et al* [14] reinforces inflammation signalling as intrinsic to the methylome signature of blood proteins. This also suggests that pQTMs may be common across

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

ancestries. Second, we observed a substantial inflation for PAPPA and PRG3 proteins. While comprehensive adjustment for estimated immune cells was performed and the remainder of CpGs involved in pQTMs did not show high correlations (Supplementary Fig. 2), concurrently measured blood components such as haemoglobin, red blood cells and platelets were not available. Future studies should seek to resolve signals with more detailed blood-cell phenotyping and immune cell estimates [63]. Third, 89% 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 blood immune cell DNAm, 2) underpowered analyses, or 3) the presence of indirect associations that are not captured by our MWAS approach. Fourth, the extent of non-specific and cross-aptamer binding with SOMAmer technology has not been fully resolved [64]. Fifth, there are likely unknown genetic influences on pQTMs. Further characterisation of pQTLs and advances in multi-omic modelling techniques [15] will aid in the separation of genetic and environmental influences on epigenetic signatures. Sixth, differences in blood and brain DNAm and pQTLs are emerging; these indicate that blood-based markers may not fully align to biology of brain degeneration [65,66]. However, our ChIP-seq analysis of chromatin regulation suggested that some regulatory states may persist between blood and brain. Seventh, profiling DNAm signatures alone cannot capture the full role of the epigenome in brain health. Integration of more diverse epigenetic markers will be critical to further resolve these relationships. Finally, though we have incorporated a wide portfolio of brain health measures, we recognise that these are not extensive. Increasing triangulation across modalities, as we have shown here, will be useful in identifying candidate markers.

In conclusion, by integrating epigenetic and proteomic data with cognitive scoring, brain morphology and *APOE* e4 status, we identify 191 protein markers of brain health. We characterise DNAm signatures for all 4,058 proteins included in the study, uncovering 35 associations between differential DNAm and the levels of 17 of the protein markers of brain health outcomes. These data

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

identify candidate targets for the preservation of brain health and may inform risk stratification approaches.

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 [67]. 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 [68]. Of these, N=1,065 individuals had proteomic data available and N=778 of these had DNAm data available. Supplementary Table 2 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) [69]. SOMAmers were processed for 1,065 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. Full details of quality control stages are provided in Supplementary Methods. In the final dataset, 4,235 SOMAmer epitope measures were available in 1,065 individuals and these

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

corresponded to 4,058 unique proteins (classified by common Entrez gene names). Supplementary Table 2 provides annotation information for the 4,235 SOMAme measurements that were available.

DNAm measurement

Measurements of blood DNAm in the STRADL subset of GS subset were processed in two sets on the Illumina EPIC array using the same methodology as those collected in the wider Generation Scotland cohort. Quality control details have been reported previously [70–72] and further details are provided in Supplementary Methods. Briefly, samples were removed if there was a mismatch between DNAm-predicted and genotype-based sex and all non-specific CpG and SNP probes (with allele frequency > 5%) were removed from the methylation file. Probes which had a beadcount of less than 3 in more than 5% of samples and/or probes in which >1% of samples had a detection P>0.01 were excluded. 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 joined together for use in the MWAS, with 476 individuals included in set 1 and 298 individuals in set 2. DNAm-specific technical variables (measurement batch and set) were adjusted in all MWAS and PheWAS models.

Phenotypes in Generation Scotland

All phenotypes in Generation Scotland MWAS and PheWAS samples are summarised in Supplementary Table 2. An epigenetic score for smoking exposure, EpiSmokEr [73] was calculated for all individuals with DNAm. The meffil [74] implementation of the Houseman method was used to calculate estimated white blood cell (WBC) proportions for Sets 1 and 2. Blood reference panels were sourced from Reinius et al [75]. The 'blood gse35069 complete' panel was used to imputed measures for Monocytes, Natural Killer cells, Bcells, Granulocytes, $CD4^+$T cells and $CD8^+$T cells. Eosinophil and Neutrophil estimates were also sourced through the 'blood gse35069' panel. Body mass index (body weight in kilograms, divided by squared height in metres) was available for all

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

individuals, alongside depression status (defined using a research version of the Structured Clinical Interview for DSM disorders (SCID) assessment), which was coded as a binary variable of no history of depression (0) or lifetime episode of depression (1). Five individuals did not have depression status information and were excluded from MWAS and PheWAS analyses, where appropriate. *APOE* e4 status was available for 1,050 individuals. *APOE* e4 status 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 (Supplementary Fig. 4; Supplementary Table 2) measured at the clinic visit for the STRADL subset of GS were considered. Full details for the specific scores has been detailed previously [68] and further details can be found in Supplementary Methods. Briefly, these included the Wechsler Logical Memory Test (maximum possible score of 50), the Wechsler Digit Symbol Substitution Test (maximum possible score of 133), the verbal fluency test (based on the Controlled Oral Word Association task), the Mill Hill Vocabulary test (maximum possible score of 44) and the Matrix Reasoning test (maximum possible score of 15). Outliers were defined as scores >3.5 standard deviations above or below the mean and were removed prior to analysis. The first unrotated principal component combining logical memory, verbal fluency, vocabulary and digit symbol tests 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 model. 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 greater relevance for assessing cognitive decline in ageing individuals.

The derived brain volume measures (Supplementary Fig. 5; Supplementary Table 2) were recorded at two sites (Aberdeen and Edinburgh) [68]. Brain volume data included total brain volume (ventricle volumes excluded), global grey matter volume, white matter hyperintensity volume and total

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

intracranial volume. Intracranial volume was treated as a covariate to adjust for head size in all tests including brain volume associations. The derived global white matter integrity measures included global fractional anisotropy (gFA) and global mean diffusivity (gMD). The protocols applied to derive the brain volume measures from T1-weighted scans, and white matter integrity measures from diffusion tensor imaging (DTI) scans have been previously described [35,68,76] and additional details are also provided in Supplementary Methods. 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 [3]. This estimate was regressed on chronological age to produce a measure of Relative Brain Age (residuals from the linear model). Outliers for all imaging variables were defined as measurements >3.5 standard deviations above or below the mean and were removed prior to analysis.

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 [77]. 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* e4 status variable (e2 = 0, e3 = 1, e4 = 2), cognitive and brain imaging phenotypes were included as scaled predictors. Continuous variables were scaled to mean of zero and variance one. 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

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

458 phenotypes in STRADL [68]. Clinic study site and protein lag group (storage time before proteomic
459 sequencing) were included as covariates in all models. Missing data were excluded from lmekin
460 models.

461 Three regression models were considered. For the analyses with age and sex as the predictors of
462 interest, two coefficients (β1 and β2) were extracted.

463 Protein level ~ Intercept + β1 age + β2 sex + depression + study site + lag group + (1|individual) +
464 (1|kinship)

465 For the analyses with one of cognitive scoring, *APOE* e4 status or non-volumetric brain imaging as
466 the predictor of interest, the β3 coefficient was extracted.

467 Protein level ~ Intercept + age + sex + β3 phenotype + depression + study site + lag group +
468 (1|individual) + (1|kinship)

469 Finally, for the models with volumetric brain imaging measures as the predictor of interest the β4
470 coefficient was extracted.

471 Protein level ~ Intercept + age + sex + β4 phenotype + depression + study site*ICV + lag group +
472 edited + batch + (1|individual) + (1|kinship)

473 All analyses of brain volume measures included adjustment for intracranial volume and study site as
474 main effects, in addition to the interaction between these variables. ICV was used to account for head
475 size. Imaging data processing batch, and presence or absence of manual intervention during quality
476 control (edited) variables were also included as covariates, wherever appropriate.

477 P-values for all PheWAS models were calculated in R using effect size estimates (beta) and standard
478 errors (SE) as follows: pchisq((beta/SE)^2, 1, lower.tail=F). The Prcomp package [78] was used to
479 generate principal components for the 4,235 SOMAmer measurements (N=1,065). 143 components

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

explained >80% of the cumulative variance in protein levels (Supplementary Fig. 1 and Supplementary Table 3); these components were used to derive the PheWAS multiple testing adjustment threshold of $P < 0.05 / 143 = 3.5x10^{-4}$. This method was chosen due to the presence of high intercorrelations within the protein data.

Epigenome-wide association study of protein levels

Prior to running the MWAS, protein levels for 774 individuals with complete phenotypic information were log transformed and regressed on 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 and taken forward as protein level data. Methylation data were in M-value format and were pre-adjusted for age, sex, processing batch, methylation set, depression status [73] and known pQTL effects (from a previous genome-wide association study of 4,034 SOMAmers targeting 3,622 proteins from Sun *et al*) [24] in the basic MWAS. A second model further adjusted for estimated white blood cell proportions (Monocytes, $CD4^+$T cells, $CD8^+$T cells, BCells, Natural Killer cells, Granulocytes and Eosinophils). While Neutrophil estimates were available, they were excluded due to high correlation (>95%) with Granulocyte proportions **(**Supplementary Fig. 22**)**. Finally, the fully-adjusted model further regressed DNAm onto an epigenetic score for smoking, EpiSmokEr[73] and body mass index (BMI).

Omics-data-based complex trait analysis (OSCA) [79] Version 0.41 was used to run EWAS analyses. Within OSCA, a genetic relationship matrix (GRM) was constructed for the STRADL population. 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 random effect to adjust for

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

relatedness between individuals [79]. CpG sites were the dependent variables and the 4,235 proteins were the independent variables.

Five fully-adjusted models did not converge (NAGLU, CFHR2, DAP, MST1, PILRA) and were excluded. A threshold for multiple testing correction ($P < 4.5x10^{-10}$) was based on 143 independent protein components with cumulative variance >80% (Supplementary Fig. 1 and Supplementary Table 3) ($P < 0.05/(143x772,619)$ CpGs). A more conservative threshold based on total number of SOMAmers was also considered ($P < 0.05/(4,235x772,619) = 1.5x10^{-11}$) and is detailed in Supplementary Tables 4-6. pQTMs were classified as *cis* if the CpG was on the same chromosome as the protein-coding gene and fell within 10Mb of the transcriptional start site (TSS) of the protein gene. pQTMs involving a CpG located on a different chromosome to the protein-coding gene, or >10Mb from the TSS of the protein gene were classed as *trans*.

Circos plots were created with the circlize package (Version 0.4.12) [80]. BioRender.com was used to create Figs. 1, 2, 3 and 6. All analyses were performed in R (Version 4.0) [81].

Functional mapping and tissue expression analyses

Functional mapping and annotation (FUMA) [82] gene set enrichment analyses were conducted for genes corresponding to protein markers that were identified through the PheWAS study, in addition to genes linked to either CpGs or proteins in the neurological pQTM subset. Protein-coding genes were selected as the background set and ensemble v92 was used with a false discovery rate (FDR) adjusted $P < 0.05$ threshold for gene set testing. For the genes corresponding to protein markers in the PheWAS a minimum overlapping number of genes was set to 3, whereas this was set to 2 for the genes involved in neurological pQTMs for the purposes of visualisation. The STRING [83] database was queried to build a protein interaction network based on all proteins that had associations in the PheWAS. mQTL and eQTL lookups were performed using the GoDMC [59] and eQTLGen databases

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

526 [84], respectively. UCSC database searches were used to profile the positional information relating to
527 CpGs in the pQTMs involving proteins associated with brain health.

528 Although inter-chromosomal chromatin interactions are unlikely to be stable and persistent, seven
529 proteins with *cis* pQTMs involving CpGs located intra-chromosomally to the proximal protein-
530 coding gene were considered for ChIP-seq and promoter-capture Hi-C mapping to interrogate local
531 chromatin interactions and states that might form the basis for co-regulation of these loci. ChIP-seq
532 data from peripheral blood mononuclear cells (PBMCs) and brain hippocampus were selected from
533 the ENCODE project [86], with accession identifiers available in Supplementary Table 21. Processed
534 promoter-capture Hi-C data for brain hippocampal tissue was integrated from Jung et al, [60] and is
535 available at NCBI Geo with accession GSE86189. Data concerning both promoter-prometer
536 interactions and promoter-other interactions were concatenated and all regions subsequently
537 visualised on the WashU epigenome browser [87].

538 **Ethics declarations**

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

543 **Data availability**

544 Datasets generated in this study are made available in Supplementary Tables. Source data are
545 provided with this paper.

546 Fully-adjusted MWAS summary statistics for protein levels are available through Zenodo [insert
547 details once accepted files are uploaded] and hosted on the MRC-IEU EWAS catalog [insert details
548 once accepted files are uploaded] [25].

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

A YouTube video summarising the findings of the study and detailing how to access files can be viewed at https://www.youtube.com/channel/UCxQrFFTIItF25YKfJTXuumQ.

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, 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. R.E.M is supported by Alzheimer's Society major project grant AS-PG-19b-010. Generation Scotland

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

572 received core support from the Chief Scientist Office of the Scottish Government Health Directorates
573 (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping and DNA methylation
574 profiling of the GS samples was carried out by the Genetics Core Laboratory at the Clinical Research
575 Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK
576 and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression
577 Longitudinally" ([STRADL; Reference 104036/Z/14/Z]). Proteomic analyses in STRADL were
578 supported by Dementias Platform UK (DPUK). DPUK funded this work through core grant support
579 from the Medical Research Council [MR/L023784/2]. C.H. is supported by an MRC University Unit
580 Programme Grant MC_UU_00007/10 (QTL in Health and Disease). L.S. is funded by DPUK
581 through MRC (grant no. MR/L023784/2) and the UK Medical Research Council Award to the
582 University of Oxford (grant no. MC_PC_17215). L.S. also received support from the NIHR
583 Biomedical Research Centre at Oxford Health NHS Foundation Trust.

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

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

592 The authors acknowledge the work of Rebecca Madden, Marco Squillace and Laura Klinkhamer
593 who aided in the quality control of volumetric brain imaging data.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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., S.X., C.G., and J.M.W processed the brain imaging data. D.A.O., G.M.T., and C.R., provided scientific counsel. T.C., and N.R., consulted on chromatin analyses. 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. R.F.H. has received consultant fees from Illumina. 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

References

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

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

3. 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.

4. Cox, S. R., Ritchie, S. J., Fawns-Ritchie, C., Tucker-Drob, E. M. & Deary, I. J. Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence* **76**, 101376 (2019).

5. 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).

6. 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).

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

633 associated with dementia risk. *Nat. Aging* **1**, 473–489 (2021).

634 10. Sebastiani, P. *et al.* A serum protein signature of APOE genotypes in centenarians. *Aging Cell*
635 **18**, e13023 (2019).

636 11. Berson, A., Nativio, R., Berger, S. L. & Bonini, N. M. Epigenetic Regulation in
637 Neurodegenerative Diseases. *Trends Neurosci.* **41**, 587 (2018).

638 12. Al-Mahdawi, S., Virmouni, S. A. & Pook, M. A. The emerging role of 5-
639 hydroxymethylcytosine in neurodegenerative diseases. *Front. Neurosci.* **8**, 397 (2014).

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

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

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

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

648 17. Conole, E. L. S. *et al.* DNA Methylation and Protein Markers of Chronic Inflammation and
649 Their Associations With Brain and Cognitive Aging. *Neurology* **97**, e2340–e2352 (2021).

650 18. Gadd, D. A. *et al.* Epigenetic scores for the circulating proteome as tools for disease
651 prediction. *Elife* **11**, (2022).

652 19. Hillary, R. F. *et al.* Genome and epigenome wide studies of plasma protein biomarkers for
653 Alzheimer's disease implicate TBCA and TREM2 in disease risk. *medRxiv*

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

2021.06.07.21258457 (2021) doi:10.1101/2021.06.07.21258457.

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

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

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

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

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

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

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

27. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet. 2021 5312* **53**, 1712–1721 (2021).

28. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).

29. 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).

30. Shi, L. *et al.* Identification of plasma proteins relating to brain neurodegeneration and vascular pathology in cognitively normal individuals. *Alzheimer's Dement. Diagnosis, Assess. Dis.*

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Monit. **13**, (2021).

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

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

33. 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).

34. 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.

35. 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).

36. Banks, W. A., Reed, M. J., Logsdon, A. F., Rhea, E. M. & Erickson, M. A. Healthy aging and the blood–brain barrier. doi:10.1038/s43587-021-00043-5.

37. Montagne, A. *et al.* Blood-Brain Barrier Breakdown in the Aging Human Hippocampus. *Neuron* **85**, 296 (2015).

38. Beaubien, F., Raja, R., Kennedy, T. E., Fournier, A. E. & Cloutier, J. F. Slitrk1 is localized to excitatory synapses and promotes their development. *Sci. Reports 2016 61* **6**, 1–10 (2016).

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

40. Madetko, N. *et al.* Anti-IgLON5 Disease – The Current State of Knowledge and Further

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Perspectives. *Front. Immunol.* **0**, 777 (2022).

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. Lohoff, F. W. *et al.* Epigenome-wide association study of alcohol consumption in N = 8161 individuals and relevance to alcohol use disorder pathophysiology: identification of the cystine/glutamate transporter SLC7A11 as a top target. *Mol. Psychiatry* (2021) doi:10.1038/S41380-021-01378-6.

45. Wood, G. C. *et al.* A multi-component classifier for nonalcoholic fatty liver disease (NAFLD) based on genomic, proteomic, and phenomic data domains. *Sci. Rep.* **7**, (2017).

46. Zhuang, J., Deane, J. A., Yang, R. B., Li, J. & Ricardo, S. D. SCUBE1, a novel developmental gene involved in renal regeneration and repair. *Nephrol. Dial. Transplant.* **25**, 1421–1428 (2010).

47. Ho, J. C. Y. *et al.* Down-regulation of retinol binding protein 5 is associated with aggressive tumor features in hepatocellular carcinoma. *J. Cancer Res. Clin. Oncol.* **133**, 929–936 (2007).

48. Hop, P. J. *et al.* Genome-wide study of DNA methylation shows alterations in metabolic, inflammatory, and cholesterol pathways in ALS. *Sci. Transl. Med.* **14**, 36 (2022).

49. Park, M. H., Jin, H. K. & Bae, J. sung. Potential therapeutic target for aging and age-related

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

neurodegenerative diseases: the role of acid sphingomyelinase. *Exp. Mol. Med.* **52**, 380–389 (2020).

50. 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).

51. 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).

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

53. 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).

54. 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).

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

56. 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).

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

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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

740 59. Min, J. L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA
741 methylation. *Nat. Genet.* **53**, 1311–1321 (2021).

742 60. Jung, I. *et al.* A Compendium of Promoter-Centered Long-Range Chromatin Interactionsin the
743 Human Genome. *Nat. Genet.* **51**, 1442 (2019).

744 61. Héberlé, É. & Bardet, A. F. Sensitivity of transcription factors to DNA methylation. *Essays*
745 *Biochem.* **63**, 727 (2019).

746 62. Arechederra, M. *et al.* Hypermethylation of gene body CpG islands predicts high dosage of
747 functional oncogenes in liver cancer. *Nat. Commun. 2018 91* **9**, 1–16 (2018).

748 63. Salas, L. A. *et al.* Enhanced cell deconvolution of peripheral blood using DNA methylation
749 for high-resolution immune profiling. *Nat. Commun. 2022 131* **13**, 1–13 (2022).

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

752 65. Yang, C. *et al.* Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins
753 implicated in neurological disorders. *Nat. Neurosci. 2021 249* **24**, 1302–1312 (2021).

754 66. Braun, P. R. *et al.* Genome-wide DNA methylation comparison between live human brain and
755 peripheral tissues within individuals. *Transl. Psychiatry 2019 91* **9**, 1–10 (2019).

756 67. Smith, B. H. *et al.* Cohort profile: Generation scotland: Scottish family health study (GS:
757 SFHS). The study, its participants and their potential for genetic research on health and illness.
758 *Int. J. Epidemiol.* **42**, 689–700 (2013).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

68. 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).

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

70. 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).

71. 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).

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

73. 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).

74. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).

75. Reinius, L. E. *et al.* Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS One* **7**, e41361 (2012).

76. 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).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

77. 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).

78. prcomp function - RDocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp.

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

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

81. (2017), R. C. T. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

82. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun. 2017 81* **8**, 1–11 (2017).

83. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

84. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet. 2021 539* **53**, 1300–1310 (2021).

85. Kent, W. *et al.* UCSC Genome Browser: The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

86. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

(2016).

87. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU Epigenome Browser update 2019. *Nucleic Acids Res.* **47**, W158–W165 (2019).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figure legends

Figure 1. Integrated methylome and phenome study of the plasma proteome in relation to brain health. Study design and key results are presented in this flow diagram. 1,065 individuals from Generation Scotland had the levels of 4,058 plasma proteins (corresponding to 4,235 SOMAmers) measured. A methylome-wide association study (MWAS) of the 4,058 plasma protein levels was conducted in 774 individuals that represented a nested subset of the full sample with DNAm measurements available. This identified 2,928 CpG-protein (pQTM) associations. A phenome-wide protein association study (Protein PheWAS) identified 191 protein levels that were associated with a minimum of one brain health trait (N>=909). Integration of the protein MWAS and PheWAS results identified 35 pQTMs that involved the levels of 17 protein markers of brain health and 31 unique CpGs located in 20 genes.

Figure 2. Methylome-wide studies of 4,058 plasma proteins. a Summary of MWAS results for 4,058 protein levels in Generation Scotland (N=774). MWAS pQTMs that had $P < 4.5x10^{-10}$ in the basic, white blood cell proportion (WBC)-adjusted and fully-adjusted models. *Cis* associations (purple) and *trans* associations (green) are summarised for each model. Covariates used to adjust DNAm are described for each model. Protein levels were adjusted for age, sex, 20 genetic principal components (PCs) and technical variables and normalised prior to running MWAS. **b** Flow diagram showing the distinction between the highly pleiotropic PAPPA and PRG3 protein pQTMs and the 825 pQTMs that involved the levels of a further 189 proteins. **c** Genomic locations for 825 of the 2,928 fully-adjusted pQTMs, excluding highly pleotropic associations for PAPPA and PRG3 protein levels. Chromosomal location of CpG sites (x-axis) and protein genes (y-axis) are presented. The 434 *cis* pQTMs (purple) lay on the same chromosome and ≤ 10Mb from the transcriptional start site (TSS) of the protein gene, whereas the 391 *trans* pQTMs (green) lay > 10Mb from the TSS of the

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

protein gene or on a different chromosome. A list of the full association counts for each protein and CpG site can be found in Supplementary Tables 8-9.

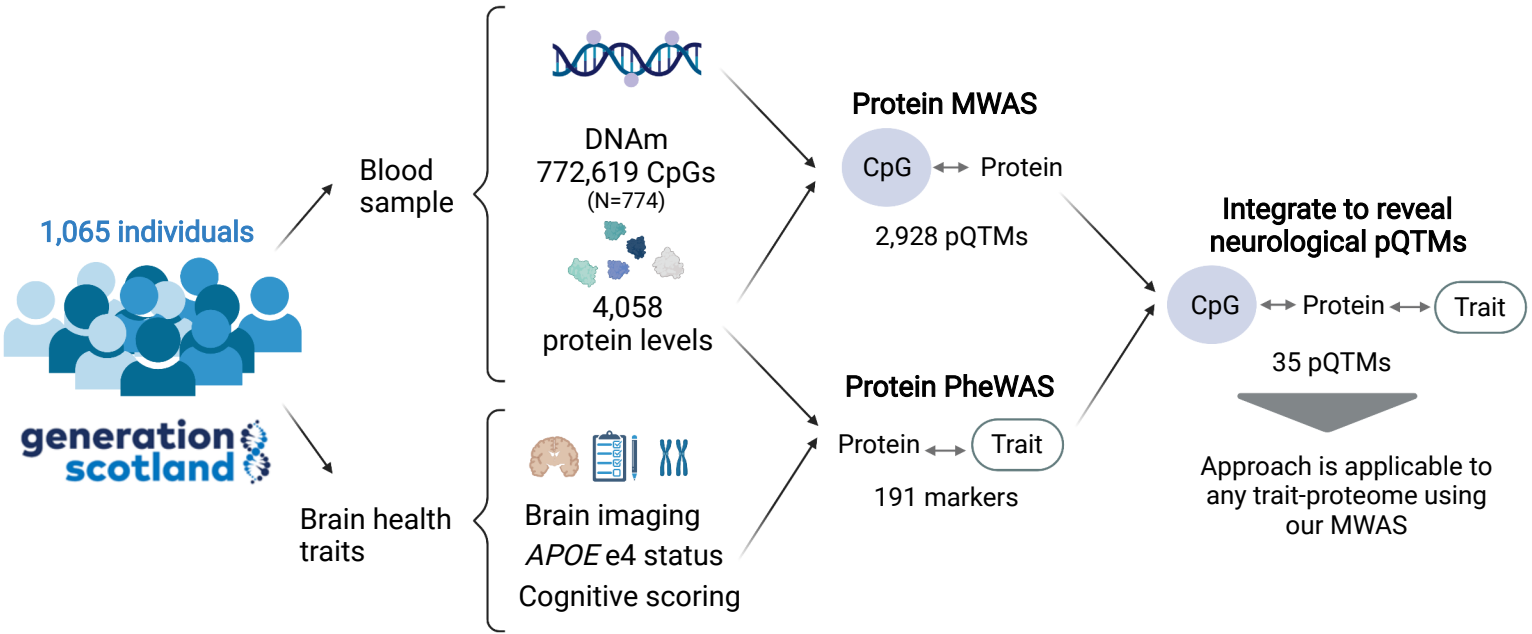
Figure 3. Pleiotropic associations in the fully-adjusted MWAS. a pQTMs that had $P < 4.5x10^{-10}$ in the fully-adjusted MWAS are plotted as individual points with chromosomal locations of the 191 protein genes (upper) and the 1,837 CpGs (lower) on the x-axis. 19 proteins with ≥ 10 associations with CpGs are highlighted in turquoise and labelled on the upper plot. Nine CpGs with ≥ 6 associations with protein levels are highlighted in turquoise on the lower plot. **b** Summary of genes with highly pleiotropic CpG signals in the fully-adjusted MWAS. The fully-adjusted MWAS pQTMs can be accessed in Supplementary Table 6. A list of the full association counts for each protein and CpG site can be found in Supplementary Tables 8-9.

Figure 4. PheWAS of 4,058 plasma proteins and brain health. a Number of protein marker associations with $P < 3.5x10^{-4}$ for each of the 15 traits related to brain health in the phenome-wide protein association studies (protein PheWAS). These studies included a maximum sample of 1,065 individuals with protein measurements from Generation Scotland and tested for associations between 15 phenotypes and the levels of 4,058 plasma proteins. Cognitive score (turquoise), brain imaging (light blue) and *APOE* e4 status (dark blue) associations are summarised. **b** Heatmap of standardised beta coefficients for 77 of the 405 protein PheWAS associations ($P < 3.5x10^{-4}$) indicated by an asterisk. They include three proteins that had associations with both *APOE* e4 status and one or more cognitive scores, in addition to 22 proteins that had associations with both a brain imaging measure and a cognitive score. Negative and positive beta coefficients are shown in blue and red, respectively. A heatmap describing the full 405 associations for *APOE* e4 status, cognitive scores and brain imaging measures is available in Supplementary Fig. 6. Full summary statistics for the 405 associations are presented in Supplementary Table 17 and the subset of 77 associations shown in part b can be accessed in Supplementary Table 18.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figure 5. pQTMs involving protein markers of brain health. Circular plot showing 15 *trans* pQTM associations between DNAm at 11 CpG sites and the levels of nine proteins that were associated with one of more of the neurological phenotypes ($P < 3.5x10^{-4}$). Chromosomal positions are given on the outermost circle. Full details of the 35 pQTMs, including 20 *cis* associations are reported in Supplementary Table 20.

Figure 6. Integration of candidate marker associations. a Three *trans* associations with the CpG site cg06690548 in the *SLC7A11* gene, which encodes a synaptic protein that is linked to environmental mediation in Parkinson's disease and is involved in glutamate transmission and oxidative stress. **b** Five *trans* associations between CpGs in the *ITIH3/ITIH1/NEK4* region on chromosome 3 and the levels of RBL2, which was associated with reductions in Global Grey Matter Volume. **c** Two *trans* associations between DNAm at cg11294350 in the *CHPT1* gene and two proteins with lysosomal-associated function (SMPD1 and HEXB) that were associated with higher Relative Brain Age and lower General Fractional Anisotropy. Associations with a positive beta coefficient are denoted as red connecting lines, whereas associations with a negative beta coefficient are denoted as blue connecting lines. The full 15 *trans* associations and 20 *cis* associations can be found in Supplementary Table 20.


1,065 individuals
generation scotland
Blood sample
Brain health traits
DNAm
772,619 CpGs
(N=774)
4,058
protein levels
Brain imaging
APOE e4 status
Cognitive scoring
Protein MWAS
CpG
Protein
2,928 pQTMs
Protein PheWAS
Protein
Trait
191 markers
Integrate to reveal neurological pQTMs
CpG
Protein
Trait
35 pQTMs
Approach is applicable to any trait-proteome using our MWAS

a

Basic MWAS

DNAm adjusted for: Age, Sex, Depression, DNAm Set/Batch

238,245 pQTMs

2,107 *cis*

236,138 *trans*

White blood cell (WBC)-adjusted MWAS

+ estimated WBC proportions

3,213 pQTMs

453 *cis*

2,760 *trans*

Fully-adjusted MWAS

+ BMI and a DNAm-based smoking score (EpiSmokEr)

2,928 pQTMs

451 *cis*

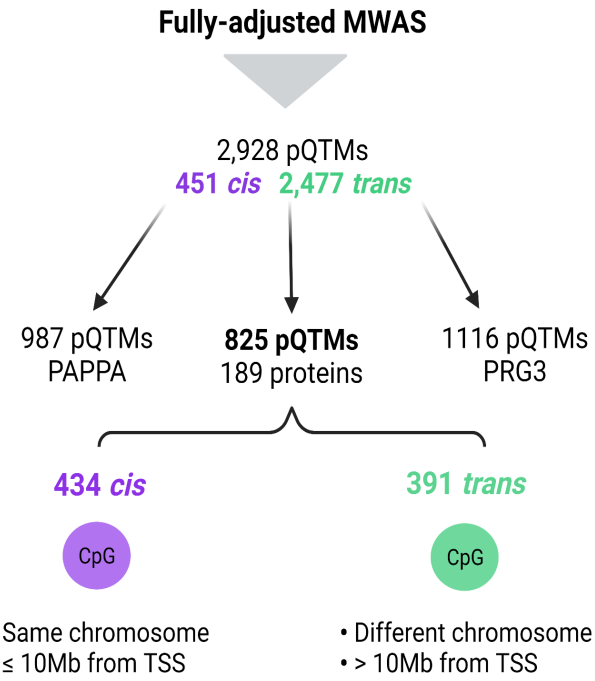
2,477 *trans*

Protein levels adjusted for: Age, Sex, pQTLs, technical covariates, 20 genetic PCs

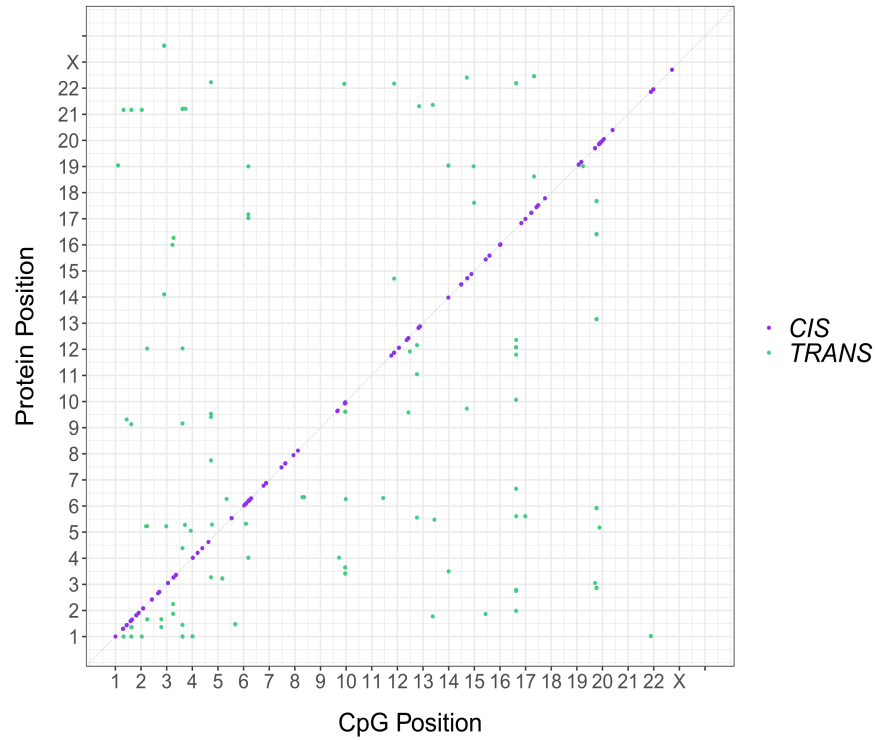
Association threshold: $P < 4.5 \times 10^{-10}$

N=774 individuals, N=4,058 protein levels (4,235 SOMAmer measurements)

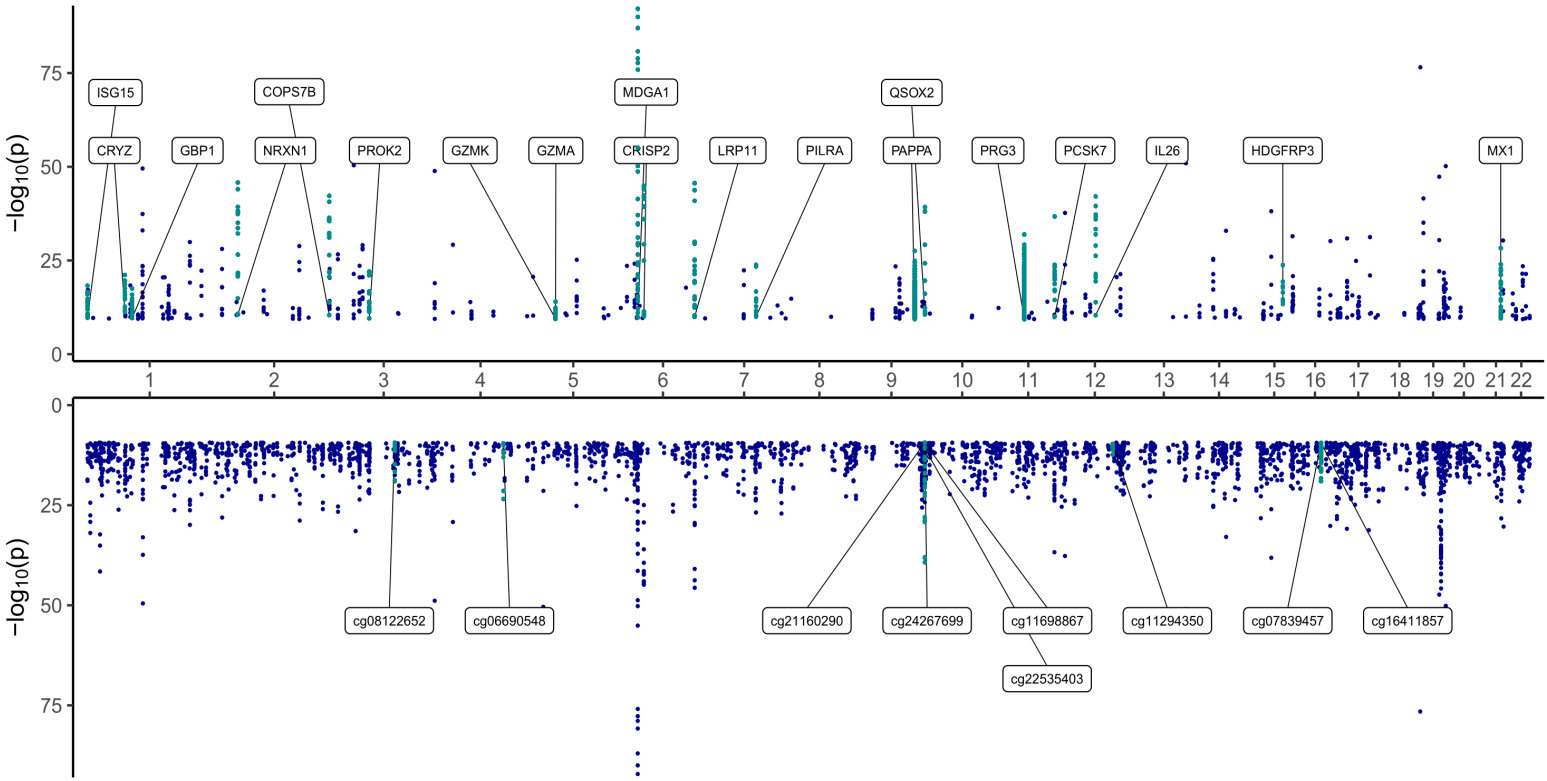
b



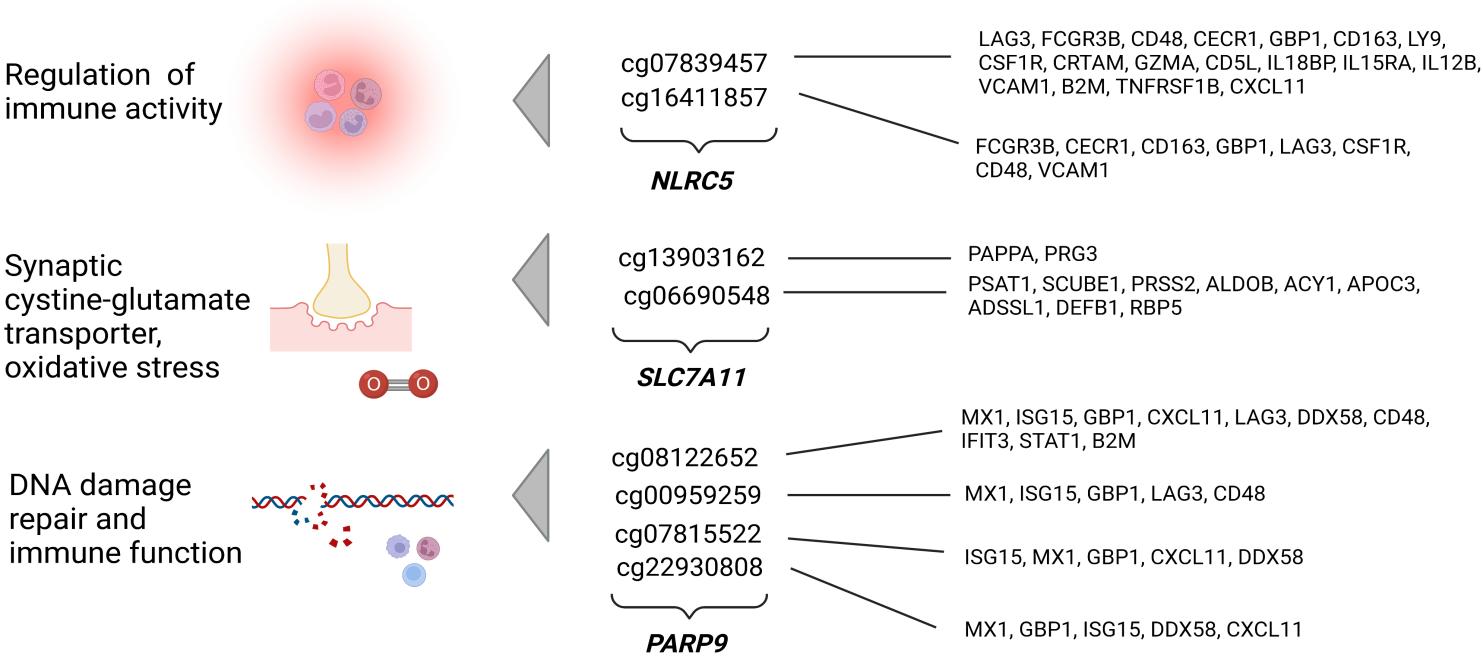
c

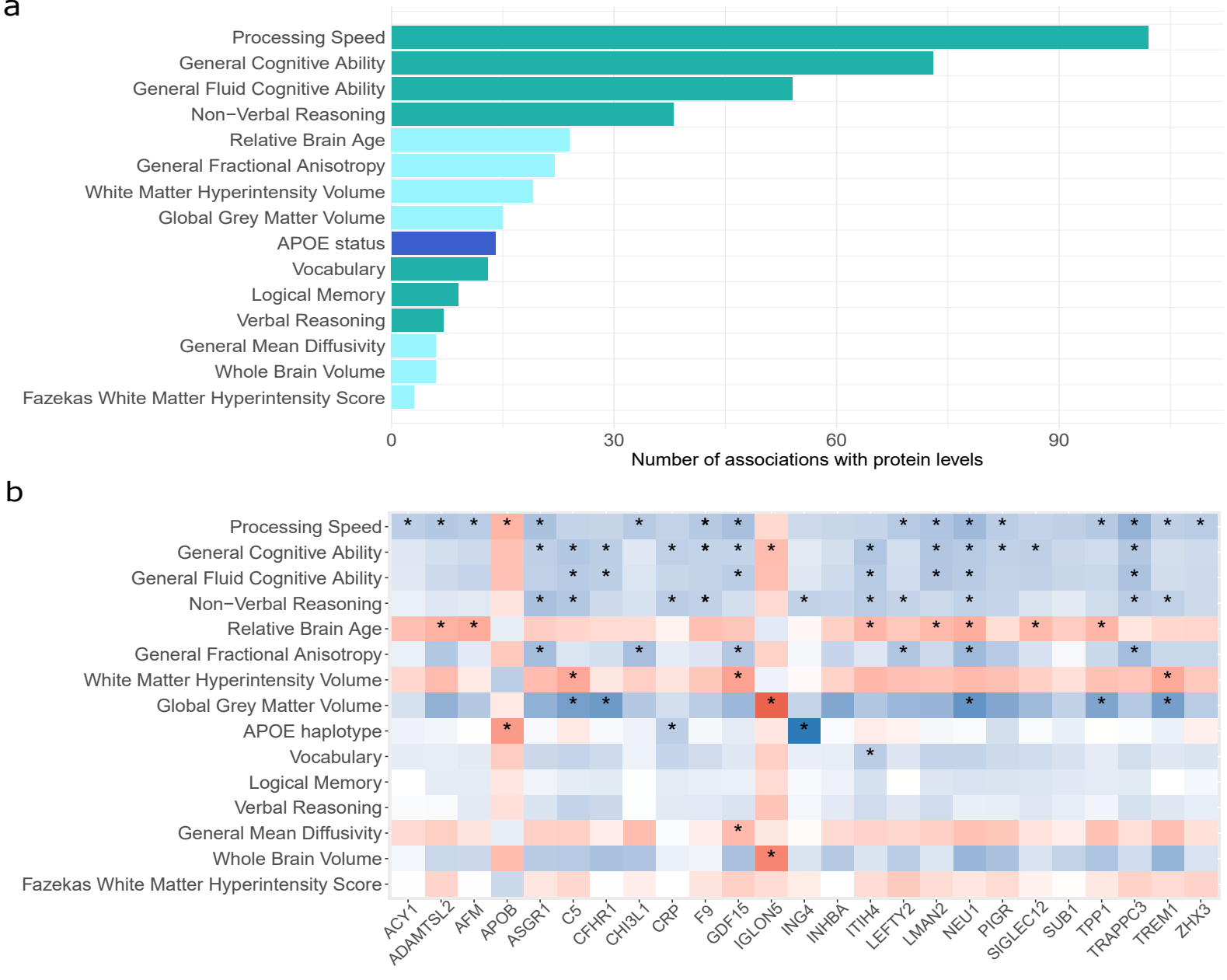


a

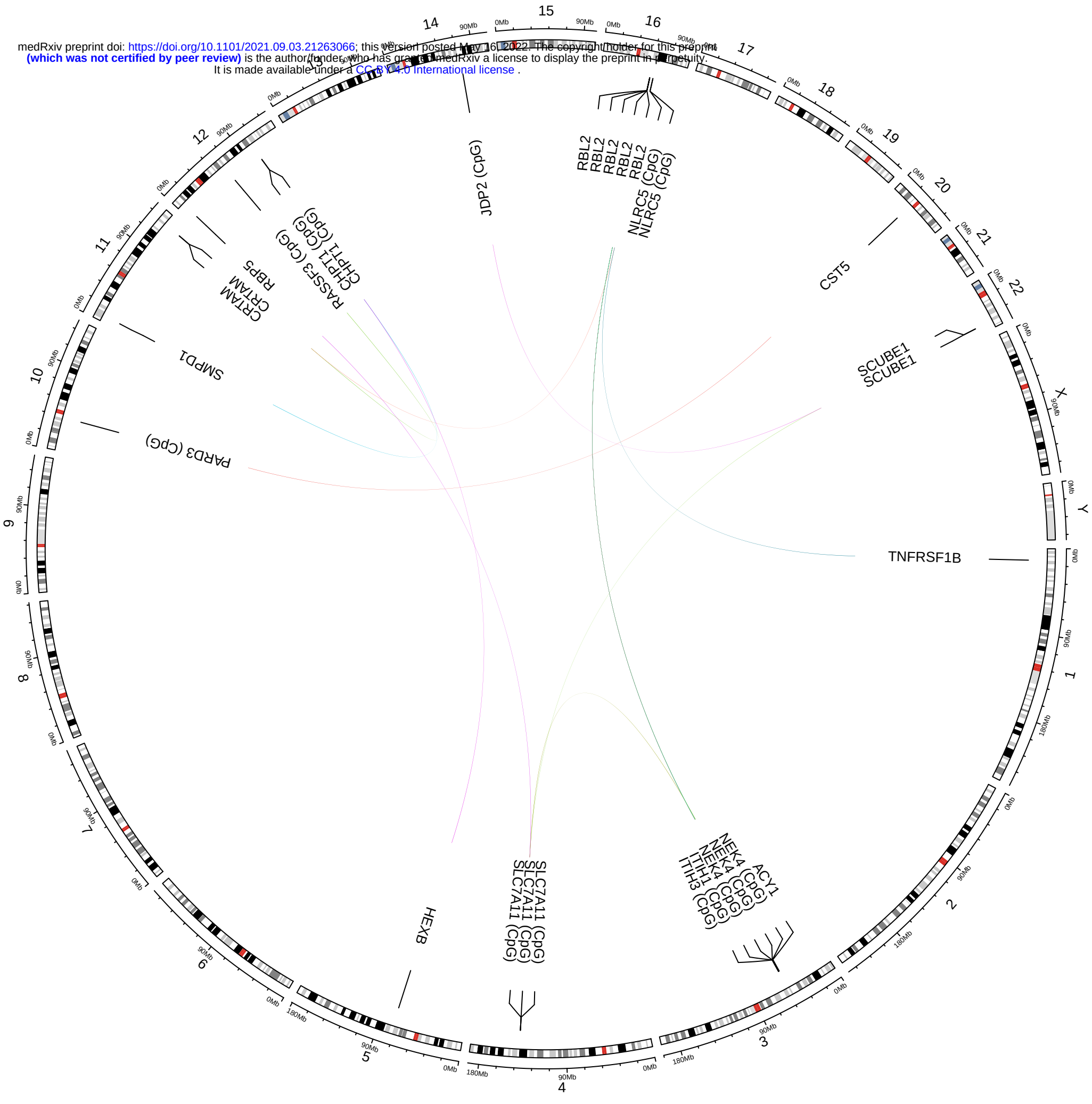

ISG15
CRYZ
GBP1
COPS7B
NRXN1
PROK2
GZMK
GZMA
MDGA1
CRISP2
LRP11
PILRA
QSOX2
PAPPA
PRG3
PCSK7
IL26
HDGFRP3
MX1
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
cg08122652
cg06690548
cg21160290
cg24267699
cg11698867
cg22535403
cg11294350
cg07839457
cg16411857

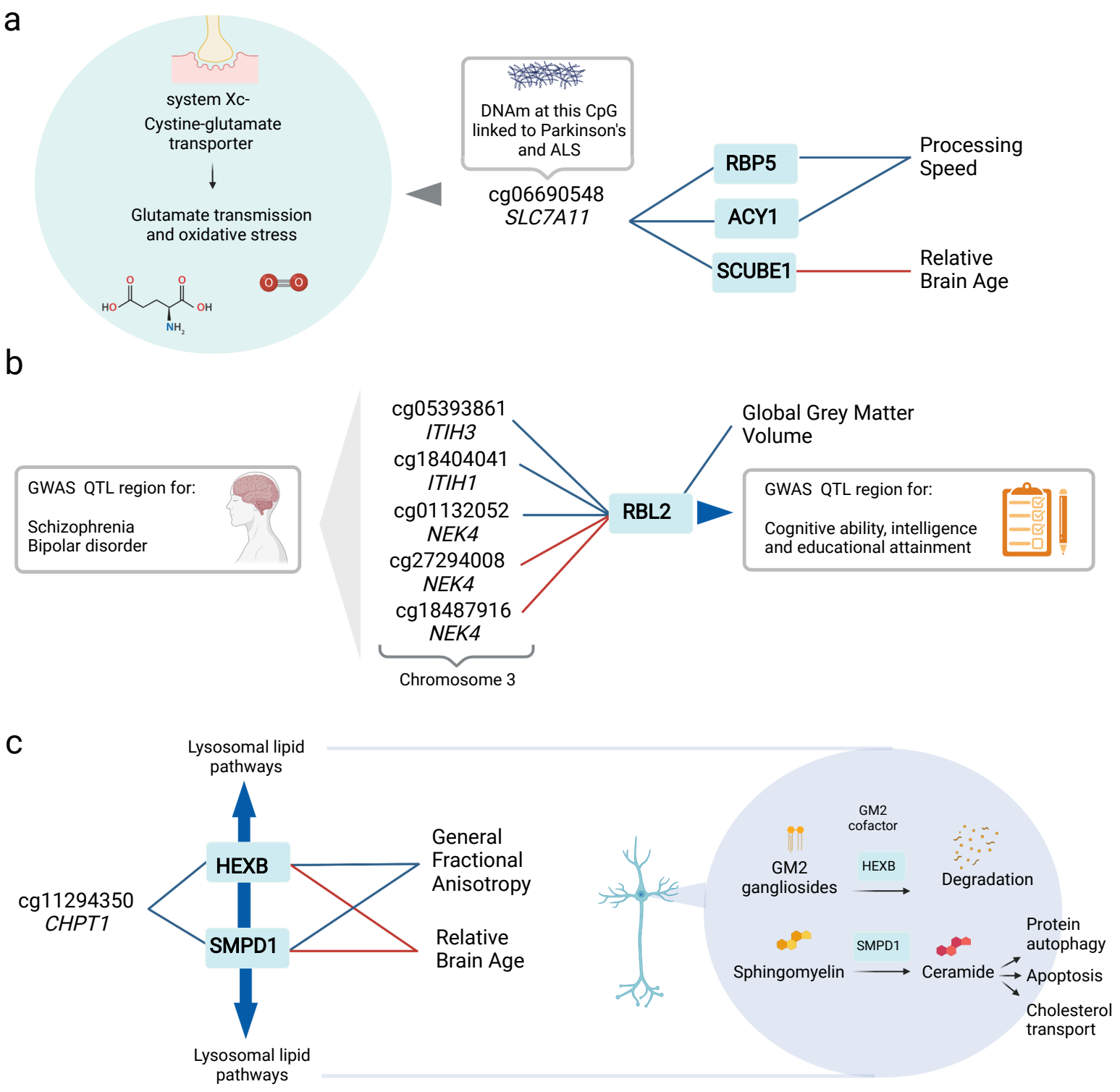
b


Regulation of immune activity
cg07839457
cg16411857
NLRC5
LAG3, FCGR3B, CD48, CECR1, GBP1, CD163, LY9, CSF1R, CRTAM, GZMA, CD5L, IL18BP, IL15RA, IL12B, VCAM1, B2M, TNFRSF1B, CXCL11
FCGR3B, CECR1, CD163, GBP1, LAG3, CSF1R, CD48, VCAM1
Synaptic cystine-glutamate transporter, oxidative stress
cg13903162
cg06690548
SLC7A11
PAPPA, PRG3
PSAT1, SCUBE1, PRSS2, ALDOB, ACY1, APOC3, ADSSL1, DEFB1, RBP5
DNA damage repair and immune function
cg08122652
cg00959259
cg07815522
cg22930808
PARP9
MX1, ISG15, GBP1, CXCL11, LAG3, DDX58, CD48, IFIT3, STAT1, B2M
MX1, ISG15, GBP1, LAG3, CD48
ISG15, MX1, GBP1, CXCL11, DDX58
MX1, GBP1, ISG15, DDX58, CXCL11


a
Processing Speed
General Cognitive Ability
General Fluid Cognitive Ability
Non−Verbal Reasoning
Relative Brain Age
General Fractional Anisotropy
White Matter Hyperintensity Volume
Global Grey Matter Volume
APOE status
Vocabulary
Logical Memory
Verbal Reasoning
General Mean Diffusivity
Whole Brain Volume
Fazekas White Matter Hyperintensity Score
0
30
60
90
Number of associations with protein levels
b
Processing Speed
General Cognitive Ability
General Fluid Cognitive Ability
Non−Verbal Reasoning
Relative Brain Age
General Fractional Anisotropy
White Matter Hyperintensity Volume
Global Grey Matter Volume
APOE haplotype
Vocabulary
Logical Memory
Verbal Reasoning
General Mean Diffusivity
Whole Brain Volume
Fazekas White Matter Hyperintensity Score
ACY1
ADAMTSL2
AFM
APOB
ASGR1
C5
CFHR1
CHI3L1
CRP
F9
GDF15
IGLON5
ING4
INHBA
ITIH4
LEFTY2
LMAN2
NEU1
PIGR
SIGLEC12
SUB1
TPP1
TRAPPC3
TREM1
ZHX3

medRxiv preprint doi: https://doi.org/10.1101/2021.09.03.21263066; this version posted May 16, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .




a
system Xc-
Cystine-glutamate transporter
Glutamate transmission and oxidative stress
DNAm at this CpG linked to Parkinson's and ALS
cg06690548
SLC7A11
RBP5
ACY1
SCUBE1
Processing Speed
Relative Brain Age
b
GWAS QTL region for:
Schizophrenia
Bipolar disorder
cg05393861
ITIH3
cg18404041
ITIH1
cg01132052
NEK4
cg27294008
NEK4
cg18487916
NEK4
Chromosome 3
RBL2
Global Grey Matter Volume
GWAS QTL region for:
Cognitive ability, intelligence and educational attainment
c
Lysosomal lipid pathways
cg11294350
CHPT1
HEXB
SMPD1
General Fractional Anisotropy
Relative Brain Age
Lysosomal lipid pathways
GM2 cofactor
GM2 gangliosides
HEXB
Degradation
Sphingomyelin
SMPD1
Ceramide
Protein autophagy
Apoptosis
Cholesterol transport