A multi-omics approach to understand the role of plasma proteins in cognitive ageing and dementia

Robert Francis Hillary

Doctor of Philosophy

The University of Edinburgh

2021

Abstract

The global burden of age-related cognitive decline and dementia will continue to rise in tandem with our ageing population. This necessitates the discovery of novel biomarkers and candidate drug targets to combat cognitive dysfunction. Blood proteins are important drug targets, and blood samples can be acquired routinely in clinical settings and epidemiological studies. Whereas hundreds of blood proteins are associated with cognitive ability and dementia, we do not understand whether these associations represent correlation or causation. Genome-wide association studies (GWAS) are required to define variants that are associated with blood protein levels. These variants can proxy for candidate disease-markers and assess their causal associations with health outcomes in analysis methods such as Mendelian randomisation. DNA methylation is an epigenetic mechanism that regulates gene expression and is influenced by genetic and environmental factors. Studying the relationship between DNA methylation and protein levels could reveal whether genetic variation or environmental factors likely mediate associations between blood proteins and disease states. The first aim of this thesis is to conduct GWAS and epigenome-wide association studies (EWAS, using DNA methylation) on plasma levels of 422 unique proteins. Using these data, I apply causal inference approaches to determine whether blood proteins are causally associated with Alzheimer's disease risk.

Several strategies have been proposed to estimate biological age by leveraging inter-individual variation in DNA methylation profiles. Epigenetic measures of ageing correlate strongly with chronological age. Recently, a novel epigenetic measure of ageing termed 'DNAm GrimAge' was developed to predict one's risk of mortality. DNAm GrimAge is a composite biomarker that incorporates methylation-based predictors of seven blood protein levels and smoking. The relationship between this biomarker of ageing and cognitive decline or dementia is not known. Therefore, the second aim of this thesis is to examine whether DNAm GrimAge associates with measures of brain health and Alzheimer's disease. To conduct these aims, I utilise data from two cohort studies: the Lothian Birth Cohort 1936 (n ≤ 906, LBC1936) and Generation Scotland (n ≤ 9,537, GS).

In Chapters 1-3, I provide an overview of cognitive ageing and dementia. I describe GWAS and EWAS on blood protein levels and the development of DNAm GrimAge. In Chapter 4, I detail the population cohorts and main methodologies that are used in this thesis.

In Chapter 5, I conduct GWAS and EWAS on plasma levels of 92 neurology-related proteins (n ≤ 750, LBC1936). I identified 41 independent genetic and 26 epigenetic loci that associate with 33 and 9 proteins, respectively. I showed that an immune-related protein, poliovirus receptor (PVR), is causally associated with Alzheimer's disease risk. In Chapter 6, I use a novel Bayesian framework termed BayesR+ to perform an integrated GWAS/EWAS on plasma levels of 70 inflammation-associated proteins (n = 876, LBC1936). Many GWAS and EWAS use linear models, which examine every measured genetic or epigenetic site in isolation. BayesR+ accounts for intercorrelations among genetic and epigenetic sites and the reciprocal influences of these data types. I estimated the contribution of genetic and epigenetic variation towards inter-individual differences in inflammatory protein levels, considered alone and together. There was no evidence for causal associations between blood inflammatory proteins and the risk of Alzheimer's disease. In Chapter 7, I perform a systematic literature review to identify known blood protein correlates of Alzheimer's disease. I then use BayesR+ to conduct an integrated GWAS and EWAS on plasma levels of 282 Alzheimer's disease-associated proteins (n ≤ 1,064, GS). I observed strong evidence for causal associations between two proteins, TBCA and TREM2, and Alzheimer's disease risk.

In Chapter 8, I examine associations between DNAm GrimAge and measures of brain health (n ≤ 709, LBC1936). A higher-than-expected DNAm GrimAge associated with poorer performance on cognitive tasks and neurostructural correlates of dementia at age 73. I observed weak evidence to suggest that DNAm GrimAge assessed at age 70 predicts cognitive decline up to age 79. In Chapter 9, I assess whether DNAm GrimAge and other measures of epigenetic ageing predict the prevalence and incidence of common disease states, including Alzheimer's disease (n ≤ 9,537, GS). Epigenetic ageing measures did not predict the prevalence or incidence of Alzheimer's disease. In Chapter 10, I discuss the major findings from this thesis in light of their limitations.

The work presented in this thesis helps to detail the molecular regulation of 422 plasma protein levels and their causal associations with Alzheimer's disease. This work also highlights the performance of DNAm GrimAge in predicting indices of cognitive performance and common disease states. By incorporating genetic, epigenetic and protein data in two large-scale epidemiological studies, my findings inform our understanding of relationships between blood proteins and cognitive ageing and dementia.

Lay Summary

As we grow older, we may experience a decline in our thinking, or cognitive, skills. This is a major global health problem and might precede dementia. We do not have the tools to reliably predict who will develop cognitive decline and dementia. Blood samples can be used as a relatively pain-free way to measure a person's risk of disease. Our blood contains many proteins that keep us healthy and these proteins are the targets of many drugs. The levels of some proteins in our blood are linked to cognitive decline and dementia. However, we do not know why these changes occur. If we understood whether changes in blood protein levels are a cause or consequence of poorer brain health, we might identify new biomarkers and drug targets for cognitive decline and dementia.

In this thesis, I use blood data from volunteers in two large population studies (the Lothian Birth Cohort 1936 and Generation Scotland). My first aim is to study the biological factors that control blood levels of over 400 proteins. Genes provide the instructions to make our proteins. Therefore, I examine how our genetic differences affect the levels of proteins in our blood. Genes must be switched on to make these proteins. They are switched on and off by an epigenetic mechanism called DNA methylation. Methylation is controlled by both genetic and lifestyle factors such as diet and stress. Studying DNA methylation can therefore tell us more than genetics alone about why we differ in our protein levels. In addition to genetics, I examine how our differences in DNA methylation relate to differences in blood protein levels. Using these data, I apply several statistical techniques to determine whether any of these blood proteins likely cause Alzheimer's disease, which is the most common cause of dementia.

DNA methylation patterns can be used to predict someone's biological age. If someone's biological age is higher than their actual age, it might mean they are at risk of health problems. A major biological age measure is called DNAm GrimAge. It is based on a blood test that examines DNA methylation, seven blood proteins and whether someone is a smoker. My second aim is to examine whether DNAm GrimAge can predict measures of people's brain health and dementia.

In my first three studies, I investigate how genetic and epigenetic variation might affect blood levels of 422 proteins. In the first of these studies, I examine 92 blood proteins that relate to brain health. I found that higher levels of a protein called poliovirus receptor might increase the risk of getting Alzheimer’s disease. In my second study, I analyse 70 proteins that are involved in inflammation, which is associated with dementia and other brain diseases. My findings did not suggest that changes in the levels of inflammatory proteins affect our risk of Alzheimer’s disease. In my third study, I examine 282 proteins that are associated with Alzheimer’s disease. I found that higher blood levels of two proteins, called TREM2 and TBCA, are associated with an increased risk of Alzheimer’s disease.

In my fourth study, I test whether DNAm GrimAge predicts poorer thinking skills and signs of ‘wear-and-tear’ in brain scans at age 73 years. I found that DNAm GrimAge associates with poorer brain health at age 73, and might predict a decline in thinking skills from age 70 to 79 years. In my fifth study, I investigate whether DNAm GrimAge and five other measures of biological ageing predict ten common diseases, including Alzheimer’s disease. I found that DNAm GrimAge was more useful than other biological ageing measures in predicting several diseases including lung and heart disease. However, no measures of biological ageing predicted Alzheimer’s disease. In my last chapter, I discuss the main limitations of my studies and how future studies can use my findings to progress our understanding of dementia.

The work in this thesis tells us about the biology that underlies blood levels of over 400 proteins and their relationships with Alzheimer’s disease. My findings also show how well current measures of biological ageing predict the future risk of poor brain health and dementia. Together, my findings suggest that studying the relationship between genetics, DNA methylation and proteins in our blood might improve our understanding of cognitive decline and dementia.

Declaration of Originality

I declare that this thesis is my own composition and that it has not been submitted for any other degree or professional qualification at this university or any other institution. Parts of the work comprising this thesis have been previously published. The included publications are my own work, except where indicated otherwise.

The work presented in Chapter 5 has been published in *Nature Communications*. Author contributions are as follows: **R.F.H** and R.E.M conceived and designed the research; **R.F.H**, D.L.Mc.C, D.C.L and R.E.M conducted the statistical analyses; **R.F.H** and R.E.M drafted the article; all authors reviewed the manuscript.

The work presented in Chapter 6 has been published in *Genome Medicine*. Author contributions are as follows: **R.F.H**, M.R.R and R.E.M conceived and designed the research; **R.F.H** conducted the statistical analyses; **R.F.H**, M.R.R and R.E.M drafted the article; D.T.B, A.K, D.L.Mc.C, Q.Z, D.C.L and S.E.H contributed to the data preparation; S.E.H, N.R.W, A.F.M, P.M.V and I.J.D were responsible for the data collection; all authors reviewed the manuscript.

The work presented in Chapter 7 has been submitted for publication. Author contributions are as follows: **R.F.H** and R.E.M conceived and designed the research; **R.F.H** conducted the statistical analyses; **R.F.H** and R.E.M drafted the article; all authors reviewed the manuscript.

The work presented in Chapter 8 has been published in *Molecular Psychiatry*. Author contributions are as follows: **R.F.H**, A.J.S and R.E.M conceived and designed the research; **R.F.H**, A.J.S and S.R.C conducted the statistical analyses; **R.F.H** and R.E.M drafted the article; all authors reviewed the manuscript.

The work presented in Chapter 9 has been published in *Clinical Epigenetics*. Author contributions are as follows: **R.F.H**, A.J.S and R.E.M conceived and designed the research; **R.F.H**, A.J.S, D.L.Mc.C and R.E.M conducted the statistical analyses; A.C, R.M.W, C.H, A.M.M, D.J.P, I.J.D and K.L.E contributed to data collection and preparation; **R.F.H** and R.E.M drafted the article; all authors reviewed the manuscript.

Signed: R Hillary Date: 14/07/2021

Acknowledgements

Firstly, I would like to thank my supervisor Dr Riccardo Marioni for his mentorship, guidance and support throughout my PhD studies. I will forever be grateful to Riccardo for allowing me to join his lab as a PhD rotation student in my first year with little bioinformatics experience and for teaching and encouraging me during my time in his lab to become a confident and competent researcher. I would also like to thank my co-supervisors Dr Kathy Evans, Prof Craig Ritchie and Prof Ian Deary, and thesis chair Prof Caroline Hayward for their advice and guidance during my studies.

Thank you to the participants of the Lothian Birth Cohort and Generation Scotland studies for making the work in this thesis possible. I also acknowledge Wellcome for funding my PhD studies. Thank you to Dr Sarah Harris for her mentorship and for her expertise with proteomic data. Thank you to Prof David Porteous, Dr Gail Davies and Dr Simon Cox, to name a few, for their help and assistance during a number of my projects. Thank you also to Prof Matthew Robinson and members of the Robinson group - Daniel, Thanasis, Marion and Sven - for hosting me during a research visit to Lausanne in the summer of 2019.

A huge thank you to Dr Daniel McCartney and Dr Anna Stevenson for their mentorship, patience and for making every day in the office an utter joy. Thank you to all lab members, past and present, Anne Seeboth, Danni Gadd and Yipeng Cheng for their friendship and expertise.

Thank you to my friends and fellow cohort members in the Translational Neuroscience PhD programme - Caoimhe, Bex, Fenia, Anders and Rana - for their unwavering support, friendship and help throughout these past four years. Thank you to the directors of my PhD programme and our co-ordinator Dr Jane Haley for an enjoyable and enriching learning experience. To India, thank you for encouraging me to move to Edinburgh to pursue my PhD and joining me in that journey. Your enduring support throughout the PhD has meant a great deal to me. Thank you to Jason, Keira, Barry, Teddy, Vincent and Cooper for providing comfort and breaks from the PhD in my visits back home to Ireland. I extend my gratitude to my grandparents, Kathleen, Fachnan, Ann and my late grandfather Frank Sr. for their unconditional support. To Paul, thank you for all of your support, enabling my tea addiction and for always cracking a great joke when I need it the most. Finally, thank you to my parents, Valerie and Frankie, for a lifetime of encouragement, support and laughs. It is my greatest honour to be called your son.

Table of Contents

List of Abbreviations

AD	Alzheimer’s disease
AHRR	Aryl-hydrocarbon receptor repressor
APOE	Apolipoprotein E
ASM	Acid sphingomyelinase
CHI3L1	Chitinase 3 like 1
CI	Confidence interval
CKD	Chronic kidney disease
CNS	Central nervous system
COJO	Conditional and joint analysis
COPD	Chronic obstructive pulmonary disease
CpG	Cytosine-Guanine dinucleotide
CrI	Credible interval
CRP	C-reactive protein
CSF	Cerebrospinal fluid
CVD	Cardiovascular disease
DALY	Disability-adjusted life years
DARC	Duffy antigen/chemokine receptor
DNA	Deoxyribonucleic acid
DNAm	DNA methylation
DNMT	DNA methyltransferase
DSM	Diagnostic and statistical manual of mental disorders
EEAA	Extrinsic epigenetic age acceleration
eGFR	Estimated glomerular filtration rate
eQTL	Expression quantitative trait locus
EWAS	Epigenome-wide association studies

F2RL3	F2R-like thrombin or trypsin receptor 3
FA	Fractional anisotropy
FDR	False discovery rate
FER	Forced expiratory rate
FEV1	Forced expiratory volume in one second
FTD	Frontotemporal dementia
FUMA	Functional annotation and mapping
FVC	Forced vital capacity
GO	Gene ontology
GWAS	Genome-wide association studies
HEIDI	Heterogeneity in dependent instruments
hQTL	Histone quantitative trait locus
HR	Hazard ratio
IBD	Inflammatory bowel disease
ICV	Intracranial volume
IEAA	Intrinsic epigenetic age acceleration
IHD	Ischemic heart disease
IL	Interleukin
InSIDE	Instrument strength independent of direct effect
IQ	Intelligence quotient
ITIH1/4	Inter-alpha-trypsin inhibitor heavy chain H1/H4
IV	Instrumental variable
JLIM	Joint likelihood mapping
KEGG	Kyoto encyclopaedia of genes and genomes
LASSO	Least absolute shrinkage and selection operator
LBC1936	Lothian Birth Cohort 1936

LBD	Lewy body dementia
LD	Linkage disequilibrium
MAPKAPK5	MAP kinase-activated protein kinase 5
MD	Mean diffusivity
MDS	Multidimensional scaling
MHC	Major histocompatibility complex
MMP	Matrix metalloproteinase
MMSE	Mini-mental state examination
MOMENT	multi-component MLM-based omic association excluding the target
mQTL	Methylation quantitative trait locus
MR	Mendelian randomisation
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
MS	Mass spectrometry
MS4A4A/A6A	Membrane spanning 4-domains A4A/A6A
NLRC5	NOD-like receptor family CARD domain containing 5
ns	Non-significant
OLS	Ordinary least squares
OR	Odds ratio
OSCA	OmicS-data-based Complex trait Analysis
PCA	Principal component analysis
PCR	Polymerase chain reaction
PEA	Proximity extension assay
PEF	Peak expiratory flow
PIP	Posterior inclusion probability
PLA	Proximity ligation assay

PoA	Pace of ageing
PP	Posterior probability
pQTL	Protein quantitative trait locus
PVR	Poliovirus receptor
RFU	Relative fluorescence units
SCID	Structured Clinical Interview for DSM-IV
SD	Standard deviation
SE	Standard error
SELEX	Systematic evolution of ligands by exponential enrichment
SIMD	Scottish index of multiple deprivation
SMR	Scottish morbidity records
SNP	Single-nucleotide polymorphism
SOMAmers	Slow Off-rate Modified Aptamers
sQTL	Splicing quantitative trait locus
STRADL	Stratifying Resilience and Depression Longitudinally
TBCA	Tubulin-specific chaperone A
TREM2	Triggering receptor expressed on myeloid cells 2
TSS	Transcription start site
VaD	Vascular dementia
VEGFA	Vascular endothelial growth factor A
VGF	VGF nerve growth factor inducible

List of Figures

List of Tables

Introduction to Thesis

In the absence of effective therapies and biomarkers, the global burden of cognitive decline and dementia will continue to rise with our ageing population. Blood proteins are important drug targets and attractive disease biomarkers as they can be assessed through non-invasive and repeated sampling. This is of particular importance in neurological disease states in which routine access to neural tissue is infeasible. The development of effective drug therapies and biomarkers is predicated on our understanding of a given protein's role in disease processes. Understanding the molecular factors that underpin circulating protein levels can help to disentangle whether perturbations in protein levels are a cause or consequence of disease. Therefore, the first aim of this thesis is to utilise data from two large cohort studies and investigate the relationship between genome-wide genetic and epigenetic factors (using blood DNA methylation) and 422 plasma protein levels. Using these data, I assess causal relationships between plasma proteins and Alzheimer's disease (AD). DNA methylation profiles can capture long-term changes in lifestyle behaviours and act as surrogate markers for circulating protein levels. Therefore, the second aim of this thesis is to examine whether an existing DNA methylation-based biomarker of seven plasma protein levels and cigarette smoking termed 'DNAm GrimAge' associates with cognitive health and AD. This composite biomarker – often referred to as a measure of biological ageing – was trained to predict the risk of all-cause mortality.

In Chapter 1, I introduce the distinction between normal cognitive ageing and dementia. Further, I discuss the motivation for identifying blood protein markers of disease states. In Chapter 2, I introduce statistical methods for studying relationships between genetics, epigenetics and protein levels. I also describe the development of DNAm GrimAge and other DNA methylation-based biomarkers of ageing. In Chapter 3, I carry out structured literature reviews of genome-wide and epigenome-wide association studies on plasma protein levels and associations. I also detail associations between DNAm

GrimAge and health outcomes as well as the main aims of this thesis. In Chapter 4, I describe the two cohort studies used in this thesis: the Lothian Birth Cohort 1936 and Generation Scotland. In Chapters 5-7, I perform genome-wide and epigenome-wide association studies on plasma levels of 92 neurology-related proteins, 70 inflammatory proteins and 282 AD-associated proteins, respectively. There are 422 unique proteins across these chapters. In Chapters 8 and 9, I investigate associations between DNAm GrimAge and measures of brain health and AD, respectively. In Chapter 10, I discuss the main findings of this thesis in light of its limitations along with recommendations for future work.

1 The ageing brain

The rapid ageing of the global population has prompted a sharp increase in the personal and societal burden of age-associated disease and disability. The number of individuals living with dementia is expected to triple from 50 million to 152 million by 2050 (1). Differentiating individuals with age-related cognitive changes from those with dementia will help to develop targeted preventative and treatment strategies. In this chapter, I describe normal cognitive ageing and the diseases that cause dementia. Given that access to live neural tissue is limited to some surgical interventions, I also outline the motivation for developing blood-based markers of non-pathological and pathological cognitive decline.

1.1 Cognitive ageing

1.1.1 Experience of normal cognitive ageing

Cognitive impairment in older adults is a significant health and social issue, and age-associated cognitive decline may precipitate dementia and illness (2). Normal cognitive ageing manifests through age-related deficits in distinct strata of human cognitive ability. Crystallised intelligence reflects acquired knowledge and experience. Crystallised intelligence rises gradually in early adulthood and remains stable until age 60 years after which it may begin to decline (Figure 1-1) (3). Fluid intelligence involves cognitive abilities that are related to mental manipulation, reaction time, access to and utility of working memory, processing speed and performance on non-verbal tasks (4). In adulthood, fluid intelligence peaks in the third decade of life and declines thereafter (3).

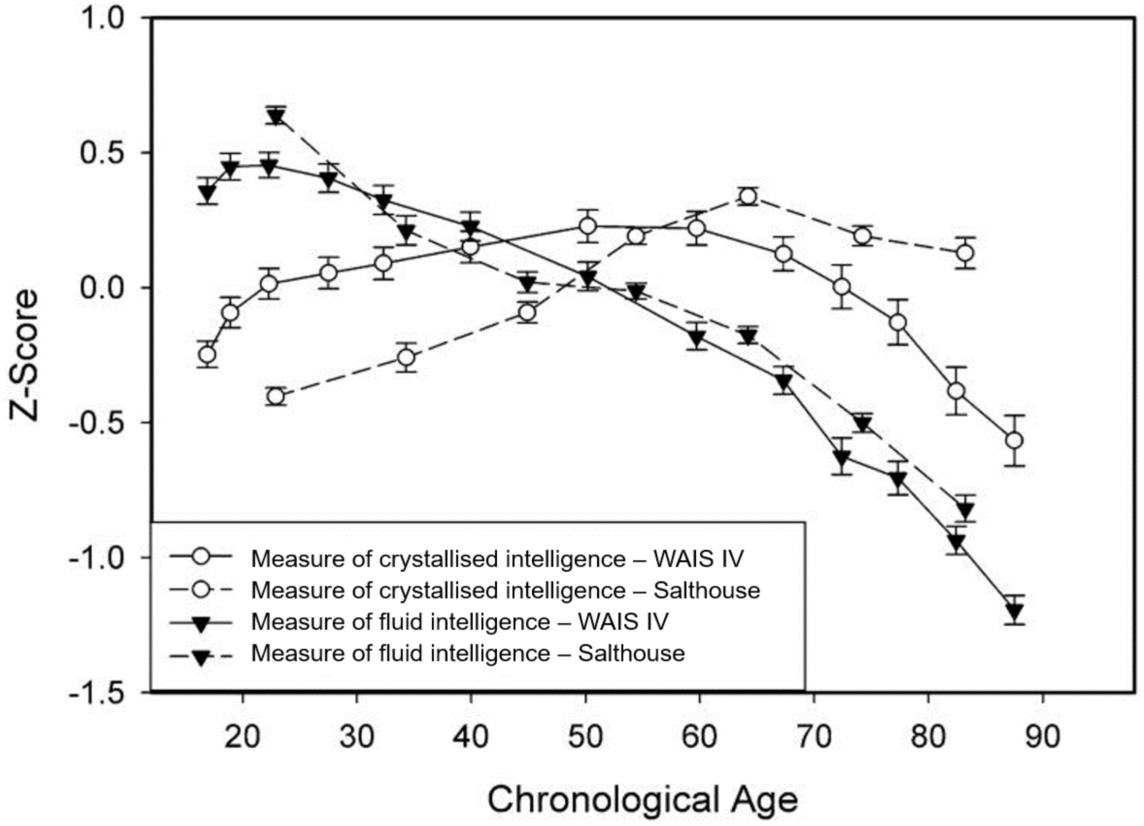

Figure 1-1. Changes in crystallised and fluid intelligence across the lifespan. Means of composite scores for crystallised intelligence and fluid intelligence are shown using white circles and inverted black triangles, respectively. Standard errors of means are shown by vertical lines. Dotted lines indicate tests from the Salthouse study whereas continuous lines show scores derived from an intelligence test: Wechsler Adult Intelligence Scale (WAIS) IV. Fluid intelligence tends to decline across adulthood. Crystallised intelligence increases in early adulthood and begins to decline after age 60-70 years. Figure adapted from Salthouse (2012) (3). Copyright: Annual Reviews.

The brain decreases in size in older age along with an enlargement of its ventricles (5). Furthermore, the loss of white matter in older age is more substantial than grey matter changes (6). The most affected areas include anterior white matter tracts, which underlie executive functioning, and the corpus callosum, which facilitates information transfer between brain hemispheres (7, 8). Grey matter changes begin after age 20 years and the most prominent alterations occur in the prefrontal cortex (9). Grey matter alterations reflect changes in dendritic morphology, axons and a significant loss of synapses (10). Synaptic loss is the greatest correlate of cognitive decline (11). Symptoms of dementia may appear upon loss of 40% of synapses (9, 12, 13).

1.1.2 Individual cognitive trajectories

On average, individuals experience age-related declines in certain cognitive domains. However, there is significant heterogeneity in individual trajectories of cognitive decline. Early-life cognitive ability explains around 50% of the variance in later-life cognitive ability (14, 15). The remainder of inter-individual variability might reflect several, non-exclusive factors including genetic influences, modifiable lifestyle factors, socioeconomic status and co-morbidities such as cardiovascular disease (CVD) (15).

Molecular correlates of cognitive decline are outlined in Sections 1.5 and 1.6. Lifestyle factors such as diet and smoking are associated with cognitive ageing, but the directionality of these associations is unclear (16, 17). Lifestyle factors have causal influences on age-related diseases. However, individuals with better cognitive abilities might have more favourable health-related behaviours and greater adherence to public health advice. Cognitive function and disease risk mechanisms might also have shared underlying genetic or biological pathways (18). Studies in cognitive epidemiology suffer from selection bias as those entering the study might exhibit systematic differences from the general population, including healthier lifestyles and more years of education. These studies might also be affected by attrition or survival bias, and practice effects on cognitive tasks in longitudinal study designs (19).

1.2 Dementia

Some individuals experiencing age-related cognitive decline progress to developing dementia. Dementia is a neurological disorder characterised by impaired memory, personality changes, language deficits, impaired reasoning and other cognitive disruptions (20). There are four major subtypes of diseases that cause dementia: Alzheimer’s disease or AD, vascular dementia (VaD), Lewy body dementia (LBD) and frontotemporal dementia (FTD). In most subtypes, an accumulation of toxic proteins creates aggregates that disrupt neuronal and synaptic function (Figure 1-2). Dementias are therefore known as “proteinopathies”. The diseases differ according to the proteins that aggregate, the brain regions that are affected, clinical symptomatology and severity (21).

Biological and molecular correlates of dementia are outlined in the following sections. Of note, the *APOE* gene on chromosome 19 possesses three major versions (alleles) - ε2, ε3 and ε4. Relative to ε3, ε2 is protective against AD whereas ε4 confers a greater risk for developing AD. Possessing one copy of the ε4 allele leads to a 3-times greater odds for developing AD relative to individuals with two ε3 alleles. Individuals with two ε4 copies have a 14-times greater odds for developing AD than those with two copies of the ε3 allele (22). Several modifiable risk factors are associated with late-life dementia: excess alcohol use, higher body mass index, smoking, high blood cholesterol, low physical activity and poor diet (23, 24). A meta-analysis estimated that the relative risk of dementia was 1.20 (95% confidence interval (CI) = 1.04-1.39) for individuals with one modifiable risk factor, 1.65 (95% CI = 1.40-1.94) for those with two risk factors and 2.21 (95% CI = 1.78-2.73) for those with three or more risk factors, relative to those with no risk factors (25). Depression, anxiety, diabetes and low social engagement are also established risk factors for developing dementia (26, 27). In the next section, I describe the different strategies that are used to assess cognitive and neurological function. These methods are required to refine our knowledge of pathways that underlie cognitive ageing and those that cause dementia.

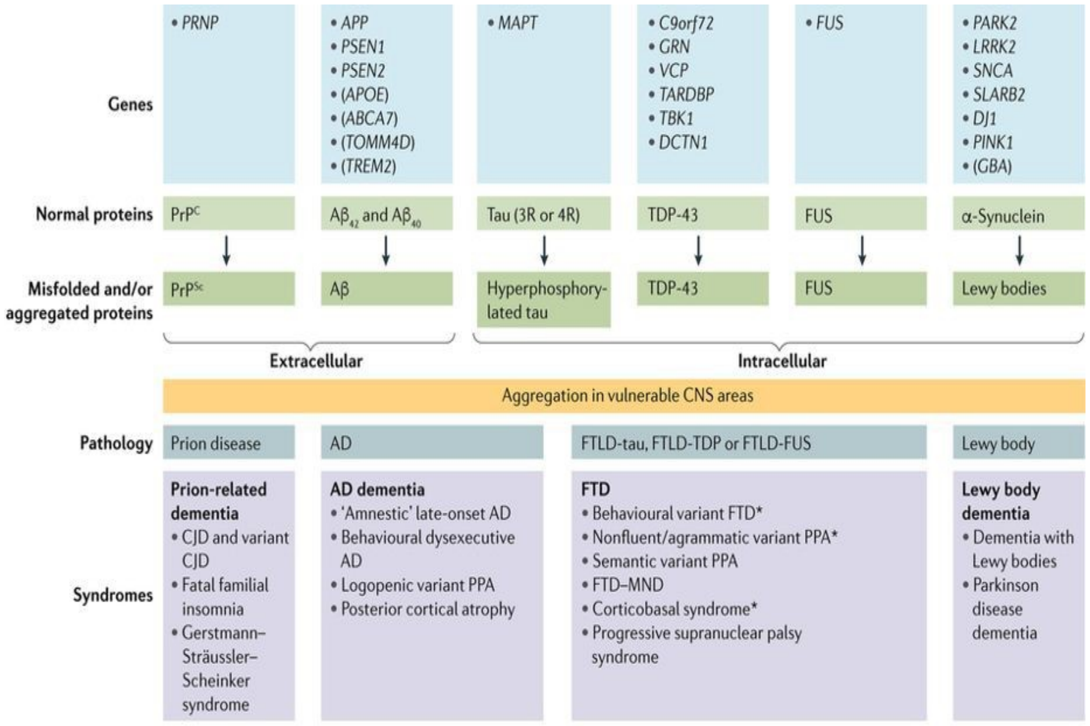

Figure 1-2. Molecular correlates of dementia-related proteinopathies. This figure shows genes with full penetrance that influence the misfolding or aggregation of six proteins: cellular prion protein (PrPC), amyloid-β (Aβ)-42 (and, to a lesser extent, Aβ40), tau, TAR DNA-binding protein 43 (TDP-43), fused in sarcoma (FUS), and α-synuclein. Causative genes are reported without parentheses and risk genes are shown in parentheses. In addition to prion disease, three major dementia-related proteinopathies are recognised: Alzheimer's disease (AD), frontotemporal dementia (FTD) and Lewy body dementia (LBD). Vascular dementia may also feature an accumulation of Aβ plaques in small vessel walls. These plaques mostly consist of Aβ40. Several clinical syndromes may occur within each subtype. Asterisks indicate FTD syndromes that may also be associated with AD neuropathology. Aβ, amyloid-β; CJD, Creutzfeldt–Jakob disease; CNS, central nervous system; FTD–MND, FTD with motor neuron disease; PPA, primary progressive aphasia. Figure taken from Elahi and Miller (21). Copyright: Nature publishing group.

1.3 Measures of cognitive health

1.3.1 Cognitive tests

Approximately 50% of the inter-individual variance on intelligence tests can be explained by a latent factor of general cognitive ability termed the *'g'* factor (28). In 1904, Charles Spearman observed that scores from school-children on different intelligence examinations were highly correlated within individuals, and that these scores agreed well with the teacher's ratings of student cognitive abilities (29). This 'positive manifold' provided a basis for a general factor of cognitive ability, or *'g'*. The *'g'* factor does not reflect an absolute indicator of intelligence, but rather a comparator against other individuals in a given dataset of cognitive test scores. Refinements to Spearman's model have been made by Horn & Cattell (30), Carroll (31), Vernon (32) and Johnson & Bouchard (33) to allow for more complex representations of human cognitive abilities and distinct cognitive domains.

The most commonly used clinical test to measure cognitive decline is the Mini-Mental State Examination (MMSE) (34). The MMSE is a 30-point questionnaire that assesses attention, memory, orientation to time and space, language and visual construction. Scores below 23 or 24 points are used to indicate significant cognitive impairment (35). The MMSE had a specificity and sensitivity of 81% and 89%, respectively, in a recent meta-analysis of dementia screening tests (n = 49,000 individuals). The Mini-Cog and Addenbrooke's Cognitive Examination Revised tests showed specificities of 91% and 92% and sensitivities of 86% and 89%, respectively (36).

1.3.2 Neuroimaging methods

The majority of neuroimaging biomarkers for cognitive ageing and dementia focus on AD pathology. The hallmarks of AD include abnormal accumulations of hyperphosphorylated tau and beta-amyloid plaques. These aberrations are detectable in brain scans some 15 years before the onset of dementia symptoms (37). Amyloid burden can be quantified *in vivo* using carbon-based radiolabelled tracers such as Pittsburgh Compound-B in positron emission tomography (PET) scans (38). Amyloid PET imaging is useful for staging AD progression and enabling patient stratification in secondary prevention trials (39-42). Amyloid PET scans would cost the UK National Health Service an additional £113 million per annum if they are used to predict AD conversion from those with mild cognitive impairment (43). Fluorine-based PET tracers against tau show increased uptake in early AD (44). Imaging methods are of great importance in VaD as they can detect large vessel strokes, small vessel disease and microvascular changes (45). However, neuroimaging methods are costly and resource-intensive.

1.3.3 Cerebrospinal fluid (CSF) biomarkers

CSF surrounds brain and spinal cord tissue and performs key roles in shock absorption, waste removal and the regulation of solute levels in the central nervous system (CNS). Ependymal cells within the choroid plexus of brain ventricles produce the majority of CSF (46). Given the proximity of CSF to brain tissue, changes in the CSF may reflect concomitant alterations in brain metabolism or neuropathology (47, 48). Changes in CSF beta-amyloid and tau levels are detected 15 years before the expression of cognitive decline (37). A further 27 proteins in CSF were associated with AD diagnosis in a recent systematic review of 47 independent studies. CSF levels of the neuropeptide VGF (VGF nerve growth factor inducible), which is involved in synaptic plasticity (49), were consistently lower in AD patients relative to controls. The glial pro-inflammatory glycoprotein CHI3L1 (Chitinase 3 Like 1, YKL-40) was consistently upregulated in AD cases (50). Collecting CSF through lumbar puncture is an invasive process and might induce post-procedure headaches (51).

1.4 Genetics of cognitive ageing and dementia

1.4.1 Structure and variation of the human genome

The haploid human genome consists of 23 chromosomes (22 autosomal and one sex-determining) and approximately three billion base pairs of deoxyribonucleic acid (DNA). Four nucleotides are present in the nuclear genome (adenine, cytosine, guanine and thymine) and the specific order of these nucleotides dictates an individual's genetic code. The genetic code provides a framework for the expression and assembly of approximately 20,000 distinct proteins. Individuals differ, on average, by ~0.1% in their DNA sequences. The majority of this variation does not directly impact protein structure or function as approximately 1.5% of the genome is protein-coding. Other portions of the genome might indirectly affect protein structure, function and abundance (52, 53).

There are two broad categories of human genetic variation. First, single-nucleotide variations refer to different versions of nucleotide bases at a given point in the genome. Single-nucleotide polymorphisms (SNPs) are single-base variations in at least 1% of individuals (54, 55). On average, SNPs occur every ~300 bases along the genome and over 80 million SNPs have been characterised (56). Second, structural differences between genomes induce genetic variation. Structural alterations include gene copy numbers, insertions and deletions (indels), inversions, translocations and block substitutions (57, 58). SNPs are the most common form of genetic variation and are the focus of this chapter and the empirical work in this thesis.

1.4.2 Genome-wide association studies

Genome-wide association studies (GWAS) are observational studies of associations between SNPs across the genome and traits of interest (59). The inception of chip-based genome-wide SNP microarrays and reference genetic variation maps have permitted large-scale GWAS (60, 61). There are over 70,000 SNP-trait associations across >5,500 GWA studies as of 2019 (62).

Association studies must consider linkage disequilibrium (LD), which relates to the statistical non-independence of two or more loci that occur together within the population. The locus that is causal for a given trait may arise from an unobserved, historical mutation event. However, nearby genotyped SNPs can tag this causal variant in GWAS (63). Unknown genotypes are imputed or predicted from population-specific reference LD panels. Imputation accuracy is high for common variants but it decreases for rare and low-frequency SNPs (1% <minor allele frequency (MAF)< 5%) (64). Imputation increases the power of association studies, resolves genotyping errors and facilitates fine-mapping studies (65).

Population stratification might confound GWAS and results from imbalances in allele frequencies between comparison groups, i.e. cases versus controls. These imbalances occur due to systematic differences in ancestry (66, 67). A common approach to control for population stratification is to apply dimension reduction methods and reduce SNP data to a much smaller set of orthogonal components. Dimension reduction methods include principal component analysis (PCA) and multidimensional scaling (MDS). The components are then used as covariates in GWA designs to account for population structure (68). GWAS require corrections for the multiple tests that are performed. Hundreds of thousands of genotyped SNPs, or millions of imputed SNPs, might be correlated with the trait of interest. The significance threshold is often corrected to account for ~1 million effective independent common variants across the genome ($\alpha = 0.05/1 \times 10^{6} = 5 \times 10^{-8}$) (69). Methods including Bayesian analyses and permutation testing are also used to control for multiple testing in GWAS (70).

GWAS on cognitive function

In a recent GWAS, 148 independent loci were associated with general cognitive ability (n = 300,486, age range = 16-102 years) (71). These loci included variants in the genes *GATAD2B* and *SLC39A1*, which associate with intellectual disability and AD, respectively (72, 73). The study estimated that the heritability of general cognitive ability captured by common SNPs is 25% (standard error (SE) = 0.06%). This is in agreement with previous GWAS on cognitive function, which provided SNP-based heritability estimates of 20-30% (n = 35,298 - 78,308) (74-77). Genome-wide analyses of cognitive decline have been less described due to a relative paucity of longitudinal cognitive data. The *APOE* locus was significantly associated with cognitive decline in healthy individuals and those with mild cognitive impairment (78-81).

GWAS on dementia

Large GWAS on AD and family history of AD have identified dozens of possible risk loci (82-85). Sample sizes ranged from 314,278 to 1,126,563 participants. The number of loci associated with clinically diagnosed AD or AD risk ranged from 26 to 75 loci. These studies have highlighted variation in the *TOMM40-APOE-APOC2* locus and amyloid processing, inflammatory and microglial pathways as strong candidate risk mechanisms in AD (82-85). Only two GWAS on VaD have been conducted to date. One variant near the androgen receptor gene on the X chromosome was associated with VaD at genome-wide significance (86, 87). GWAS on FTD and LBD are also limited (88, 89).

1.5 DNA methylation in cognitive ageing and dementia

1.5.1 Molecular mechanisms of DNA methylation

DNA methylation is an important epigenetic mechanism in eukaryotic cells. Epigenetic mechanisms contribute to differential patterns of gene expression across different cell types (90). DNA methylation is influenced by genetic sequence variations and environmental variables (91-93). Smoking is the strongest lifestyle correlate of DNA methylation (94). Environmental influences such as diet, alcohol and pollutants also associate with DNA methylation patterns (95-97).

DNA methylation involves the addition of methyl groups to the fifth carbon of cytosine nucleotides. In humans, this typically occurs in the context of cytosine-guanine dinucleotides (CpG site). DNA methylation is directly involved in protein regulation (98-100). DNA methylation is also responsible for imprinting of genes in parent-of-origin specific manners (101-103) and X-chromosome inactivation (103, 104).

There are approximately 28 million CpG sites across the genome (105, 106) and 60-80% of these CpG sites are methylated in mammals (107). The absence of DNA methyltransferases (DNMTs), which catalyse the addition of methyl groups to DNA, results in embryonic lethality. Therefore, DNA methylation is essential for life (108-110).

1.5.2 Measurement of DNA methylation

The most popular method for distinguishing between methylated DNA and unmethylated DNA is bisulfite conversion. DNA is treated with sodium bisulfite, which deaminates unmethylated cytosines to uracil and leaves methylcytosines intact (111-113). Uracil bases are then converted to thymines by polymerase chain reaction. Whole-genome bisulfite sequencing is the ‘gold-standard’ approach for assessing DNA methylation levels. However, it is expensive and labour-intensive (114). Methylation arrays are popular alternatives for detecting methylation signals from bisulfite-treated DNA. These arrays permit a rapid and cost-effective quantification of methylation levels across representative subsets of CpG sites in the genome (Figure 1-3) (114).

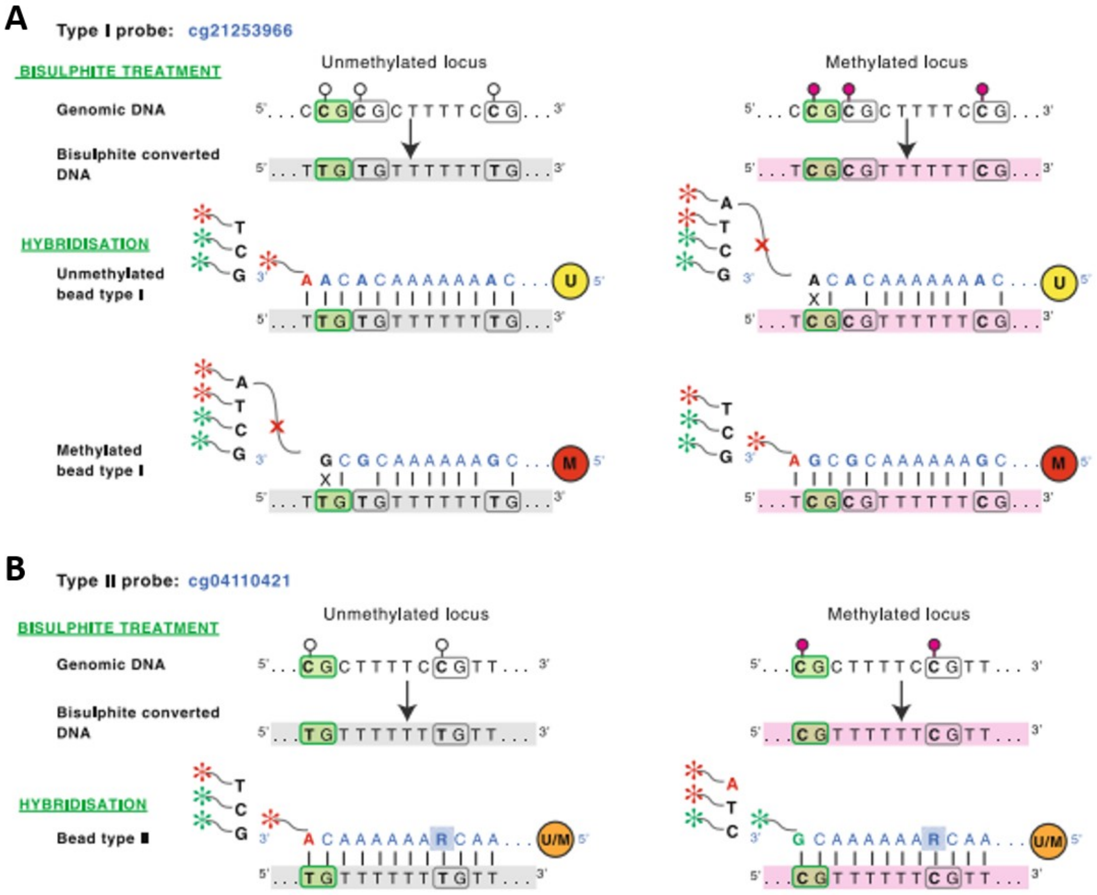

Figure 1-3. Infinium chemistry for assessing methylation in sodium bisulfite-treated DNA. A) In type I probes, there are two bead types. Unmethylated-bead types recognise unmethylated versions of CpG sites (thymines). The adjacent nucleotide is a template to extend the unmethylated probe by a single-base (single-base extension). If an adenine or thymine is incorporated, a red fluorescent signal is emitted. If cytosines or guanines are added, the signal is green. Similarly, methylated-type beads will only bind to methylated loci. Single-base extension induces the emission of fluorescent signals as in the unmethylated-bead type design. B) In type II probes, there is only one bead type. This probe does not use single-base extension to differentiate between methylated and unmethylated sites. The methylated cytosine or the bisulfite-converted thymine is a template for the incorporation of a complementary nucleotide into the probe sequence. Green signals are emitted for methylated sites whereas red fluorescence signals are released if the CpG site is unmethylated. DNA, deoxyribonucleic acid. Figure adapted from Pidsley *et al*. (115). Copyright: BioMed central.

1.5.3 Epigenome-wide association studies

In epigenome-wide association studies (EWAS), methylation levels at individual CpG sites are correlated with continuous or categorical outcome variables. CpG methylation may vary over time and is influenced by genetic and non-genetic factors (116). This leads to key three considerations in EWA designs. First, DNA methylation levels at individual CpG sites may vary across different cell types. Failure to account for cell type heterogeneity can lead to the identification of false positives (117). Second, an underlying molecular or environmental factor might influence DNA methylation levels and the outcome of interest thereby generating a non-causal association (118). Third, reverse causation can occur if an outcome of interest affects CpG methylation. For example, DNA methylation in the *HIF3A* gene associates with body mass index (119) and adiposity (120-123). Causal inference methods suggested that an individual's body mass index likely influenced methylation at *HIF3A* and not vice versa (124).

EWAS require stringent correction for the multiple tests that are performed. Saffari *et al.* (2018) estimated a significance threshold of $P < 3.6 \times 10^{-8}$ for the Illumina 450k array (~450,000 CpG sites) (125) and Mansell *et al.* (2019) suggested a threshold of $P < 9 \times 10^{-8}$ for the Illumina EPIC array (~800,000 CpG sites) (126). EWAS commonly utilise methodologies from GWAS where each CpG site is tested in turn. Sample relatedness, population stratification and cell type heterogeneity might induce correlations between distal probes. Correlations among probes and omitted variable bias can result in biased effect size estimation and model overfitting (127). Several statistical frameworks aim to address these potential biases by using (i) reference data to predict cell type compositions or (ii) fitting latent covariates computed by applying dimension reduction methods to DNAm data (128-131). However, the influence of confounders might not be explained in full by reference-based predictions or fixed numbers of principal components. Linear mixed-effects models and Bayesian frameworks use reference-free approaches and estimate probe effects conditional on one another. These methods account for intercorrelations between distal probes induced by unknown confounders and data structure. Mixed model and Bayesian penalised regression association studies are performed in OSCA (OmicS-data-based Complex trait Analysis) and BayesR+ software, respectively (132, 133). OSCA and BayesR+ permit the inclusion of other data sources such as genome-wide SNP data. Therefore, the individual and joint contributions of genetic and epigenetic data towards the variance in phenotypes of interest can be estimated. Linear mixed-effects models and Bayesian strategies are also used in GWAS to account for population stratification, cryptic relatedness and inflation due to polygenicity and to increase detection power (134-140). I further describe the application of linear regression, mixed-effects models and BayesR+ in GWAS and EWAS in Chapter 4.

EWAS on cognitive function

A recent meta-analysis on seven cognitive traits identified epigenome-wide significant associations between cg21450381 (intergenic, chromosome 12) and MMSE scores, and cg12507869 (*INPP5A* gene, chromosome 10) and phonemic verbal fluency (n ≤ 6,809) (141). Blood CpG methylation in *AGBL4* and *SORBS1* showed suggestive associations with a 10-year change in cognitive ability (n = 486, P = 9.01×10^{-7} and 5.28×10^{-6}, respectively) (142). *CLDN5* methylation in dorsolateral prefrontal cortex tissue associated with longitudinal changes in cognitive function (n = 636) (143).

EWAS on dementia

Over 200 CpG sites were associated with AD neuropathology in *post-mortem* tissue from 1,453 donors (144). EWAS implicate *BIN1* and *ANK1* as important correlates of AD pathology in *post-mortem* tissue (145, 146). Sanchez-Mut *et al.* (2016) provided evidence for *ANK1* dysregulation in AD and LBD using *post-mortem* dorsolateral prefrontal cortex tissue (n = 32 AD patients, 23 LBD patients and 32 controls) (147). In relation to blood-based EWAS, hypermethylation in the *HOXB6* gene associated with AD diagnosis (n = 284) (148). Further, twenty differentially methylated CpG sites were associated with FTD (149). It has been estimated that 1-10% of CpG sites show significant correlations between their brain and blood methylation levels (150, 151).

Therefore, comparisons between brain- and blood-based EWAS are challenging.

1.6 Motivation for identifying blood-based markers of disease

In Section 1.4, I outline that neuroimaging methods and CSF collection are limited by high costs, the need for specialised training and invasiveness (152). Additionally, routine access to *in vivo* neural tissue is infeasible. Several studies have identified blood-based correlates of cognitive ageing and dementia as non-invasive alternatives to lumbar punctures or imaging-based approaches (153-179). Blood-based strategies have the following advantages (180, 181):

- safe collection
- serial sampling is convenient and scalable for population studies
- serial blood sampling is more accessible for ill and older individuals than other methods
- tissue is amenable to multiple types of analysis
- blood may capture lifestyle profiles and overall health status of body

In Sections 1.5 and 1.6, I describe known molecular correlates of cognitive ageing and dementia. Variation in genetic and DNA methylation profiles drives changes in gene expression and protein levels. Identifying blood protein correlates of cognitive impairment might highlight potential therapies and biomarkers. Conducting GWAS on blood protein levels can help to determine whether proteins are causally associated with disease states. EWAS that use DNA methylation data could reveal whether genetic variation or environmental factors mediate associations between blood proteins and disease states (Figure 1-4).

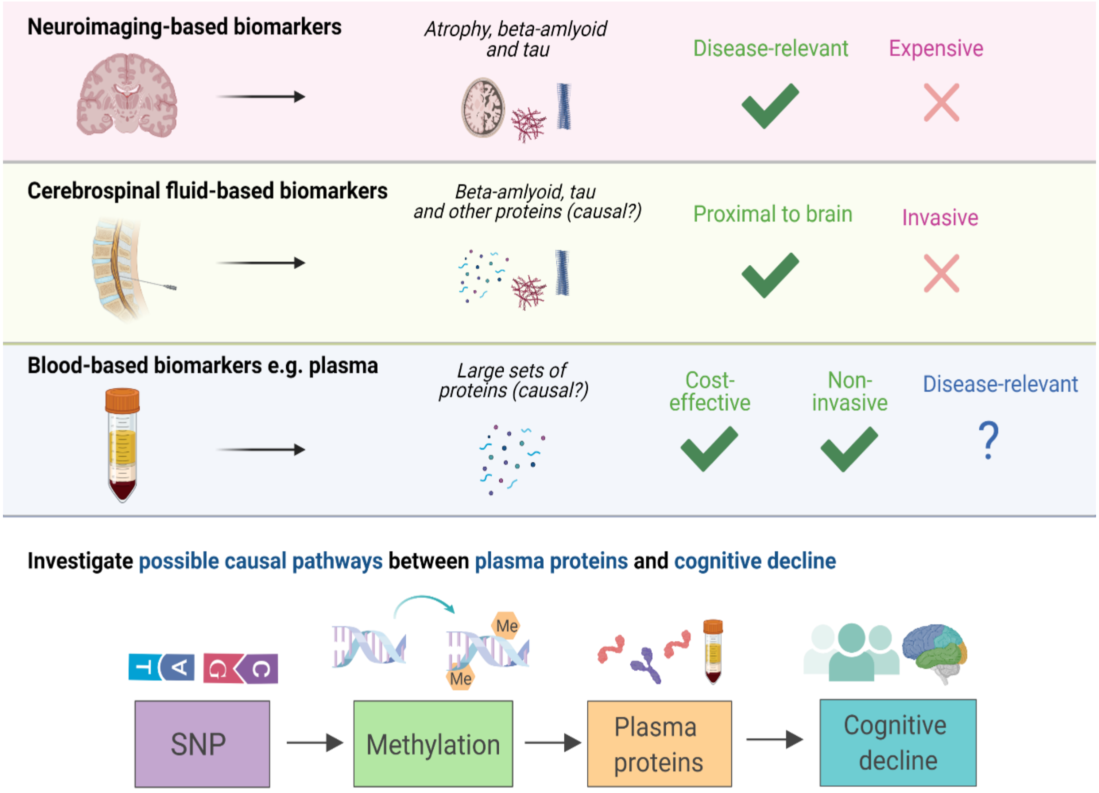
Neuroimaging-based biomarkers
Atrophy, beta-amlyoid and tau
Disease-relevant
✓
Expensive
✗
Cerebrospinal fluid-based biomarkers
Beta-amlyoid, tau and other proteins (causal?)
Proximal to brain
✓
Invasive
✗
Blood-based biomarkers e.g. plasma
Large sets of proteins (causal?)
Cost-effective
✓
Non-invasive
✓
Disease-relevant
?
Investigate possible causal pathways between plasma proteins and cognitive decline
T A G C
Me
Me
SNP
Methylation
Plasma proteins
Cognitive decline

Figure 1-4. Rationale for identifying blood-based markers of cognitive decline. First, neuroimaging methods are limited by high costs and poor scalability. Second, obtaining cerebrospinal fluid is invasive and requires specialist training. Third, identifying blood-based proteins whose levels correlate with cognitive ageing and dementia is inexpensive and non-invasive relative to other approaches. GWAS and EWAS can be combined with causal inference methodologies to assess whether there are plausible molecular pathways that link blood proteins to complex diseases. EWAS, epigenome-wide association studies; GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism. Figure created using Biorender.com.

1.7 Summary

Identifying blood-based correlates of cognitive ageing and dementia may help to predict individual risk profiles and allow for the implementation of preventative strategies. Elucidating the molecular factors that underpin protein levels will help to disentangle whether alterations in protein levels are a cause or consequence of disease. In the next chapter, I describe statistical methods for determining the contribution of genetic and epigenetic factors towards inter-individual differences in protein levels. I also outline how these data can be integrated to develop effective biomarkers of healthy ageing.

2 Genome-wide and epigenome-wide studies on protein levels

In this chapter, I describe technologies for quantifying blood protein levels. I also outline statistical methods for identifying the genetic and epigenetic factors that underpin protein levels and how these data may inform the biology and prediction of cognitive ageing and dementia.

2.1 The plasma proteome

The proteome refers to the full complement of proteins expressed by the human genome. Marc Wilkins coined this term in 1994 as an analogy to the genome (182). In 2010, the Human Proteome Project was established to collect evidence for the 20,379 proteins (2020 estimate) that are predicted to exist based on protein-coding regions of the genome. In total, 1,421 'missing' proteins remained undetected in high-throughput proteomic screens. The existence of missing proteins is inferred by evidence at the transcript level, homologous species and bioinformatics-based predictions (183).

2.1.1 Composition of the human plasma proteome

Human blood plasma provides an accessible window to an individual's proteome. Plasma constitutes 55% of total blood volume and refers to the liquid portion of blood. The Human Plasma PeptideAtlas catalogues 3,509 proteins in plasma (2017 estimate) that have two unique mapping peptides across different proteomic datasets. There are >1,300 further proteins with ambiguous evidence (184).

In contrast to plasma, serum is fluid and does not contain clotting factors such as fibrinogen (185). Plasma proteins are analysed in this thesis and are therefore the subject of this section. However, high abundance proteins may preclude the detection of low abundance proteins in both plasma and serum. Approximately 22 proteins make up 99% of the protein weight in plasma and serum (186). The dynamic range of the plasma proteome ranges from 9 to 13 orders of magnitude (187). As an illustrative example, identifying a low abundance protein such as troponin (5 ng/ml) among albumin molecules (45 mg/ml) is analogous to identifying one person within the entire human population (188). The wide dynamic range of protein concentrations is a significant source of analytical complexity for proteomic platforms, which I outline in the following section.

2.1.2 Proteomic technologies

High-throughput proteomic technologies are broadly categorised into mass spectrometry (MS) and affinity-based assays. In the next sections, I detail the measurement process for two affinity-based technologies, which are used to quantify protein levels in my empirical analyses.

Affinity-based technologies

The principles of proximity extension assays (PEAs) and proximity ligation assays (PLAs) are shown in Figure 2-1. Paired antibodies are coupled to oligonucleotides. The antibodies target different epitopes on a given protein of interest (189-191). In PEA, oligonucleotides share complementary sequences and hybridise when both detection antibodies bind to the target protein. The resultant DNA strand can be amplified and quantified through polymerase chain reaction. PLA uses a bridging aptamer (splint) that ligates the oligonucleotides. A continuous DNA strand is created, which can be amplified and quantified.

Dual protein recognition allows for complementary probe hybridisation (PEA) or ligation (PLA). This limits the occurrence of cross-reactive events when compared to traditional multiplex immunoassays (190). Olink® Bioscience (Uppsala, Sweden) employs PEA technology and offers measurement of up to 1,536 protein levels as of May 2021.

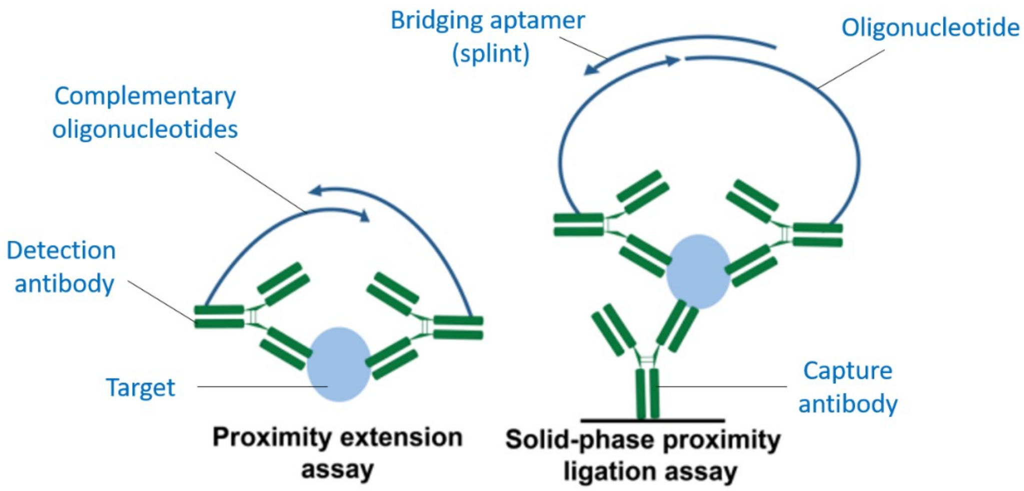
Bridging aptamer (splint)
Oligonucleotide
Complementary oligonucleotides
Detection antibody
Target
Capture antibody
Proximity extension assay
Solid-phase proximity ligation assay

Figure 2-1. Proximity extension assays and proximity ligation assays. In proximity extension assays (left), two detection antibodies recognise different epitopes on the same target protein. The detection antibodies are attached to probes that share complementary sequences. Probes hybridise when both detection antibodies bind to the target protein. In proximity ligation assays (right), a capture antibody binds to the target protein. Instead of hybridisation through complementary sequences, a bridging aptamer (or splint) ligates probes and creates a single-stranded DNA molecule. Figure adapted from Landegren *et al.* (192). Copyright: Elsevier.

Alternative methods use fluorescently tagged aptamers that bind directly to proteins of interest. Aptamers are short single-stranded molecules of DNA or RNA that can bind to protein targets, and were first described by two independent groups in 1990 (193, 194). Aptamers have three major advantages over antibodies in protein detection: (i) greater versatility, (ii) enhanced specificity through sequence modifications and (iii) greater stability (190). Systematic evolution of ligands by exponential enrichment (SELEX) is a commonly used approach to generate aptamers with high affinity for target proteins. In SELEX, large oligonucleotide libraries are generated and exposed to a target ligand. The stringency of the elution conditions is increased in successive rounds to identify highly specific aptamers (194). SomaLogic uses SELEX technology and can measure over 4,200 proteins simultaneously (Figure 2-2).

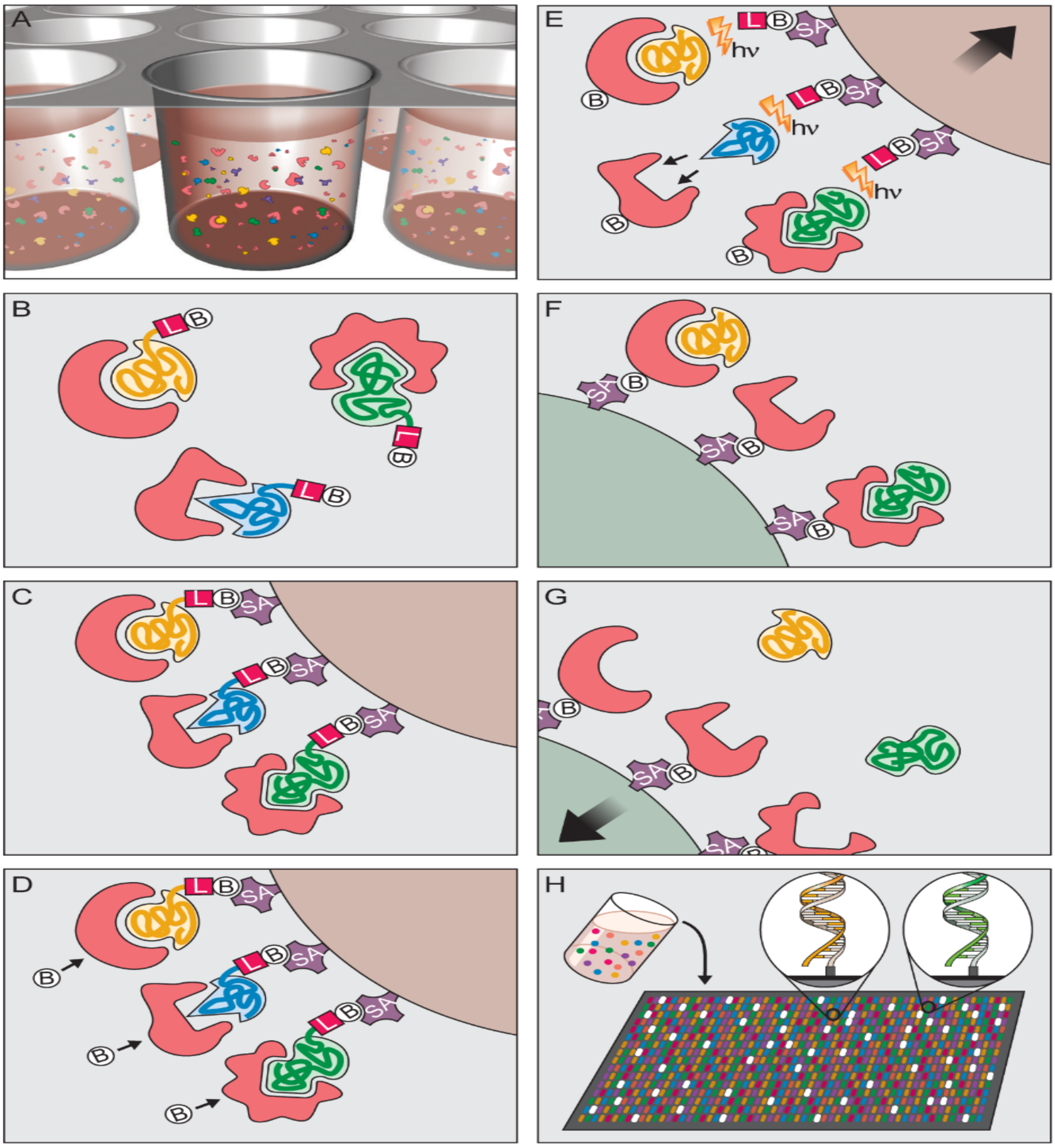

Figure 2-2. Principles of SomaLogic assay. (A) Slow Off-rate Modified Aptamers (SOMAmers®) are mixed with protein samples. (B) SOMAmers bind to cognate proteins. (C) Streptavidin (SA)-linked beads bind SOMAmer-protein complexes. Unbound proteins are washed away. (D) Proteins are tagged within biotin molecules. (E) Ultraviolet light (hv) breaks bonds between SOMAmers and SA beads. Competitors are added and non-cognate complexes dissociate (here, blue SOMAmer is lost). (F) The remaining SOMAmer-protein complexes are attached to new SA beads via the biotin molecule. Free SOMAmers are washed away. (G) At high pH levels, proteins dissociate from their SOMAmers. (H) SOMAmers are added to microarrays in which they bind to complementary single-stranded probe sequences. Protein concentrations are quantified as relative fluorescence units. SA, streptavidin. Figure taken from Gold *et al.* (195). Copyright: PLOS.

The major advantage of SomaLogic technology over other affinity-based methods is the large increase in protein coverage. However, this platform may show more off-target binding than Olink technology.

2.2 Genome-wide studies on protein levels

GWAS on protein levels detect SNPs that associate with protein abundance in a given tissue of interest. These SNPs are termed protein quantitative trait loci (pQTLs) (196). SNPs might reflect deleterious substitutions in coding regions of genes and therefore change protein structure. However, pQTLs can affect any aspect of protein regulation, including post-translational modifications and protein degradation (197). Protein QTLs may also affect gene expression or transcript levels (eQTLs), messenger RNA (mRNA) splicing (sQTLs) and histone activity, which influences transcription factor binding (hQTLs) (198-202). Later in this chapter, I discuss statistical methods that can help to mechanistically link pQTLs to protein levels and assess their causal relationships with diseases. In the following section, I briefly delineate conceptual differences between GWAS on complex human traits and GWAS on protein abundances.

Many variants associated with complex traits are non-coding and are not in LD with variants in coding regions (203). Therefore, they may have indirect and regulatory effects on protein expression, thereby motivating efforts to examine whether pQTLs overlap with variants that associate with common disease states (204-206). The polygenic nature of complex traits implies that many proteins might influence a given trait with effect sizes of varying magnitudes. The possible involvement of non-coding RNA expression cannot be ruled out and is an active, emerging area of research (207).

GWAS on complex traits, in part, aim to identify proteins with possible roles in biological and disease-related pathways. In contrast with this, GWAS on protein levels aim to detect variants that regulate the abundances of proteins in different tissues. Whereas variants associated with complex traits can have small effect sizes, variants associated with protein levels exhibit relatively large effect sizes if they are acting in *cis*. *Cis* variants are in the gene that encodes the protein and might directly impact protein structure or function (204). Protein QTLs may also represent *trans* effects where variants lie in distal loci or on a chromosome distinct from the protein-encoding gene. *Trans* effects often exhibit smaller effect sizes than *cis* variants and require larger sample sizes for their detection (208).

Existing GWAS on blood protein levels are outlined in a literature review within Chapter 3. Next, I introduce key statistical approaches that use pQTLs to examine whether proteins have causal roles in disease states.

2.3 Mendelian randomisation - causal inference

Identifying biomarkers for disease states is a major goal of epidemiological research. These approaches are limited by the possibility that associations between proteins and disease may reflect correlation and not causation. There are two main challenges in testing for causation: confounding factors and reverse causation (209). Methods that are robust to these complicating factors are essential in observational epidemiology.

2.3.1 Assumptions in Mendelian randomisation

Analogous to randomised control trials, Mendelian randomisation (MR) assesses the effect of long-term exposure to a given variable on health outcomes (Figure 2-3). Genetic variants are used as instrumental variables (IV), or instruments, to proxy for the effect of a given risk factor or exposure of interest (210). MR relies on Mendel’s law of segregation. At conception, an individual’s genotype is randomly assorted from parental genotypes. We assume that confounding influences, such as smoking or socioeconomic status, are randomly distributed between genotype groups. Complex disease states should not influence an individual’s germline genotype. Therefore, we assume that MR is robust to reverse causation.

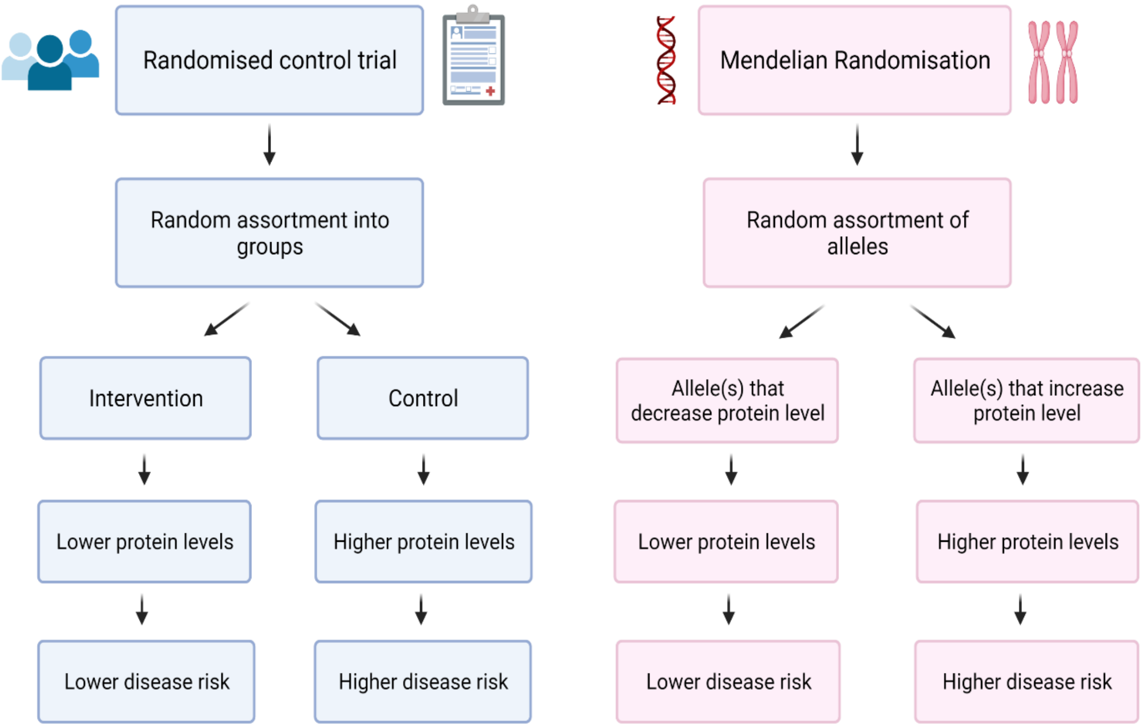

Figure 2-3. Conceptual similarities between randomised control trials and Mendelian randomisation. In both methods, individuals are randomly allocated to one of two groups. In randomised control trials, individuals are allocated to a control or treatment arm. Mendelian randomisation rests on Mendel's second law of segregation, which states that individual genotypes at a given locus are randomly assigned according to the paternal genotypes. One group may, on average, have a higher or lower exposure to a given risk factor. The long-term effect of the risk factor on an outcome is assessed by comparing the proportion of outcome events in both groups. We assume that confounders are balanced between groups. Reverse causation should also be absent. Figure created using Biorender.com.

There are three key assumptions for valid instruments (211) (Figure 2-4):

1. **Relevance assumption:** the variant is directly associated with the exposure (typically at genome-wide significance)

2. **Exclusion restriction assumption:** the variant affects the outcome through the exposure variable i.e. a protein biomarker

3. **Independence assumption:** the variant is not associated with confounders for the exposure-outcome relationship

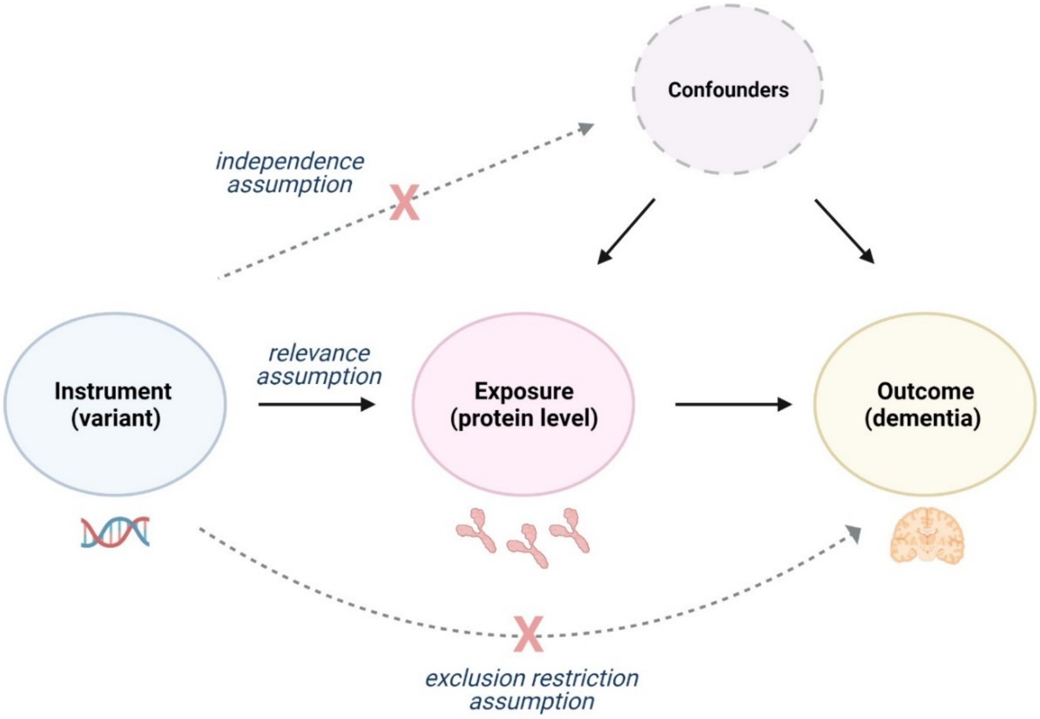

Figure 2-4. Directed acyclic graph showing the Mendelian randomisation paradigm. A genetic variant proxies for the intermediate phenotype, or risk factor, of interest. Given that alleles are randomly assigned at conception, the influence of confounders should be minimal (independence assumption). The variant should be strongly associated with the exposure variable (relevance assumption). The effect of the exposure on the outcome variable is assessed by using genetic variants as instrumental variables. The instrument should only affect the outcome through the exposure (exclusion restriction assumption). Figure created using Biorender.com.

Concerning the relevance assumption, weak instruments are identified by assessing the R^2 or F statistic of the genotype-exposure relationship. Instruments with an F statistic <10 are excluded from MR analyses (212, 213). In GWAS, the relationship between a variant and exposure might be overestimated due to 'Winner's curse'. This phenomenon leads to biased MR estimates if weak instruments are used (214).

Multiple instruments (> 2 SNPs) are required to perform formal tests of the exclusion restriction assumption. First, outliers may be identified by examining Cook's distance, which measures the influence of each estimate on MR regression slopes (339, 340). Second, the exclusion restriction assumption is violated when variants influence the outcome through pathways distinct from the exposure (horizontal pleiotropy). MR-Egger regression allows for variants to have pleiotropic effects only if the magnitude of pleiotropic effects is independent of the magnitude of the genotype-exposure effect (341). This is known as the InSIDE assumption (INstrument Strength Independent of Direct Effect). The intercept from MR-Egger regression provides an estimate of the average pleiotropic effect of instruments (215). A non-zero intercept implies that the average pleiotropic effect differs from zero or that the InSIDE assumption has been violated. MR-Egger regression is limited in that the exact nature of pleiotropy is uncertain. Several alternative methods including Bayesian and maximum likelihood approaches have been developed to address issues surrounding pleiotropy (215-221).

In a single sample, the independence assumption can be investigated through bias component plots. These plots show the relative bias in exposure-outcome regressions and instrument-outcome regressions using measured covariates (222, 223). As described in Section 1.4.2, confounding due to population stratification can be attenuated by controlling for principal components of ancestry (224). In two-sample studies, researchers can assess whether instruments are associated with baseline covariates in large existing GWAS.

2.3.2 Protein QTLs as instruments in Mendelian randomisation

Protein QTLs have three key advantages as instruments when compared to IV for conventional risk factors such as body mass index: (i) stronger instrument-exposure relationship, (ii) lower likelihood of horizontal pleiotropy and (iii) it is often easier to test relationships between IV and the exposure (e.g. *cis* pQTLs in *in vitro* studies).

Locally-acting pQTLs might influence distal protein levels at the genetic level through chromosomal interactions and the expression of regulatory microRNAs (225). Further, the relationships between *trans* pQTLs and protein levels merit elucidation in causal inference frameworks. Knowledge of the molecular mechanisms that link pQTLs to protein levels will allow us to fine-map causative pathways in disease. In the next section, I discuss statistical methods for investigating the molecular mechanisms that explain inter-individual differences in protein levels.

2.4 Colocalisation methods: linking protein QTLs to protein levels

Understanding the relationships between proteins and disease states is important for developing effective drug therapies. GWAS alone cannot reveal how pQTLs affect protein levels. The majority of diploid cells within the body have the same underlying genetic code, except for some B and T cells (226, 227). Therefore, it is essential to determine which cell types express the protein of interest and underlie the protein's role in disease states. Protein QTLs can be cross-referenced against known variants associated with other traits, such as gene expression, DNA methylation and disease outcomes (228). For instance, a variant is identified as an eQTL for a gene in immune cells and a pQTL for blood levels of the corresponding protein. Hence, the same underlying variant might influence cell type-specific gene expression and protein levels i.e. the two traits colocalise (229). Conceptually, this can be extended to any two, or more, traits with shared association signals.

2.4.1 Overview of colocalisation methodologies

Online tools such as Genevar were developed to allow for visual comparisons of overlaps between association signals for complex traits and gene expression (230). However, this does not represent a formal test for colocalisation. Early regression frameworks were limited by pre-specifying subsets of variants, which can introduce bias through 'Winner's curse' (231, 232). They also required individual-level genotype data, which may not be available in external datasets such as large eQTL databases. He *et al.* (2013) developed the Sherlock algorithm, which matches association signals from GWAS and eQTL data using summary-level statistics (233). Using a relaxed significance threshold, genome-wide eQTLs for a given gene are compared against risk loci for a disease of interest. However, eQTLs that fail to surpass the significance threshold in a reference sample might provide evidence against colocalisation if they associate strongly with the trait of interest.

In light of these limitations, Giambartolomei *et al.* (2014) developed a popular Bayesian R package called *coloc* to test for colocalisation between two traits (234). Coloc uses summary statistics for two distinct traits across a pre-defined genomic region (typically 400 kilobases). The locus is centred on a genome-wide significant variant that associates with both traits.

Coloc integrates over multiple possible configurations to assign posterior probabilities to five distinct hypotheses. The hypotheses are: hypothesis 0 (no causal variants), hypotheses 1 and 2 (causal variant for one of the traits) hypothesis 3 (distinct causal variants) hypothesis 4 (one common causal variant). By default, the method assumes that one in every 10,000 SNPs is causal for each trait. Furthermore, one in every 100 SNPs that associate with one trait is assumed to associate with the other trait. Priors may need to be adjusted depending on the study design. For instance, *trans*-acting QTLs may require more stringent thresholds. A posterior probability of greater than 95% provides strong evidence in favour of a given hypothesis. Posterior probabilities are a measure of correlation as opposed to causation. A single eQTL may be associated with the expression of multiple genes in a given locus. Several eQTLs may be colocalised between genes and also with the trait of interest due to LD. High posterior probabilities will be present for all genes. Nevertheless, coloc provides a useful overview of colocalisation methodologies. Figure 2-5 provides an illustrative example of the coloc framework. Alternative methods are discussed in the next section.

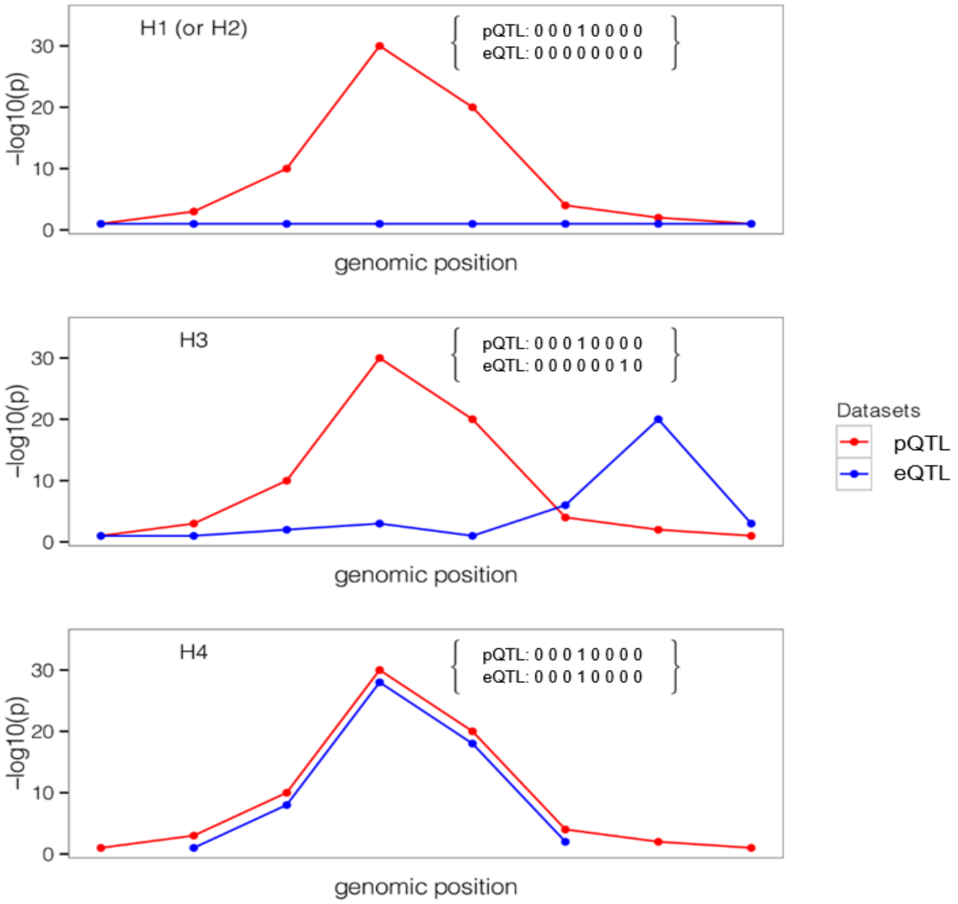

Figure 2-5. Illustrative example of colocalisation between two traits. A variant is associated with two distinct traits in GWAS, for instance gene expression (eQTL) and protein levels (pQTL). Evidence of this overlap is insufficient to conclude that the variant causally influences both traits. Therefore, formal tests of colocalisation are applied. In this example, a Bayesian test of colocalisation is used to examine whether a single variant is causal for gene expression and protein levels. Five distinct hypothesis are tested: hypothesis 0 states that no variant in a given locus is causal for either trait, hypotheses 1 and 2 state that there is a causal variant but only for either protein levels or gene expression, respectively (Panel 1), hypothesis 3 posits that there are two distinct causal variants for these traits (Panel 2) and hypothesis 4 states that one variant causally associates with both traits i.e. these traits colocalise (Panel 3). eQTL, expression quantitative trait locus; pQTL, protein quantitative trait locus. Figure adapted from Giambartolomei *et al.* (234). Copyright: PLOS.

Genetic colocalisation methods are outlined in Table 2-1. Here, I briefly summarise commonly used colocalisation methods and their differences. Coloc and other methods including HEIDI (heterogeneity in dependent instruments) and JLIM (joint likelihood mapping) assume that there is one causal variant for two traits in a locus (235, 236). The methods are satisfactory if the model is biologically plausible. Biased estimates might arise if multiple causal variants are present. A Bayesian fine-mapping and colocalisation approach termed eCAVIAR allows for multiple causal variants, but imposes an upper bound on the number of causal SNPs (typically at 6) (237). Enloc also allows for multiple variants and uses a Bayesian hierarchical model for colocalisation, fine-mapping and enrichment analyses (238). An iterative Bayesian stepwise selection procedure termed SuSiE (Sum of Single Effects regression) can provide more accurate inferences than coloc when multiple causal variants are present (239). Moloc and HyPrColoc test for the probability of a single causal variant affecting multiple traits (240, 241). The conditional regression approaches Jointsum (242) and Primo (243) permit the presence of multiple causal variants and the testing of multiple traits.

2.4.2 Considerations for colocalisation analyses

Selecting colocalisation methods requires assumptions about the underlying genetic architecture. For gene expression, allelic heterogeneity implies multi-variant models might be appropriate (244-246). A strong lead *cis*-acting pQTL might explain most of the variance in the levels of a given protein, and therefore justify a single variant model. Sun *et al.* (2018) showed that 75% of 1,478 plasma proteins have one conditionally independent pQTL (204). He *et al.* (2020) estimated the median number of pQTLs per protein is four in liver tissue (247). Furthermore, the testing of multiple traits depends on individual study designs. Colocalisation does not imply causation and assessing the direction of effect is challenging. The complementary use of MR and colocalisation methods, for instance in generalised summary-data-based MR, can provide directional evidence for an effect of one trait on another (248). These methods can aid in assessing whether mechanisms such as DNA methylation and gene expression are responsible for pQTL effects on protein levels. In the next section, I focus on the complementary use of DNA methylation for understanding the mechanisms that underlie protein regulation.

Table 2-1. Methods for assessing colocalisation between two or more traits.

Method	Year	Approach	No. of Variants	No. of Traits	Ref.
QTLMatch	2009	Proportionality of coefficients	No assumption	2	(231)[a]
Regulatory trait concordance	2010	Conditional regression	1	2	(229)[a]
Proportionality test	2012	Proportionality of coefficients	No assumption	2	(232)[a]
Sherlock	2013	Bayesian	Multiple (*cis* + *trans*)	2	(233)
Coloc	2014	Bayesian	1	2	(234)
gwas-pw	2016	Bayesian	1	2	(249)
eCAVIAR	2016	Bayesian fine-mapping	Multiple	2	(250)
HEIDI	2016	Proportionality of coefficients	1	2	(236)
Enloc	2017	Bayesian enrichment and fine-mapping	Multiple	2	(238)
JLIM	2017	Joint likelihood mapping	1	2	(235)[a]
Moloc	2018	Bayesian	1	Multiple	(240)
Simple Sum	2019	Frequentist colocalisation	Multiple	2	(251)
PICCOLO	2019	Bayesian, computes colocalisation when only top SNP is available	1	2	(252)
POEMColoc	2020	Bayesian + imputation of missing summary statistics	1	2	(253)
Primo	2020	Conditional regression	Multiple	Multiple	(243)
fastEnloc	2020	Bayesian, implemented in C++	Multiple	2	(254)
Jointsum	2020	Conditional regression	Multiple	Multiple	(242)
MRLocus	2020	Bayesian + Mendelian randomisation	Multiple	2	(255)
HyPrColoc	2021	Bayesian	1	Multiple (>100)	(241)
SuSiE	2021	Sum of single effects regression	Multiple	2	(239)

[a] These methods require individual-level genotype data. The remainder require summary-level data.

2.5 Epigenome-wide studies on protein levels

DNA methylation in promoter sequences mediates repression of transcriptional activity (Figure 2-6) (98-100). DNA methylation of the first exon in genes is also inversely correlated with protein levels (256). DNA methylation in gene bodies might activate transcription. Here, DNA methylation displaces repressive complexes or suppresses alternative intragenic or cryptic promoters in cell-specific or tissue-specific patterns (257-264). Overall, DNA methylation reduces gene expression, but it is important to consider that its relationship with protein levels is dynamic and cell- or tissue-specific.

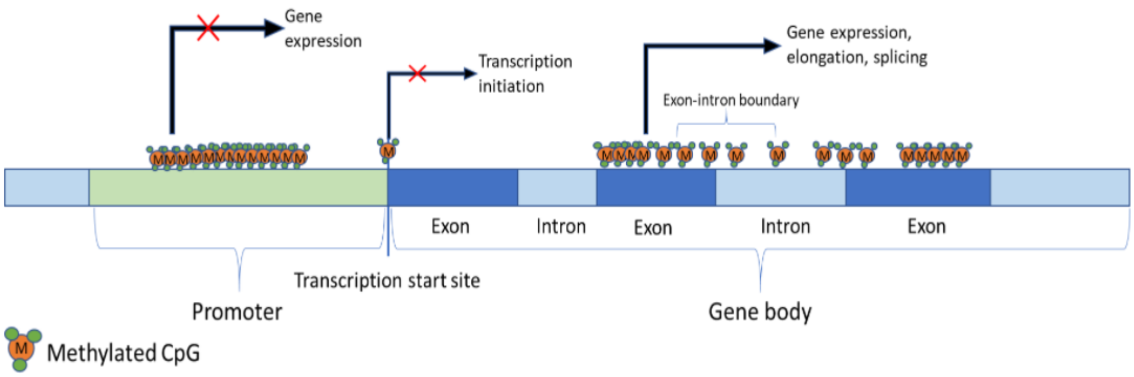

Figure 2-6. The relationship between DNA methylation and gene expression in different regions of the gene. Increased cytosine methylation in promoter sequences and in the first exon of genes associates with gene silencing. DNA methylation in gene bodies may increase transcriptional activity by regulating the inclusion and exclusion of some exons and silencing alternative, intragenic promoters. Figure taken from Aquino *et al.* (265). Copyright: LIDSEN Publishing Inc.

2.5.1 Methylation QTLs

As outlined in Section 1.6, DNA methylation profiles are influenced by environmental and genetic factors. Cross-referencing EWAS on protein levels with EWAS on lifestyle factors, such as smoking, may reveal shared methylation signatures and interrelationships between these traits. Furthermore, CpG correlates of protein levels might represent underlying genetic effects termed methylation QTLs or mQTLs, which are identified through GWAS on DNA methylation profiles (266-274).

In the largest mQTL study to date (n = 32,851), Min *et al.* (2020) conducted GWAS meta-analyses of 420,509 blood CpG sites present on the Illumina 450k array (275). Methylation QTLs were identified for 45.2% of tested CpG sites (190,102/420,509 CpG sites). There was a median of 2 independent mQTLs per CpG site (interquartile range = 4 variants). Min *et al.* showed that *cis*-acting mQTLs were enriched in active chromatin states and genic regions whereas *trans*-acting mQTLs were enriched in heterochromatin. In an mQTL study of the EPIC array (n = 1,111), Hannon *et al.* (2018) reported that 34.5% of CpG sites had one underlying mQTL whereas each mQTL was associated with a median of two CpG sites (interquartile range = 4 CpG sites) (267).

Cis-acting mQTLs are enriched in GWAS signals (273, 276, 277). Methylation QTLs have been used as instruments in MR analyses to demonstrate possibly causal associations between DNA methylation and complex traits or lifestyle factors such as smoking (275, 278-280). Colocalisation analyses have also revealed associations between mQTLs and schizophrenia (281, 282) and between mQTLs and eQTLs (i.e. between DNA methylation and gene expression) (283). Colocalisation between mQTLs, pQTLs and GWAS signals for complex diseases might identify causative molecular pathways between SNPs, CpG methylation, protein levels and health outcomes.

2.5.2 DNA methylation signatures of human traits and protein levels

The short-term variability of adult DNA methylation levels is relatively stable (284-286). Therefore, DNA methylation represents a potentially useful signature of exposure to cellular and external environmental factors. As a result, DNA methylation-based predictors have been developed for multiple health and lifestyle factors (287-291). DNA methylation-based predictors of some traits, such as cigarette smoking, provide more accurate measurements than self-reported information (292). The predictors might improve disease risk prediction and risk stratification paradigms. Methylation-based predictors require sets of CpG sites that are informative for predicting a given trait. The effect size of an association between a given CpG site and the trait in training samples provides a weight for the importance of the CpG site in estimating trait values. EWAS correlate each CpG site, in turn, with the trait of interest. Methods such as penalised regression model all CpG sites simultaneously. These methods produce parsimonious solutions that account for correlations between CpG sites (e.g. least absolute shrinkage and selection operator or LASSO regression). LASSO regression selects a subset of CpG sites as informative for predicting a given trait (293, 294). Ridge regression uses all tested CpG sites to predict a given trait (295). Elastic net regression is a commonly used intermediate of LASSO and ridge regression (296). The linear combination of CpG sites provides an estimate for a trait of interest.

EWAS and penalised regression models have been utilised to predict circulating protein levels. Inflammatory proteins such as C-reactive protein (CRP) show rapid changes in their blood levels in response to injury and infection. Serum levels of CRP can increase up to 1000-fold and resolve to baseline concentrations over 7-12 days after injury or infection (297, 298). Therefore, single time-point blood CRP measurements might not accurately reflect inflammatory profiles (299). A methylation-based predictor of CRP levels provided a more reliable signature of chronic inflammation than serum CRP across the eighth decade of life (300). The CRP predictor was based on seven CpG sites from a large meta-analysis EWAS on CRP levels (n = 8,863 in discovery sample) (301, 302). Epigenetic measures of CRP levels showed greater effect sizes than serum CRP levels in their associations with cognitive abilities, early-life mental health outcomes and structural neuroimaging phenotypes (300, 302, 303). Recently, a methylation-based predictor of interleukin 6 (IL6) levels was developed using elastic net regression (303). The epigenetic predictor of IL6 levels was based on 35 CpG sites and outperformed serum IL6 in its associations with cognitive functioning (serum IL6, n = 417, β = -0.06, SE = 0.05, P = 0.19 and epigenetic IL6, n = 7,028, β = -0.16, SE = 0.02, $P < 2 \times 10^{-16}$) (304). This approach could be extended to other blood proteins if DNA methylation processes capture a substantial proportion of the variance in their circulating levels.

In the next section, I describe an existing composite biomarker termed ‘DNAm GrimAge’, which utilises methylation-based proxies for blood protein levels to predict the risk of all-cause mortality.

2.6 Epigenetic measures of ageing

2.6.1 DNAm GrimAge

Ageing is the greatest risk factor for many age-related disease states such as dementia (305). However, individuals of the same chronological age exhibit different rates of biological ageing (306). In the past decade, several methylation-based predictors of biological ageing have been developed. These epigenetic measures of ageing are termed ‘epigenetic clocks’ (307) and an accelerated biological age as indexed by these clocks associates with mortality and multiple age-related morbidities (308-312).

In 2019, Lu *et al.* developed DNAm GrimAge by using elastic net regression models to predict mortality (313). This is in contrast with other epigenetic clocks which use chronological age as the dependent variable in elastic net models. In the first stage, Lu *et al.* created methylation-based proxies for 88 plasma protein levels and cigarette smoking pack years (n = 1,731). Pack years is calculated by multiplying the number of cigarette packs smoked per day by the number of years a person has smoked. Proxies that exhibited a correlation coefficient > 0.35 with their respective phenotypes were retained. The predictors of smoking pack years and twelve plasma protein levels met this criterion. In the second stage, an elastic net Cox regression model was used to regress time-to-death due to all-cause mortality onto chronological age, sex and methylation-based surrogates for smoking pack years and twelve plasma protein levels. The following features were selected: chronological age, sex and methylation-based surrogates for smoking pack years and seven plasma protein levels. The seven plasma proteins were: adrenomedullin

(DNAm ADM), beta-2-microglobulin (DNAm B2M), cystatin C (DNAm Cystatin C), growth differentiation factor 15 (DNAm GDF15), leptin (DNAm Leptin), plasminogen activation inhibitor 1 (DNAm PAI-1) and tissue inhibitor metalloproteinase (DNAm TIMP-1). DNAm GrimAge outperformed other epigenetic clocks in predicting mortality risk and associated with a number of peripheral, lifestyle and cardiometabolic traits. However, relationships between DNAm GrimAge and cognitive traits or neurological disease outcomes were not examined in the original study.

2.6.2 Additional epigenetic measures of ageing

Horvath Age

In 2013, Horvath developed a multi-tissue epigenetic clock termed ‘Horvath Age’, which is calculated from the linear combination of 353 CpG sites (n = 7,844 samples). This clock is independent of age-related changes in blood composition and therefore gives rise to a measure of biological ageing termed “intrinsic epigenetic age acceleration” or IEAA. Horvath Age was developed by regressing chronological age onto 21,369 CpG sites that are present on both the Illumina 27k and 450k arrays (314). Subsequently, IEAA was derived by regressing Horvath Age onto chronological age and methylation-based estimates of eight immune cell counts.

Hannum Age

Hannum Age gives rise to a measure of biological ageing termed “extrinsic epigenetic age acceleration” or EEAA as it captures age-related changes in blood cell composition (n = 482) (315). EEAA was generated in a two-step process. First, Hannum Age was derived by regressing chronological age onto CpG sites present on the 450k array. Seventy-one CpG sites were selected. A weighted average was obtained for Hannum Age and three immune cell types whose abundances change with age (naive cytotoxic T-cells, exhausted cytotoxic T-cells, and plasmablasts) (316). Second, EEAA was derived by regressing this weighted average onto chronological age.

DNAm PhenoAge

DNAm PhenoAge was developed in a two-stage design by Levine *et al.* (2018) (317). First, the hazard of mortality was regressed on 42 health-related markers from the third National Health and Nutrition Examination Survey in an elastic net regression model (n = 9,926). This gave rise to a measure of 'Phenotypic Age'. The model selected age and nine haematological and biochemical markers for predicting 'Phenotypic Age'. The markers were: albumin, alkaline phosphatase, creatinine, C-reactive protein (CRP), mean cell volume, percentage lymphocytes, red cell distribution width, serum glucose (as indexed by glycated HbA1c) and white blood cell count. Second, 'Phenotypic Age' was regressed on 20,169 CpG sites present on the 27k, 450k and EPIC arrays in an elastic net model. In total, 513 CpG sites were selected.

DNAm Telomere Length

Lu *et al.* (2019) used elastic net regression to develop a methylation-based predictor of telomere length in leukocytes (318). Telomeres are DNA-protein complexes at the end of chromosomes and shorten with age (319-321). A linear combination of 140 CpG sites was selected to predict telomere length (n = 2,256). In other epigenetic clocks, a higher age-adjusted measure of the clock is expected to correlate with poorer health outcomes. In contrast with this, a lower age-adjusted measure of telomere length is anticipated to correlate with adverse health outcomes as this reflects the age-related process of cellular telomere shortening.

DunedinPoAm – Pace of Ageing

Epigenetic clocks are 'state' measures in that they estimate how much biological ageing has occurred in an individual up to the time of measurement. An epigenetic 'speedometer' was developed to measure the pace of biological ageing in individuals. This measure is termed the Dunedin 'Pace of Ageing' (PoA) and was trained on longitudinal changes in 18 blood-chemistry and organ-system-function biomarkers at three distinct time points from age 26 to 38 years (n = 964) (322). Epigenetic clocks are also trained on individuals born within different years. This may result in confounding due to the possibility that individuals born in different years will have been exposed to different early-life environmental factors (306). To mitigate this issue, Dunedin PoA was developed using data from individuals who were all born between April 1972 and March 1973 in Dunedin, New Zealand (322). Mixed-effects growth modelling was used to calculate an individual's rate of change (slope) for each biomarker using data points from age 26, 32 and 38 years. An individual's PoA was estimated by calculating the sum of these random slopes. Belsky *et al.* (2020) employed elastic net regression models to derive a methylation-based predictor of the PoA measure (323). PoA estimates for Dunedin study participants were regressed on methylation data at age 38 years. The resultant measure of an individual's pace of biological ageing is termed 'DunedinPoAm' and is estimated from 46 unique CpG sites.

2.7 Summary

I discussed statistical methods that use data from GWAS and EWAS on protein levels to probe causal molecular pathways between plasma proteins and disease states. I also described DNA methylation signatures of complex traits and blood protein levels. Methylation-based predictors of plasma protein levels are incorporated into a leading measure of biological ageing termed DNAm GrimAge. This composite biomarker may capture important facets of cognitive decline as ageing is the greatest risk factor for cognitive ageing and dementia. In the next chapter, I conduct a literature review of existing GWAS and EWAS on blood protein levels. I also further describe associations between DNAm GrimAge and disease states, and outline the main aims of this thesis.

3 Literature review and aims

In this chapter, I perform structured literature reviews to identify existing studies that report GWAS and EWAS on blood protein levels. I describe stand-alone GWAS and EWAS, and studies that have combined genetic and epigenetic data to identify molecular determinants of blood protein levels. I outline associations between DNAm GrimAge and health outcomes before stating the aims of this thesis.

I queried MEDLINE (Ovid interface, Ovid MEDLINE in-process and other non-indexed citations and Ovid MEDLINE 1946 onwards), Embase (Ovid interface, 1980 onwards), Web of Science (core collection, Thomson Reuters) and medRxiv/bioRxiv to identify relevant articles indexed as of 14 July 2021. Search terms are outlined under each review.

3.1 GWAS on blood protein levels

I first conducted a structured literature review of existing studies that report GWAS with multiplexed blood protein levels. The following search terms or their synonyms were used to screen for relevant articles:

1. GWAS OR genome-wide association study.mp OR genetic variants.mp OR protein quantitative trait locus.mp OR pQTLs.mp OR whole-genome.mp OR genetic.mp
2. protein levels.mp OR proteomics.mp OR analyte levels.mp OR protein biomarkers.mp OR protein biomarker levels.mp OR exp plasma /
3. exp proteome / OR exp plasma protein / OR exp serum protein / OR proteins.mp
4. 1 AND 2 AND 3

Four hundred and eighty-four unique articles were identified. Thirty-one articles met inclusion criteria: (i) genome-wide association study, (ii) multiplexed protein levels, (iii) proteins were measured in blood, (iv) original research article and (v) reported methods of protein measurement. Seven further articles were identified through a supplementary manual literature search (324-330). The 38 articles included in this literature review are presented in Table 3-1. The study population, demographics, sample size, number of proteins, number of pQTL associations, sample sources and proteomic platform are reported for each study.

Of the 38 studies identified in the literature review, one study had a sample size ≤ 100, twelve had a sample size between 101 and 1,000, 22 had a sample size between 1,001 and 10,000 and three studies had a sample size ≥ 10,000 individuals. Sample sizes ranged from 96 (331) to 30,931 (332). Seven studies measured serum protein levels and 28 studies reported plasma protein levels. Three studies analysed a mixture of serum and plasma samples. The number of proteins that were assayed ranged from 16 (by multiplex immunoassay) (333) to 4,782 (SOMAscan®) (330, 334).

SOMAscan and Olink technologies are used to measure plasma protein levels in Chapters 5-7. Therefore, I describe in detail the existing GWAS that used these platforms to quantify blood protein abundances. However, studies that utilised other multiplexed assays to conduct GWAS on blood proteins are outlined in Table 3-1. The SOMAscan platform was the most frequently employed technology in GWAS on blood protein levels (13/38 studies, 32.41%). Olink was the second most commonly used platform (11/38 studies, 28.95%) (Figure 3-1).

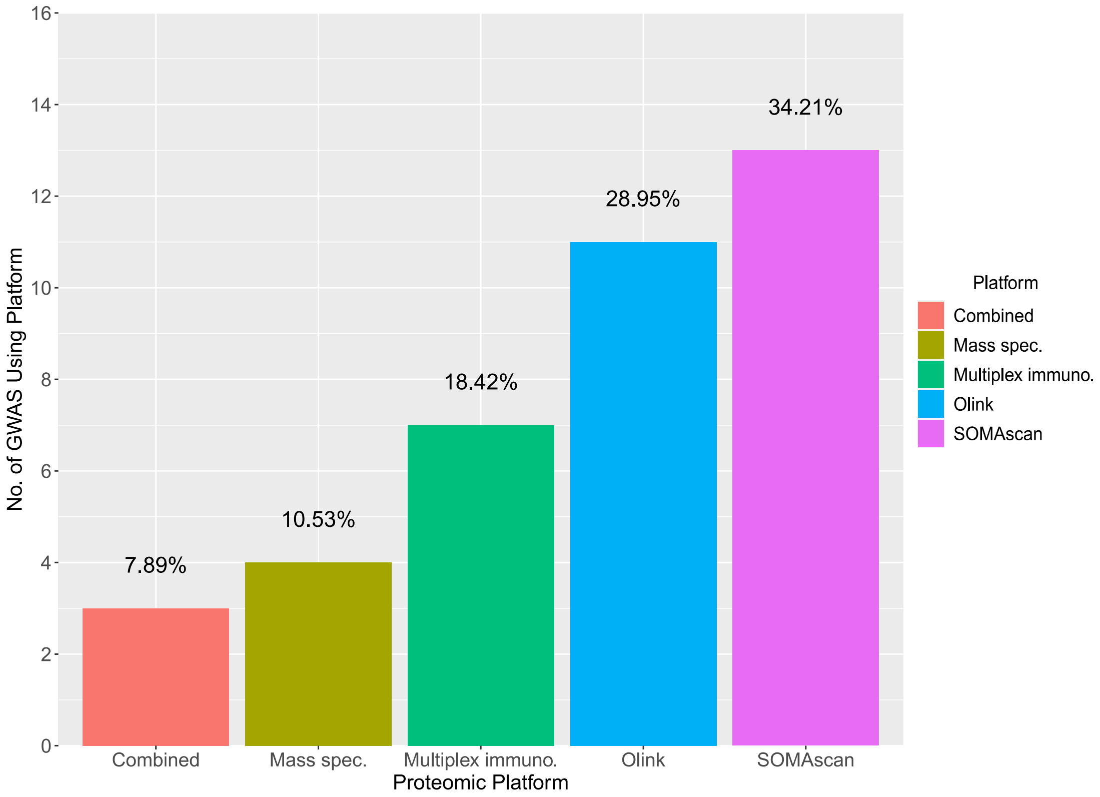

Figure 3-1. Proportion of studies reporting a given proteomic technology in GWAS on blood protein levels. Three studies reported a combination of platforms in their discovery GWAS (mass spectrometry and SOMAscan, SOMAscan and Olink, multiplex immunoassay and SOMAscan). GWAS, genome-wide association studies; Mass spec., mass spectrometry; Multiplex immuno., multiplex immunoassay.

SOMAscan

As shown in Figure 3-2, SOMAscan and Olink platforms have become increasingly popular in recent years. Lourdusamy *et al.* (2012) conducted the first GWAS on blood protein levels using SomaLogic technology (version 1, n = 778 plasma proteins). Ninety-six cognitively normal individuals from the AD-related AddNeuroMed study were included. This is the smallest sample size among GWAS on blood protein levels. The authors tested *cis* associations as they had limited power to identify *trans* effects, which typically have smaller effect sizes (331). Di Narzo *et al.* (2017) employed a newer version of SOMAscan (version 3) to measure serum levels of 1,128 proteins in two cohorts enriched for individuals with inflammatory bowel disease (335). Carayol *et al.* (2017) and Benson *et al.* (2017) used the SOMAscan platform (version 3) to measure 1,129 protein levels. Linear mixed-effects models identified 55 and 161 pQTLs using data from 494 (obese) and 2,180 individuals, respectively (336, 337). Gurinovich *et al.* (2021) applied mixed-effects models in a sample of 224 older adults (mean age = 83 years). Twenty-one pQTLs were associated with longitudinal changes in serum protein levels (338). Emilsson *et al.* (2018) measured serum levels of 4,137 proteins in 5,457 Icelanders (AGES cohort, mean age = 76.6 years). Only *cis*-regions were tested. Large networks of serum proteins were identified, which associated with complex diseases including cardiovascular and metabolic disease states (339). Gudjonsson *et al.* (2021) expanded on this work by performing genome-wide analyses in the same sample. Five proteins colocalised with AD risk, including BIN1 (Bridging integrator 1) and TREM2 (Triggering receptor expressed on myeloid cells 2). BIN1 and TREM2 are among the lead risk loci for AD (340, 341). The remaining three proteins were GLTPD2, PILRA and PLXDC2 (330). In the AGES cohort, Emilsson *et al.* (2021) regressed serum levels of 4,782 proteins on 54,469 low-frequency (MAF < 5%) and common exome-array variants. The authors detected a low-frequency *cis* pQTL in *TREM2* that associated with serum TREM2 levels. A common *trans* pQTL for TREM2 (MAF ~ 40%) was detected in the *MS4A6A* (Membrane Spanning 4-Domains A6A) locus (334).

In a seminal study, Sun *et al.* (2018) identified genetic correlates of the human plasma proteome using blood samples from 3,301 individuals in the Interval study. Linear regression models were used to examine associations between 10,572,788 imputed autosomal SNPs and 2,994 plasma protein levels. In total, 1,927 pQTLs associated with 1,478 proteins at a Bonferroni-corrected significance threshold of $P < 1.5 \times 10^{-11}$. The Olink platform was used to test for pQTL replication, and 106/163 pQTLs available for testing replicated across technologies. The authors showed that *cis* pQTLs were significantly enriched for *cis* eQTLs acting within the same gene ($P < 0.0001$). MR analyses suggested that higher matrix metalloprotease (MMP)-12 levels were associated with a lower risk of coronary artery disease ($P = 2.8 \times 10^{-13}$). Multivariate MR analyses were conducted to determine the proteins in the

IL1RL1–IL18R1 locus that might mediate atopic dermatitis risk. Higher levels of IL1RL2 and IL18R1 were negatively and positively associated with atopic dermatitis risk (P = 1.1 x 10^{-69} and 1.5 x 10^{-28}, respectively) (204). Furthermore, Suhre *et al.* (2017) identified 539 pQTLs that were associated with 284 plasma protein levels (n = 1,335 individuals). Plasma levels of 1,124 proteins were measured using version 3 of the SOMAscan platform. Protein QTLs were queried against eQTL and mQTL databases, and 179 and 122 pQTLs overlapped with known *cis* eQTLs and *cis* mQTLs, respectively. The authors also demonstrated that variation in the *SLAMF7* gene might influence individual responses to Elotuzumab, a monoclonal antibody therapy for multiple myeloma (206). These seminal studies demonstrated how pQTL data help to elucidate molecular mechanisms linking the blood proteome to complex disease states. Yang *et al.* (2021) conducted GWAS on 713, 931 and 1,079 proteins in CSF, plasma and brain tissue, respectively. One hundred and twenty-seven pQTLs were detected in plasma. Approximately 75% of *cis* pQTLs and 25% of *trans* variants in plasma were replicated in CSF or brain tissue. There was strong evidence across MR and coloc analyses for causal associations between AD risk and eight plasma proteins. The proteins were AIMP1, CD33, CTSF, EPHA5, KLK13, MICA, PDE4D and SPARCL1. AIMP1 and F11 levels were causally associated with the age-of-onset and progression of AD, respectively (342).

Pietzner *et al.* (2020) and Zhang *et al.* (2021) employed the most recent version of the SOMAscan platform (version 4) to measure 4,775 and 4,665 plasma protein levels in 10,708 and 9,084 individuals, respectively (343, 344). Pietzner *et al.* (2020) identified 220 *cis*-acting variants for 97/179 host proteins related to SARS-CoV-2 infection. Protein QTLs explained up to 70% of inter-individual variation in plasma levels of host proteins. Replication was tested using twelve Olink arrays (n = 485 individuals). Thirty-three proteins were available for cross-platform comparisons. Effect sizes for *cis*-acting variants were strongly correlated (29 SNPs, correlation coefficient r = 0.75) whereas *trans* pQTLs showed a weaker correlation (96 SNPs, r = 0.54) (343). Zhang *et al.* (2021) showed that the median heritability of protein levels explained by *cis*-acting pQTLs was similar across 7,213 European American and 1,871 African American participants from the Atherosclerosis Risk in Communities study. The respective heritability estimates were 9% and 10% for these ethnic groups.

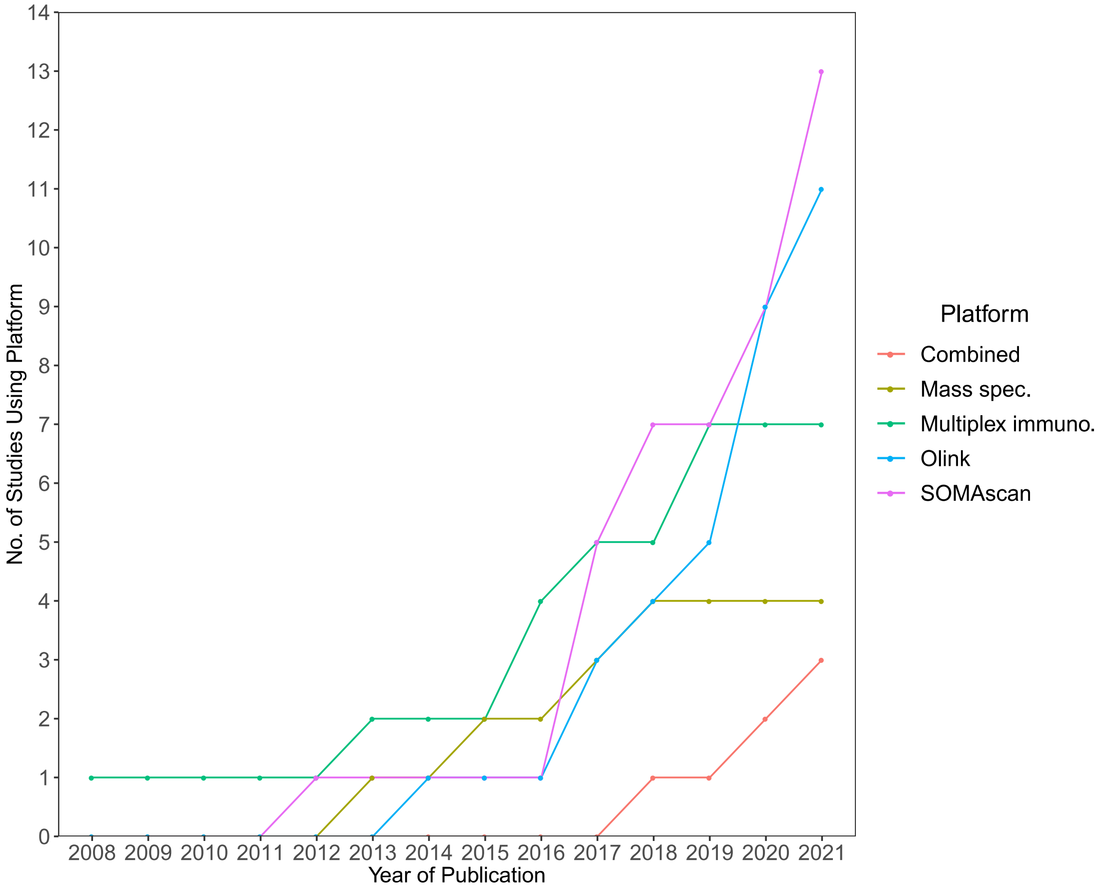

Figure 3-2. Time series data for GWAS on multiplexed blood protein levels. The first GWAS on blood levels measured with a multiplex assay was conducted in 2008. SOMAscan and Olink platforms have become increasingly popular owing to their scalability and high protein coverage. GWAS, genome-wide association studies; Mass spec., mass spectrometry; Multiplex immuno., multiplex immunoassay.

Olink

Enroth *et al.* (2014) conducted the first GWAS on blood protein levels using Olink technology. Seventy-seven plasma proteins on the Olink Oncology I panel were measured in 970 individuals. Heritability estimates for protein levels ranged from 0% (EPO) to 78% (CCL24) (324). Folkersen *et al.* (2017)

measured 83 plasma protein levels present on the Olink CVD I array in a large sample of 3,394 individuals. Eight proteins, including MMP-12 and IL6R, were causally associated with coronary artery disease risk (345). Ahsan *et al.* (2017) combined data from the Olink CVD I and Oncology I arrays and performed both GWAS and EWAS. This study is discussed in the following sections. Höglund *et al.* (2019) also measured plasma protein levels using the Olink CVD I and Oncology I arrays (n = 1,005 individuals). The authors employed whole-genome sequencing and identified 18 novel pQTLs, which were not detected with genotyped or imputed SNPs. Five variants had a MAF < 5% (346). Zhernakova *et al.* (2018) utilised non-parametric Spearman's rank correlation and measured the plasma levels of 92 proteins on the Olink CVD II array (n = 1,264 individuals). In total, 224 pQTLs were identified at FDR-adjusted P < 0.05 (329). Folkersen *et al.* (2020) measured plasma levels of 90 proteins on the Olink CVD I array in 30,931 individuals from 13 different cohorts (SCALLOP consortium). This is the largest sample size, to date, in GWAS on blood protein levels. Four hundred and fifty-one pQTLs associated with 85 plasma protein levels. MR analyses of 38 traits were performed using this large pQTL resource. The authors replicated protective associations between MMP-12 and stroke, and between IL6R and coronary artery disease. The novel associations included: higher MMP-12 levels and eczema risk, higher TRAIL-R2 levels and prostate cancer risk and higher CD40 levels and rheumatoid arthritis risk. GDF15 levels were associated with AD risk (β = 0.18, P = 4.4 x 10^{-3}), but this association did not survive multiple testing correction (332).

Two studies employed three Olink arrays in their GWAS. Gilly *et al.* (2020) measured 257 serum protein levels that were spread across the Olink CVD II, CVD III and Metabolism panels (n = 1,328 Cretan individuals). Linear mixed-effects models detected 131 pQTL associations. Polygenic risk scores based on pQTL summary statistics explained up to 45% of variation in serum protein levels (347). Bretherick *et al.* (2020) used the CVD II, CVD III and Inflammation panels. In this study, 249 plasma protein levels were measured in up to 1,992 individuals. Multiple inflammatory proteins, including IL2RA and IL4R, were causally associated with asthma. Five proteins (IL6R, FABP2, FGF5, LPL and

LTA) were causally associated with coronary artery disease risk. The levels of three proteins (SHPS1, CD40 and FCGR2B) were associated with schizophrenia risk (327).

Viñeula *et al.* (2021) measured plasma levels of 373 unique proteins present on four Olink arrays: CVD I, CVD II, Development and Metabolism (n = 3,029 individuals). In this study, 1,592 *cis* pQTLs and 533 *trans* pQTLs were identified using linear regression models (348). Zhong *et al.* (2020) quantified plasma levels of 794 proteins across eleven different Olink platforms including the Neurology and Inflammation panels. Whole-genome sequencing was used to analyse 7.3 million variants. Linear mixed-effects models revealed 144 independent pQTLs for 107 protein levels. This study had the second smallest sample size among all GWAS on blood protein levels (n = 101 individuals) (349). Recently, Zhong *et al.* (2021) expanded on this work by using a custom Olink Explorer assay. The authors increased their coverage to 1,463 plasma proteins (350). Finally, Pietzner *et al.* (2021) combined twelve Olink arrays, including the Neurology and Inflammation panels, and the SOMAscan platform (version 4). Plasma levels of 871 unique proteins were quantified in up to 10,708 individuals from the Fenland study. Sixty-four percent of all pQTL associations were shared across Olink and SOMAscan. However, pQTLs effect sizes were not strongly correlated between platforms (*cis* r = 0.41, *trans* r = 0.34) (351).

Summary

GWAS on blood protein levels are heterogeneous in terms of their sample sizes, protein coverage, proteomic platforms and statistical methods for conducing GWAS. Twenty-nine studies employed linear regression models to test for associations between genetic variants and blood protein levels. Due to LD between variants, it is necessary to perform additional methods such as LD clumping or conditional analyses following GWAS to identify independent variants. Linear models also assume that observations are independent from one another. Population stratification, relatedness among individuals, nonadditive genetic effects and unmeasured environmental confounders can bias estimates of effect sizes and reduce power in GWAS (352-355). Linear mixed-effects models are used to account for non-independence in GWAS (356). Eight studies performed linear mixed-effects GWAS to account for relatedness (6 studies) (338, 346, 347, 357-359), repeated blood sampling (1 study) (335) and population structure (1 study) (336). One study by Ruffieux *et al.* (2020) used a Bayesian approach, which accounts for correlations between markers and issues pertaining to data structure. The authors compared the performance of the Bayesian strategy against linear mixed-effects models with simulation data as a reference for ground truth. The Bayesian strategy identified more pQTLs and accounted for more variation in protein levels than linear mixed-effects models (360).

Few studies have examined panels of proteins related to dementia or neurological disease states. Colocalisation between pQTLs and eQTLs was tested in several studies. However, few analyses were conducted to investigate the relationship between pQTLs and other molecular traits, such as DNA methylation. These data may provide additional information pertaining to the molecular regulation of blood protein levels. In the following section, I perform a literature review of studies that report EWAS on blood protein levels.

Table 3-1. Summary of studies that report genome-wide association studies with multiplexed blood protein levels.

Author (date)	Study population	Mean age, % female	Sample size	No. of proteins	No. of pQTLs	Source of samples	Proteomic platform	Ref.
Melzer *et al.* (2008)	Invecchiare in Chianti	68.4, 55.2%	1,200	42	9	Serum and Plasma	Multiplex immunoassay (LINCOplex)	(361)
Lourdusamy *et al.* (2012)	AddNeuroMed	72.1, 56.2%	96	778	60	Plasma	SOMAscan (version 1)	(331)
Johansson *et al.* (2013)	Northern Sweden Population Health Study	N/A, N/A	1,060	163	5	Plasma	Mass spectrometry	(359)
Kim *et al.* (2013)	Alzheimer's Disease Neuroimaging Initiative	75, 37.6%	521	132	28	Plasma	Multiplex immunoassay (Luminex 100)	(362)
Enroth *et al.* (2014)	Northern Sweden Population Health Study	N/A, N/A	970	77	18	Plasma	Olink Oncology I array	(324)
Liu *et al.* (2015)	Twins UK	57.8, 100%	113	342	18	Plasma	Mass spectrometry	(358)
Sun *et al.* (2016)	SPIROMICS COPDGene	SPIROMICS: 63.7, 48% COPDGene: 65.6, 45%	1,340	88	527	Plasma	Multiplex immunoassay (Myriad-RBM)	(363)

Deming *et al.* (2016)	KADRC ADNI	KADRC: 73.2, 61.2% ADNI: 78.3, 37.4%	818	146	56	Plasma	Multiplex immunoassay (Myriad-RBM)	(325)
Ahola-Olli *et al.* (2017)	The Cardiovascular Risk in Young Finns Study FINRISK	Young Finns: 37, N/A FINRISK: 60, N/A	8,293	48	27	Young Finns: Serum FINRISK: Plasma	Multiplex immunoassay (Bio-rad)	(364)
Di Narzo *et al.* (2017)	PURSUIT study: 88 ulcerative colitis patients CERTIFI study: 84 moderate-to-severe Crohn's disease patients 15 healthy controls	Ulcerative colitis patients: 42, 50% Crohn's disease patients: 38, 67.9% Healthy controls: 51, 66.7%	187	1,128	41	Serum	SOMAscan (version 3)	(335)
Suhre *et al.* (2017)	Discovery: KORA Replication: QMDiab	KORA: N/A, N/A QMDiab: N/A, N/A	1,335	1,124	539	Plasma	SOMAscan (version 3)	(206)
Folkersen *et al.* (2017)	The IMPROVE study	N/A, N/A	3,394	83	79	Plasma	Olink CVD I array	(345)
de Vries *et al.* (2017)	Atherosclerosis Risk in Communities study	53.6, 60%	3,424	25 (peptides)	22	Serum	Mass spectrometry	(365)

Ahsan *et al.* (2017)	Northern Sweden Population Health Study	50.4 (median), 53%	961	121	45	Plasma	Olink Oncology I and CVD I arrays	(326)
Carayol *et al.* (2017)	The Diogenes Study	N/A, 64%	494	1,129	55	Plasma	SOMAscan (version 3)	(336)
Benson *et al.* (2017)	Discovery: Framingham Heart Study Offspring cohort Replication: Malmö Diet and Cancer Study	Framingham: 56.1, 52.6% Malmö Diet and Cancer Study: 58.6, 53.6%	2,180	1,129	161	Plasma	SOMAscan (version 3)	(337)
Sun *et al.* (2018)	Interval study	43.7, 49.7%	3,301	2,994	1,927	Plasma	SOMAscan (in-house)	(204)
Emilsson *et al.* (2018)	AGES Reykjavik study	76.6, 57.3%	5,457	4,137	3,134	Serum	SOMAscan (in-house)	(339)
Yao *et al.* (2018)	Discovery: Framingham Heart Study Replication: Interval study and KORA	Framingham: 50, 53% Interval study: 43.7, 49.7% KORA: N/A, N/A	Discovery: 6,861 Replication: 4,298	71	105	Plasma	Discovery: Multiplex immunoassay (Luminex xMAP) Replication: SOMAscan	(205)
Zhernakova *et al.* (2018)	LifeLines Dutch population cohort	45, 58%	1,264	92	214	Plasma	Olink CVD II array	(329)
Solomon *et al.* (2019)	The Tromsø Study	N/A, N/A	165	664	60	Plasma	Mass spectrometry	(357)

Sliz *et al.* (2019)	Northern Finland Birth Cohort 1966	31.1, 51.9%	5,284	16	13	Plasma	Multiplex immunoassay (Bio-rad)	(333)
Published after Chapter 5								
Höglund *et al.* (2019)	Northern Sweden Population Health Study	52 (median), 50.8%	1,005	72	5,812	Plasma	Olink Oncology I and CVD I arrays	(346)
Nath *et al.* (2019)	The Cardiovascular Risk in Young Finns Study FINRISK	Young Finns: 37.7, 54.4% FINRISK97: 47.6, 51.5% FINRISK02: 60.3, 50.1%	9,267	18	8	Young Finns: Serum FINRISK: Plasma	Multiplex immunoassay (Bio-rad)	(328)
Zhong *et al.* (2020)	SCAPIS Wellness Profiling	N/A, 52.5%	101	794	144	Plasma	Eleven Olink panels: Cardiometabolic, Cell Regulation, CVD II, CVD III, Development, Immune Response, Oncology II, Inflammation, Metabolism, Neurology and Organ Damage	(349)
Ruffieux *et al.* (2020)	The Ottawa Study The Diogenes Study	Ottawa: N/A, N/A Diogenes: N/A, N/A	1,718	1,227	136	Plasma	SOMAscan (version 3, 1,097 proteins) and mass spectrometry (130 proteins)	(360)
Pietzner *et al.* (2020)	Fenland Study	48.6, 53.4%	10,708	4,775	678	Plasma	SOMAscan (version 4)	(343)

Bretherick *et al.* (2020)	Discovery: CROATIA-Vis Replication: Orkney Complex Disease Study (ORCADES)	N/A, N/A	Up to 1,992	249	154	Plasma	Olink CVD II, CVD III and Inflammation arrays	(327)
Published after Chapter 6								
Folkersen *et al.* (2020)	SCALLOP consortium	N/A, N/A	30,931	90	451	Plasma	Olink CVD I array	(332)
Gilly *et al.* (2020)	Hellenic Isolated Cohorts MANOLIS study	N/A, N/A	1,328	257	131	Serum	Olink CVD II, CVD III, and Metabolism arrays	(347)
Zhang *et al.* (2021)*	Atherosclerosis Risk in Communities cohort	European Americans (EA): N/A, 61.8% African Americans (AA): N/A, 52.5%	EA: 7,213 AA: 1,871	4,665	EA: 1,005 AA: 1,384	Plasma	SOMAscan (version 4)	(344)
Viñuela *et al.* (2021)*	DIRECT study	N/A, N/A	3,029	373	2,125	Plasma	Four Olink arrays: CVD I, CVD II, Development and Metabolism	(348)

Pietzner *et al.* (2021)*	Fenland Study	48.6, 53.4%	Up to 1,078	871	1,923	Plasma	SOMAscan (version 4), twelve Olink arrays: Cardiometabolic, CVD II, CVD III, Cell Regulation, Development, Immune Response, Inflammation, Metabolism, Neuro Exploratory, Neurology, Oncology I and Oncology II	(351)
Zhong *et al.* (2021)	SCAPIS Wellness Profiling	N/A, 52.5%	101	321	1,463	Plasma	Olink Explorer array	(350)
Gurinovich *et al.* (2021)	New England Centenarian Study	83.0, 63%	224	4,131	21	Serum	SOMAscan (version 4)	(338)
Gudjonsson *et al.* (2021)*	AGES Reykjavik study	76.6, 57.3%	5,368	4,782	4,113	Serum	SOMAscan (in-house)	(330)
Emilsson *et al.* (2021)*	AGES Reykjavik study	76.6, 57.3%	5,343	4,782	2,019 (exome)	Serum	SOMAscan (in-house)	(334)
Yang *et al.* (2021)	Washington University cohort	70.4, 44%	529	931	127	Plasma	SOMAscan (version 1.3k)	(342)

*denotes studies that are pre-prints and not yet published. AA, African Americans; ADNI, Alzheimer's Disease Neuroimaging Initiative; CVD, cardiovascular disease; EA, European Americans; KADRC, Knight Alzheimer's Disease Research Centre; KORA, Cooperative Health Research in the Region of Augsburg; pQTL, protein quantitative trait locus; QMDiab, Qatar Metabolomics Study on Diabetes; SCAPIS, Swedish CArdioPulmonary bioImage Study.

3.2 EWAS on blood protein levels

Next, I conducted a structured literature review to identify existing EWAS with multiplexed blood protein levels. Fifty-two potentially relevant articles were identified using the following search criteria:

1. DNA methylation.mp OR EWAS.mp OR epigenome-wide association study.mp
2. protein levels.mp OR proteomics.mp OR protein biomarker levels.mp OR exp plasma proteome / OR exp plasma protein / OR exp serum protein / OR proteins.mp
3. 1 AND 2

The inclusion criteria were: (i) epigenome-wide association study, (ii) used DNA methylation, (iii) multiplexed protein levels, (iv) proteins were measured in blood tissue, (v) original research article and (vi) reported methods of protein measurement. Two studies satisfied these criteria (326, 366).

Ahsan *et al.* (2017) conducted EWAS on 121 plasma protein levels that were measured using the Olink Oncology I and CVD I arrays (n = 729 individuals). Linear mixed-effects models were applied and adjusted for age, sex, white blood cell proportions, year of sampling, batch and plate effects. One hundred and eighty-eight CpG sites associated with 44 protein levels. The authors also performed GWAS on plasma protein levels. The combined genetic and epigenetic analyses are discussed in the following section. Thirteen CpG sites were associated with multiple protein levels. These associations included differential methylation in the *NLRC5* gene (NOD-like receptor family CARD domain containing 5) and CXCL11, CXCL9, IL12, and IL18 levels. NLRC5 is an important activator of the inflammasome and regulates expression of innate immune system receptors and sensors (367). Smoking-associated CpG sites in *F2RL3* (F2R-like thrombin or trypsin receptor 3) and *AHRR* (Aryl-hydrocarbon receptor repressor) were associated with IL12 and WFDC2 levels (368, 369). Adjustment for smoking attenuated these associations to non-significance.

Zaghlool *et al.* (2020) conducted the largest EWAS on blood protein levels to date, both in terms of sample size and protein coverage. The authors measured 1,123 plasma protein levels using the SOMAscan platform (version 3.2, n = 1,288 individuals). Step-wise linear regression models were used iteratively to regress out the potential confounding effects of sex, white blood cell proportions, *cis*-acting genetic variation, age, smoking, body mass index and type 2 diabetes. The authors accounted for SNPs that might explain associations between CpG sites and protein levels. However, they did not perform both GWAS and EWAS on blood protein levels in this study. Ninety-eight CpG sites were associated with 15 unique proteins. Seventy-two CpG sites were associated with pappalysin-1 (PAPPA) levels; however, 70/72 associations were attenuated to non-significance when accounting for eosinophil counts. Seven immune-related proteins (CD48, CD163, CXCL10, CXCL11, LAG3, FCGR3B and B2M) were associated with differential methylation in the *NLRC5* locus (366).

Summary

There are only two existing EWAS with multiplexed blood protein levels. Cross-chromosome correlations between measured markers are higher for epigenetic than genetic data (118). Inter-correlations between markers, unmeasured confounders and data structure can result in poor effect size estimation, model overfitting and reduced power (127). Therefore, methods that can control for these issues in EWAS on protein levels are required. Furthermore, no study has examined the relationship between DNA methylation, blood proteins and cognitive ageing or dementia. Only one study by Ahsan *et al.* performed both GWAS and EWAS on blood protein levels, which I discuss in the following section.

3.3 Combined GWAS and EWAS on blood protein levels

Ahsan *et al.* (2017) identified 45 pQTLs for 39/121 protein levels. Eighteen biomarkers harboured SNP and CpG associations. For 11/18 proteins, GWAS and EWAS signals were located within two megabases of one another. SNPs for these protein were included as covariates in EWAS for the same protein. All CpG associations were attenuated to non-significance when adjusting for underlying genetic variants. Heritability estimates ranged from 0% (for 6 proteins) to 67% for CCL24 levels. CCL24 also had the highest heritability estimate (78%) in Enroth *et al.* (324). Both studies employed Olink technology and their sample sizes were comparable. In the study by Ahsan *et al.*, pQTLs largely accounted for the contribution of DNA methylation variability towards inter-individual differences in protein levels. The authors concluded that genetic and environmental factors influenced the associations between DNA methylation and blood protein levels (326).

3.4 DNAm GrimAge and health outcomes

In their initial description of DNAm GrimAge, Lu *et al.* (2019) demonstrated that age-adjusted DNAm GrimAge, or AgeAccelGrim, was a strong predictor of all-cause mortality (hazard ratio (HR) = 1.10, P = 2.0 x 10^{-75}). AgeAccelGrim also predicted time-to-onset of coronary artery disease, congestive heart failure and cancer (range of HR = [1.08, 1.10], range of P = [4.9 x 10^{-9}, 6.2 x 10^{-24}]). AgeAccelGrim was associated with the prevalence of type 2 diabetes (odds ratio (OR) = 1.02, P = 0.01) and hypertension (OR = 1.04, P = 5.1 x 10^{-13}) at baseline. Associations between AgeAccelGrim and cognitive ageing or dementia were not tested.

In Chapter 8, I provide the first external replication of the association between AgeAccelGrim and all-cause mortality. It is also the first study after Lu *et al*. to study associations between DNAm GrimAge and health outcomes. In Chapter 9, I conduct the largest study on DNAm GrimAge, both in terms of sample size and the number of health outcomes examined. Outside of these studies, AgeAccelGrim has been associated with post-traumatic stress disorder (370-372), depression (373, 374), schizophrenia (375), cancer (376-378), cardiovascular disease (379) and Huntington’s disease (380). I discuss the relationships between my findings and those in the literature in Chapters 8-10.

3.5 Thesis aims

There are no combined GWAS and EWAS on blood proteins that associate with neurological phenotypes. Furthermore, we have not characterised causal relationships between many blood proteins and different types of dementia, including AD. Therefore, the respective aims of Chapters 5, 6 and 7 are:

Aim 1: To perform GWAS and EWAS on plasma levels of 92 neurology-related proteins in 750 healthy older adults (Olink Neurology panel, the Lothian Birth Cohort 1936).

Aim 2: To perform an integrated GWAS and EWAS on plasma levels of 70 inflammatory proteins in 876 healthy older adults (Olink Inflammation panel, the Lothian Birth Cohort 1936).

Aim 3: To perform an integrated GWAS and EWAS on plasma levels of 282 proteins that associate with AD in the literature (n ≤ 1,064, SOMAscan, Generation Scotland).

I perform GWAS and EWAS on 422 unique proteins. Forty-eight proteins overlap with those in Ahsan *et al*. In Chapters 5-7, I use MR to assess whether plasma levels of 54 distinct proteins are causally associated with AD risk. One protein (poliovirus receptor, PVR) is tested in a hypothesis-driven approach (Chapter 5). The remaining 53 proteins are tested agnostically and represent the proteins that harbour significant pQTL associations in Chapters 6 and 7.

No study has investigated the association between an accelerated DNAm GrimAge and cognitive ageing. Furthermore, no study has tested associations between DNAm GrimAge and the incidence of AD or other common disease states in a single sample. Therefore, the respective aims of Chapters 8 and 9 are:

Aim 4: To test associations between an accelerated DNAm GrimAge and measures of brain health and cognitive decline in older age (n = 709, the Lothian Birth Cohort 1936, mean age = 73 years).

Aim 5: To examine associations between six epigenetic measures of ageing, including DNAm GrimAge, and the prevalence and incidence of AD and nine other disease states (n ≤ 9,537, Generation Scotland).

3.6 Summary

I outlined the major aims of this thesis and their novelty with respect to existing GWAS and EWAS on blood protein levels and studies on DNAm GrimAge. In the following chapter, I describe the two cohort studies (the Lothian Birth Cohort 1936 and Generation Scotland) and the major methodologies used throughout Chapters 5-9.

4 Study cohorts and methods

In this chapter, I provide an overview of the original protocols for the two cohort studies used in this thesis: the Lothian Birth Cohort 1936 and Generation Scotland: the Scottish Family Health Study. I describe the measurement of genetics, DNA methylation and proteomics data within these cohorts. I outline the rationale for the main methods used in my empirical work.

4.1 The Lothian Birth Cohort 1936

The Lothian Birth Cohort 1936 (LBC1936) is a longitudinal study of ageing. The sample consists of surviving individuals who partook in the Scottish Mental Survey 1947. On June 4th 1947, 70,805 out of all 75,211 Scottish children born in 1936 undertook a cognitive test named the Moray House Test No. 12 (381). The Lothian Health Board identified potential participants who were born in 1936 and were registered with a general practitioner within the Lothian area. Between June 2004 and November 2006, 3,686 individuals were invited to take part in the LBC1936 study. In total, 1,091 participants were enrolled in the study and completed baseline testing. These individuals had also taken the Moray House Test at age 11 years. Cognitive tests, physical tests and lifestyle questionnaires were administered by a trained psychologist and a research nurse at the Wellcome Clinical Research Facility, Western General Hospital, Edinburgh. Blood draws were also taken. Participants have completed up to five waves of data collection at mean ages of 70, 73, 76, 79 and 82 years. The number of individuals in each wave was: 1,091 (Wave 1), 866 (Wave 2), 697 (Wave 3), 550 (Wave 4) and 440 (Wave 5) (382). The original protocol and follow-up testing are further described by Deary *et al.* (2007) (381), Deary *et al.* (2011) (383) and Taylor *et al.* (2018) (382).

4.1.1 Ethics and funding

Ethical approval for the LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian

Research Ethics committee (LREC/1998/4/183 and LREC/2003/2/29). All participants provided written informed consent.

Phenotypic data collection in the LBC1936 was supported by Age UK (The Disconnected Mind project) and the Medical Research Council (MR/M01311/1). Genome-wide genotyping was funded by the Biotechnology and Biological Sciences Research Council. DNA methylation typing was supported by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, the Wellcome Institutional Strategic Support Fund, The University of Edinburgh and The University of Queensland. Proteomics analyses were supported by a National Institutes of Health grant (R01AG054628).

4.1.2 Genetic data in the Lothian Birth Cohort 1936

Quality control steps were performed by Dr Lorna Houlihan. Genomic DNA was isolated from whole blood samples at the Wellcome Clinical Research Facility, Edinburgh (n = 1,071). Twenty-nine blood samples failed quality control procedures before genotyping. SNP genotyping was performed using the Illumina Human610-Quadv1 chip (n = 1,042 samples, San Diego, USA). Individuals were removed if they showed a disagreement between reported and predicted sex (n = 12) or evidence of non-European ancestry following MDS of genotype data (n = 1). Samples with a call rate ≤ 0.95 were removed (n = 16). One member from each pair of related individuals was also removed (proportion of identity by descent > 0.25, n = 8). SNPs were retained if they had a call rate ≥ 0.98, MAF ≥ 0.01 and Hardy-Weinberg equilibrium test with P ≥ 0.001. Genotype data were available for 542,050 SNPs and 1,005 individuals at Wave 1 following quality control procedures (384).

4.1.3 DNA methylation data in the Lothian Birth Cohort 1936

DNA methylation was measured using Illumina Infinium HumanMethylation 450K BeadChips in the LBC1936 at Waves 1, 2, 3 and 4. LBC1936 samples were processed in three separate batches on 13 dates using 41 plates and 309 microarrays. Quality control procedures were performed by Dr Sonia

Shah, Dr Allan McRae and Dr Qian Zhang (385). Raw intensity data were background-corrected and normalised using internal controls. Methylation beta values were generated using the *minfi* package in R (386). Probes that had a low detection rate (<95%) at P < 0.01 were removed. Samples with issues pertaining to hybridisation, nucleotide extension, staining signal and bisulfite conversion were removed following manual inspection of array control probes. Samples that had <450,000 probes detected at P < 0.01 were excluded. Individuals with a disagreement between predicted sex based on XY probes and reported sex were removed. Following quality control procedures, 450,726 autosomal probes were available for 895 individuals at Wave 1.

4.1.4 Proteomics data in the Lothian Birth Cohort 1936

Two Olink panels were used to measure plasma protein levels in members of the LBC1936: the Neurology and Inflammation panels. The Olink Neurology panel was used at Wave 2. This time-point was selected for the measurement of neurological protein biomarkers as brain magnetic resonance imaging (MRI) data were also available at Wave 2 (mean age = 73 years), but not at Wave 1. The Olink Inflammation panel was used for Wave 1 samples (mean age = 70 years). This time-point was selected to allow for the longest possible longitudinal analyses between proteins and physical, cognitive or disease phenotypes. The Olink Neurology panel consists of 92 proteins with established links to neuropathology (such as tau) and exploratory proteins with roles in cell-cell communication, development and immunology. The Olink Inflammation panel represents 92 blood proteins that are associated with inflammatory disease states and related biological processes. At Wave 2 (Neurology panel), plasma was collected in citrate tubes (n = 816, mean age ± standard deviation (SD) = 72.5 ± 0.5). At Wave 1 (Inflammation panel), plasma was extracted in lithium heparin tubes at mean age 69.8 ± 0.8 years (n = 1,047). Olink Bioscience performed in-house quality control procedures. One protein (BDNF) on the Inflammation panel was removed. Furthermore, 30 samples in the Inflammation set were excluded. Additional quality control procedures were performed by Dr Sarah Harris. For 21 proteins on the Inflammation panel, over 40% of samples fell below the lowest limit of detection. These proteins were removed and left 70 inflammatory proteins available for analyses at Wave 1 (n = 1,017 individuals). Ninety-two neurology-related proteins were available for analyses at Wave 2 (n = 816).

In both panels, one microlitre of sample was incubated with proximity antibody pairs. DNA tails combine to form an amplicon through proximity extension when pairs of antibodies bind to their cognate antigens. The abundances of amplicons, or the proteins to which they bind, can be quantified by high-throughput real-time polymerase chain reaction (191). Data were pre-processed by Olink using NPX Manager software. I then applied rank-based inverse normalisation on plasma protein levels to minimise potential false positive results caused by outlying values. Following normalisation procedures, I regressed plasma protein levels onto age, sex, Olink array plate and four genetic principal components of ancestry derived from MDS to control for population structure.

4.2 Generation Scotland: the Scottish Family Health Study

Generation Scotland: the Scottish Family Health Study (GS) is a large family-based genetic epidemiology study that includes approximately 24,000 volunteers from across Scotland. Participants were recruited using Community Health Index numbers, facilitated by Scottish Practices and Professionals Involved in Research. Between 2006 and 2010, recruitment took place in the Glasgow and Tayside regions of Scotland. Individuals were eligible for the study if they were aged between 35 and 65 years, had at least one first-degree relative aged 18 years old or over and at least one full sibling group. The family members of these individuals (‘probands’) were also invited to take part (387). In 2010, recruitment was extended to Ayrshire, Arran and the Northeast region of Scotland. The age range of probands was broadened to 18-65 years. Across all centres, 126,000 potential probands were invited to take part in the study and 6,665 individuals attended an initial appointment. A further 1,288 individuals self-volunteered without invitation. There were 16,007 family members of the probands who were enrolled in the study, resulting in a total of 23,960 participants at baseline. There were 5,573 families with a mean size of 4.05 members, and there were 1,400 individuals without any relatives in the final sample. The median age was 47 years and the sample was 59% female. In 2011, the sample size was updated to 24,084 (388). Full details on the original protocol for this study are available in publications by Smith *et al.* (2006) (387) and Smith *et al.* (2013) (389).

In 2015, a strategic award from Wellcome provided funding for 'STRADL: Stratifying Resilience and Depression Longitudinally'. The STRADL cohort consists of GS participants who were re-contacted for further assessments of mental health, specifically depression. Participants were eligible for recruitment to STRADL if they provided informed consent for re-contact and had Community Health Index numbers. Of the 24,084 GS participants, 21,525 were eligible for re-contact. There were 9,618 positive respondents. The STRADL cohort consists of 2,460 families (n = 7,158 individuals) and 2,460 unrelated individuals. The mean age at baseline was 53 years and the sample was 52% female. Participants completed self-report health and lifestyle questionnaires at GS baseline (2006-2011) and STRADL baseline (2015-2016). The participants contributed blood samples, which allowed for genome-wide genotyping and the measurement of DNA methylation and proteomics data. Ninety-eight percent of study participants consented to follow-up electronic health record linkage. Further details on baseline descriptions of the STRADL cohort are outlined by Navrady *et al.* (2018) (388).

4.2.1 Ethics and funding

All components of GS and STRADL received ethical approval from the National Health Service Scotland Tayside Committee on Medical Research Ethics (05/S1401/89 and 14/SS/0039). Research Tissue Bank status was granted by the Tayside Committee on Medical Research Ethics (20-ES-0021).

GS is supported by core funding from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). STRADL is supported by a Wellcome Strategic Award (104036/Z/14/Z). Genotyping was funded through the Medical Research Council and the Wellcome Strategic Award (104036/Z/14/Z). Funding from the

Medical Research Council and Wellcome also supported DNA methylation typing (104036/Z/14/Z), in addition to support from the Brain and Behaviour Research Foundation (27404). Proteomic analyses in STRADL were supported by Dementias Platform UK and were funded through core grant support from the Medical Research Council (MR/L023784/2).

4.2.2 Genetic data in Generation Scotland

Blood samples were collected, processed and stored using standard operating procedures at the Wellcome Clinical Research Facility, Western General Hospital, Edinburgh. DNA was quantified using the Invitrogen PicoGreen assay kit and diluted to 50 nanograms per microliter. Four microlitres of sample were used in genotyping (390). The first 9,863 samples were genotyped using the Illumina HumanOmniExpressExome-8 v1.0 BeadChip. The remaining samples were genotyped using the Illumina HumanOmniExpressExome-8 v1.2 BeadChip. Quality control was carried out in PLINK v1.9b2c by Dr Saskia Hagenaars (391, 392). SNPs with a call rate < 0.98, MAF ≤ 0.01 and Hardy-Weinberg equilibrium test with $P \leq 1 \times 10^{-6}$ were removed. There was a total of 561,125 autosomal SNPs that passed quality control. Duplicate samples were removed. Samples were excluded if they had a genotype call rate < 0.98. PCA of genotype data was performed to identify potential outliers. GS genotypes were combined with data from 1,092 individuals in the 1000 Genomes Project before PCA (393). These analyses were initially performed by Dr Carmen Amador. Outliers who were more than six standard deviations away from the mean of the first two principal components were removed (394). Following quality control there were 19,904 individuals with genotype data, consisting of 11,731 females and 8,173 males (395, 396).

4.2.3 DNA methylation data in Generation Scotland

Whole blood genomic DNA samples were normalised to 50 nanograms per microlitre and were treated with sodium bisulfite using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, California), following the manufacturer’s instructions. DNA methylation was profiled using the Infinium MethylationEPIC BeadChip. Blood samples drawn at GS baseline were processed at two separate time-points, which are designated as Set 1 and Set 2. There were 5,200 individuals in Set 1 and 4,585 individuals in Set 2. Methylation typing was performed in 31 batches within each set. Quality control procedures were conducted by Dr Rosie Walker, Dr Mairéad Bermingham and Mr Stewart Morris.

In Set 1, the *ShinyMethyl* package in R was used to compare plots of log median signal intensities across methylated and unmethylated beads in each array (397). Outliers were removed based on a visual inspection of the plots. The *wateRmelon* package in R was used to remove samples based on the following exclusion criteria: (i) >1% of probes had a detection P value > 0.05, (ii) probes had a beadcount <3 in >5% of samples and (iii) probes were non-autosomal i.e. XY probes (398). Eighty samples and 5,910 probes were excluded based on these criteria. Probes that were predicted to have off-target effects were excluded. Probes were also removed if they contained a SNP in the final five 3' bases or at the site of single-base extension in type I probes (n = 84,352) (399, 400). Twelve individuals were removed due to discordance between their methylation-based predicted sex and recorded sex. Seven further samples were removed as they were identified as genetic outliers in PCA of genotype data (Dr Carmen Amador) (394). Ten samples that were derived from saliva were excluded, along with three individuals that self-reported 'Yes' to all health conditions listed on study questionnaires at baseline. One individual was removed as their methylation data indicated that they might have an XXY genotype. Data were normalised using the dasen method in the *wateRmelon* package. The final set consisted of 5,087 participants and 760,943 loci (401). In Set 1 there are 2,578 unrelated individuals.

In Set 2, the *Meffil* package in R was used to perform initial quality control steps (402). Samples were excluded if they met the following criteria: (i) there was a mismatch between self-reported and methylation-based predicted sex, (ii) more than 1% of probes had a detection P value > 0.05, (iii) samples showed evidence of dye bias, (iv) sample were outliers at bisulfite conversion control probes and (v) the sample had a median methylated signal intensity that was ≥ 3 standard deviations lower than expected. *ShinyMethyl* was then used to exclude probes based on the same criteria as those applied in Set 1. MDS of methylation data was performed and outliers were removed based on visual inspection of the resultant plots. *Meffil* was employed again to remove poor-performing probes that met the following exclusion criteria: (i) probes had a beadcount of <3 in >5% of samples and (ii) >5% of samples had a detection P value > 0.05. In total, 8,878 poor-performing probes and 135 samples were excluded. Probes with potential off-target effects were excluded as in Set 1. Probes were also removed if they had a SNP in the final five 3' bases or in the single-base extension site (n = 84,352) (399, 400). Data were normalised using the dasen method. There were 4,450 individuals in Set 2 and 758,332 CpG sites following quality control procedures. The individuals in Set 2 are unrelated to each other and to those in Set 1.

4.2.4 Proteomics data in Generation Scotland

Plasma samples from 1,065 STRADL participants were analysed using the 5k SOMAscan platform (version 4). Plasma samples were collected in 150 microlitre aliquots at baseline (2015-2016). The samples were stored in ethylenediaminetetraacetic acid-treated tubes at -80°C. SOMAmer levels were measured using six scanners and twelve 96-well plates (70-85 samples per plate) between February 18th 2020 and March 4th 2020.

There are 5,284 SOMAmer reagents in the 5k SOMAscan platform. Seven SOMAmers are deprecated, twelve are spike-in controls and 286 are negative controls/non-human targets. There are 4,979 SOMAmers that target 4,776 unique human proteins. The 4,979 SOMAmers for human proteins are spread across three dilution bits accordingly: 160 in the 0.005% bin, 797 in the 0.5% dilution group and 4,022 in the 20% bin. In the 96-well plates there are eleven wells that are dedicated to replicate controls (five calibrator samples, three quality control samples and three buffer or no protein samples) and 85 that are reserved for biological samples. The recognition signal between SOMAmers and protein targets is measured using relative fluorescence units (RFUs) (195). The following in-house quality control criteria were applied by SomaLogic (403):

- **Hybridisation control normalisation** is performed to reduce nuisance variance within sample wells. A scaling factor is applied to each well. The scaling factor is the median ratio of reference spike-in RFUs against observed spike-in RFUs in the well. The reference RFU for a spike-in control is the median RFU of the control SOMAmer across the entire plate.
- **Intra-plate median signal normalisation** is carried out to minimise technical variability across sample wells in a plate. This is applied separately to wells of the same class (i.e. separately for each buffer, calibrator and quality control type) and within SOMAmers of the same dilution factor (0.005%, 0.5%, and 20%). This creates a number of sample-SOMAmer groupings. (1) The RFU of each SOMAmer (within a particular group) is divided by the median RFU of the SOMAmer across a plate. (2) The scaling factor for each well is calculated as the inverse of the median ratio from (1) across all SOMAmers in the well that are in a given group.
- **Calibration Normalisation** accounts for inter-plate variability that is introduced chiefly by differences in scanner intensity. RFUs for dedicated calibrator samples in a plate are each divided by a reference value. The median of this ratio in a given plate is used to calculate a single scaling factor for the plate.
- **Calibration** accounts for variability between experiments and is performed on a SOMAmer-by-SOMAmer basis. A SOMAmer-specific reference value is divided by the median of calibrator control RFUs across a given experiment (or run in a scanner). This produces a calibration scaling factor for a given SOMAmer and is applied to every plate in the scanner.
- **Adaptive Normalisation by Maximum Likelihood** is an optional quality control step in SOMAscan experiments. This procedure was applied in the STRADL sample. Median signals and median absolute deviations for each SOMAmer are taken from an external reference sample (n ~ 1,000). A scaling factor is calculated for each SOMAmer, which maximises the probability that a sample's RFU has come from the sampling distribution. It is based on the assumption that more than

30% of SOMAmers for a given sample lie within the reference distributions. This step reduces technical variability and inter-sample biological variability that may contribute to differences in total protein signal. It might not be suitable for case versus control study designs.

- **Post-calibration quality control** is carried out to ensure that quality control procedures have been appropriately applied. Three pooled quality control replicates are randomly distributed on a given 96-well plate. The accuracy of the median replicate signal on a given plate is compared against a reference value. This results in a vector of accuracy ratios across all SOMAmers on the 96-well plate. At least 85% of accuracy ratios must be between 0.8 and 1.2 in a plate prior to release. A plate will also not be released if any scaling factors were below 0.4 or above 2.5.

All samples met quality control criteria in the STRADL sample. In total, 4,235 SOMAmers for human targets passed quality control and were brought forward for analyses. I first log-transformed RFUs and regressed them onto age, sex, study site (Aberdeen or Dundee), the duration between a sample being collected and processed for proteomics analyses (factor, four levels) and 20 genetic principal components of ancestry to control for population structure. I applied rank-based inverse normalisation to residualised RFUs. I refer to these normalised RFUs as protein levels. However, in both the LBC1936 and STRADL, they reflect signal intensities that have undergone a number of quality control and pre-correction steps.

4.3 Overview of key methods

Detailed methods sections are provided within Chapters 5-9. Here, I provide a brief rationale for the methods used in molecular association studies throughout Chapters 5, 6 and 7. I also describe additional DNA methylation quality control procedures that are performed in Chapters 8 and 9 prior to the calculation of epigenetic ageing measures.

In Chapter 5, GWAS are performed on plasma levels of 92 Olink neurology proteins. An additive genetic model is assumed. Imputation is performed using the 1000 Genomes reference panel (phase 1, version 3) and an INFO score ≥ 0.6. Linear regression models are used to test for associations between 8,683,751 autosomal variants and each protein in mach2qtl (n = 750 unrelated individuals in the LBC1936, mean age = 73 years) (404, 405). EWAS are carried out using multivariable linear regression models and the R package *limma* (406).

As outlined in Section 1.5.3, confounders in EWAS can lead to biased test statistics and induce correlations between probes on different chromosomes (127). This issue is more pronounced in EWAS than in GWAS analyses (118). Therefore, I also use linear mixed-effects models to perform EWAS in Chapter 5 and to account for possible biases induced by data structure and unknown confounders. Association tests are performed using the MOMENT method (multi-component mixed-linear-model-based omic association excluding the target) in OSCA software (132). In this method, probes are split into two groups based on initial linear regression analyses between CpG sites and the trait of interest. Probes that surpass an epigenome-wide significance threshold are placed into the first group ('strongly-associated probes'). 'Weakly-associated probes' that do not surpass this threshold are considered in the second group. The groups are fitted as separate random-effects terms. This approach allows for the modelling of CpG associations that have effect sizes of varying magnitudes. In each association test, all probes that are >50 kilobases from the probe of interest are fitted as random-effect components. The exclusion window of 50 kilobases is selected to omit probes that show high correlations with the probe of interest, and therefore might bias effect size estimates. The MOMENT method accounts for the effects of known and unknown confounders and intercorrelations between distal probes that confounders may induce. This method was more robust in controlling family-wise error rates and false positive rates than linear regression methods, latent factor linear mixed-effects models and other preceding linear mixed-effects models using simulation data (132).

In Chapter 6, I conduct the first integrated GWAS and EWAS on blood protein levels using Bayesian penalised regression (BayesR+, n = 876, LBC1936)

(133). I also perform sensitivity analyses using the linear regression and mixed-effects modelling approaches outlined above. BayesR+ assumes a prior mixture of Gaussian distributions that correspond to markers with effect sizes of different magnitudes (i.e. small, medium or large effects). The prior distribution also includes a discrete spike at zero that allows for the exclusion of markers with non-identifiable effects on phenotypes. Marker effects are estimated after adjusting for all other measured omics probes and structural influences such as batch effects and population stratification. Simulated DNAm data were used to compare the performances of BayesR+, single-probe linear regression and penalised regression approaches in estimating true marker effect sizes. BayesR+ showed higher correlations between simulated and estimated marker effect sizes and lower mean squared errors when compared to other methods. Methylation data were simulated to have significant confounding due to differential blood cell type proportions. There was no correlation between the first principal component of methylation data and the difference between simulated and marker effect sizes. This suggested that BayesR+ appropriately accounted for cell type heterogeneity and intercorrelations between probes induced by confounders. BayesR+ exhibited lower mean squared errors than linear mixed-effects models when estimating the variance in complex traits accounted for by DNA methylation (133). In Chapter 6, I perform GWAS on plasma protein levels using BayesR+ and linear regression models. I use BayesR+, limma and OSCA-MOMENT to conduct EWAS on plasma protein levels. I assess the replication of pQTL and CpG associations across these methods. Furthermore, I utilise BayesR+ to estimate the proportion of variability in blood protein levels accounted for by genome-wide genotype and DNAm data. I quantify genetic, epigenetic, and combined variance component estimates for protein levels.

In Chapter 7, I use BayesR+ to perform integrated GWAS and EWAS on 282 AD-associated plasma proteins (n ≤ 1,064, GS). GS is a family-based study and therefore relatedness between samples could bias effect size estimates and reduce study power. I use samples from the LBC1936 in Chapters 5 and 6 and this cohort does not contain closely-related individuals. Therefore, methods including linear regression may be more appropriate for the LBC1936 than GS. In Chapter 7, I use BayesR+ as it implicitly accounts for data structure and relatedness (133). I also apply linear mixed-effects models as sensitivity EWAS analyses. The lmekin function in the R package *coxme* is used to perform linear mixed-effects models with a kinship matrix fitted to account for relatedness (407). I quantify independent and combined contributions of genetic and epigenetic data towards inter-individual variability in blood protein levels. In each chapter, I perform colocalisation analyses using the *coloc* package in R (234). I conduct MR analyses using the *TwoSampleMR* package in R (408). Analyses downstream from GWAS and EWAS are detailed in full within Chapters 5, 6 and 7.

In Chapter 8, DNAm GrimAge estimates for 906 LBC1936 participants are derived using Illumina 450k array data and the online Horvath age calculator (https://dnamage.genetics.ucla.edu/) (314). In Chapter 9, estimates for Horvath Age, Hannum Age, DNAm PhenoAge, DNAm GrimAge and DNAm Telomere Length are also obtained using the online calculator. EPIC array data are available for 9,537 individuals in GS. I apply an addition quality control procedure for DNAm data in the LBC1936 and GS to calculate measures of epigenetic ageing. Quality control methods are performed to eliminate missing CpG values as recommended by Horvath (314). Data are normalised using the preprocessNoob function in *minfi*, which performs normal-exponential convolution using out-of-band probes or ‘noob’ (386). Infinium Type I probes use the same colour channel to measure signals from methylated and unmethylated beads. The unused colour channel (i.e. red or green) serves as the out-of-band channel. Out-of-band intensities from designated control probes are used to measure non-specific background fluorescence (409). In noob, the background fluorescence that is estimated from out-of-band probes is removed from each sample. Dye-bias correction is then performed for background-corrected data. Noob-normalised methylation beta values are obtained using the getBeta function in *minfi* (386). Noob-normalised beta values were submitted to the online age calculator to obtain epigenetic measures of ageing. In Chapter 9, DunedinPoAm is calculated using the R package *DunedinPoAm38* and dasen-normalised beta values for CpG sites on the EPIC array (323). Association tests between epigenetic measures of ageing and phenotypes are detailed within Chapters 8 and 9.

4.4 Summary

In this chapter, I outlined the measurement of genetics, DNA methylation and blood protein data in the two cohort studies that are used in this thesis. These cohort studies are the LBC1936 and GS. I provided a rationale for the methods employed in Chapters 5-9. In the next chapter, I detail my first study in which I aim to identify genetic and epigenetic factors that associate with plasma levels of 92 neurology-related proteins in 750 LBC1936 participants.

5 Genome-wide and epigenome-wide studies on neurology-related proteins

5.1 Introduction

Blood-based biomarkers are useful for neurological disease states as access to *in vivo* neural tissue is limited to necessary surgical resections. I outline in Chapter 2 that conducting GWAS on blood protein levels allows us to apply causal modelling methods and evaluate relationships between protein biomarkers and disease states. Protein QTLs can be related to existing databases of disease-associated variants to accelerate the discovery of biomarkers and drug therapies (410). Further, polygenic risk scores based on pQTL data could stratify patients into subpopulations that are likely to respond positively and safely to certain drug therapies. EWAS on protein levels reveal genetic and lifestyle factors that influence the proteome, potentially refining patient stratification algorithms and our understanding of disease biology.

When this study was conducted, there were no GWAS or EWAS that focused on a panel of proteins that relate to brain health. Only one study performed GWAS and EWAS in a single sample on CVD- and cancer-associated proteins. In this chapter, I carry out GWAS and EWAS on plasma levels of 92 neurology-related proteins present on the Olink Neurology panel (n = 750, LBC1936). I apply colocalisation analyses to investigate the molecular mechanisms through which pQTLs influence blood protein levels. The cell-surface protein poliovirus receptor is encoded by the *PVR* gene within the AD-associated *TOMM40-APOE-APOC2* cluster on chromosome 19. SNPs within *PVR* associate with family history of AD (85). However, the causal relationship between PVR and AD risk is unknown. Therefore, I also use two-sample MR to evaluate the relationship between plasma PVR levels and AD risk.

This study was published in *Nature Communications* (411) in July 2019 and is included in full in Section 5.2. Supplementary material for Chapters 5-9 are available in the electronic file attached to this thesis or in the following repository: https://github.com/robertfhillary/s1777309_Supplementary_Material.

5.2 Genome- and epigenome-wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-11177-x OPEN

Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936

Robert F. Hillary[1], Daniel L. McCartney[1], Sarah E. Harris[2,3], Anna J. Stevenson[1], Anne Seeboth[1], Qian Zhang[4], David C. Liewald[2], Kathryn L. Evans[1,2], Craig W. Ritchie[5], Elliot M. Tucker-Drob[6,7], Naomi R. Wray[4], Allan F. McRae[4], Peter M. Visscher[4], Ian J. Deary[2,3] & Riccardo E. Marioni[1,2]

Although plasma proteins may serve as markers of neurological disease risk, the molecular mechanisms responsible for inter-individual variation in plasma protein levels are poorly understood. Therefore, we conduct genome- and epigenome-wide association studies on the levels of 92 neurological proteins to identify genetic and epigenetic loci associated with their plasma concentrations (n = 750 healthy older adults). We identify 41 independent genome-wide significant ($P < 5.4 \times 10^{-10}$) loci for 33 proteins and 26 epigenome-wide significant ($P < 3.9 \times 10^{-10}$) sites associated with the levels of 9 proteins. Using this information, we identify biological pathways in which putative neurological biomarkers are implicated (neurological, immunological and extracellular matrix metabolic pathways). We also observe causal relationships (by Mendelian randomisation analysis) between changes in gene expression (DRAXIN, MDGA1 and KYNU), or DNA methylation profiles (MATN3, MDGA1 and NEP), and altered plasma protein levels. Together, this may help inform causal relationships between biomarkers and neurological diseases.

[1] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. [2] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK. [3] Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK. [4] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia. [5] Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh EH16 4UX, UK. [6] Department of Psychology, The University of Texas at Austin, Austin, TX 78712, USA. [7] Population Research Center, The University of Texas at Austin, Austin, TX 78712, USA. Correspondence and requests for materials should be addressed to R.E.M. (email: riccardo.marioni@ed.ac.uk)

Plasma proteins execute diverse biological processes and aberrant levels of these proteins are implicated in various disease states. Consequently, plasma proteins may serve as biomarkers, contributing to individual disease risk prediction and personalised clinical management strategies[1]. Identifying circulating biomarkers is of particular importance in neurological disease states in which access to diseased neural tissue in vivo is almost impossible. Furthermore, in neurodegenerative disorders, symptomatology may appear in only advanced clinical states, necessitating early detection and intervention[2]. Elucidating the factors which underpin inter-individual variation in plasma protein levels can inform disease biology and also identify proteins with likely causal roles in a given disease, augmenting their value as predictive biomarkers. Indeed, studies have characterised genetic variants (protein quantitative trait loci; pQTLs) associated with circulating protein levels and utilised such genetic information to identify proteins with causal roles in conditions such as cardiovascular diseases[3–5]. However, studies which have aimed to examine the genetic determinants of neurology-related protein levels in human plasma are limited[6–8]. Furthermore, few studies have combined genetic with epigenetic data to provide an additional layer of information regarding the molecular mechanisms responsible for regulating blood protein levels[9]. Therefore, the goal of the present study was to characterise genetic and epigenetic (using DNA methylation) factors associated with putative neurology-related protein biomarkers in order to identify potential molecular determinants which regulate their plasma levels.

Here, genome-wide and epigenome-wide association studies (GWAS/EWAS) are carried out on the plasma levels of 92 neurological proteins in 750 relatively healthy older adults from the Lothian Birth Cohort 1936 study (mean age: 73; levels adjusted for age, sex, population structure and array plate; hereafter simply referred to as protein levels). These proteins represent the Olink® neurology panel and encompass a mixture of proteins with established links to neurobiological processes (such as axon guidance and synaptic function) and neurological diseases (such as Alzheimer's disease (AD)), as well as exploratory proteins with roles in processes including cellular regulation, immunology, and development. Following the identification of genotype-protein associations (pQTLs), functional enrichment analyses are performed on independent pQTL variants. Upon identification of epigenetic factors associated with protein levels, tissue specificity and pathway enrichment analyses are conducted to reveal possible biological pathways in which neurological proteins are implicated. Protein QTL data are integrated with publicly available expression QTL data to probe the molecular mechanisms which may modulate circulating protein levels. Finally, GWAS summary data for proteins and disease states are integrated using two-sample Mendelian Randomisation (MR) to determine whether selected proteins are causally associated with neurological disease states.

Results

Genome wide study of neurological protein biomarkers. For the GWAS, a Bonferroni P value threshold of 5.4×10^{-10} (genome-wide significance level: 5.0×10^{-8}/92 proteins) was set. The GWAS analysis in 750 older adults identified 2734 significant SNPs associated with 37 proteins (Supplementary Data 1). Conditional and joint analysis (GCTA-COJO) resulted in the identification of 41 conditionally significant pQTLs associated with the levels of 33 proteins ($P < 5.4 \times 10^{-10}$; Fig. 1a; Supplementary Data 2). Notably, while genome-wide significant associations were present for an additional four proteins (Alpha-2-MRAP, CD38, MRS1 and SMPD1), the conditional P value for these signals following COJO (n = 1 independent signal per protein) did not surpass the Bonferroni-corrected threshold of $P < 5.4 \times 10^{-10}$. Of these 41 variants, 36 (87.8%) were *cis* pQTLs (SNP within 10 Mb of the transcription start site (TSS) of the gene) and 5 (12.2%) were *trans* variants. Three of the five *trans* variants were located on chromosomes distinct from their respective Olink® gene. Furthermore, *cis* only associations were present for 28/33 proteins (84.8%), compared to *trans* only associations for 3/33 proteins (9.1%). Two proteins (6.1%) were associated with both *cis* and *trans* pQTLs (CD200R and Siglec-9). For all conditionally significant *cis* pQTLs associated with a given protein, the pQTL with the lowest P value was denoted as the sentinel variant (n = 30). The significance of *cis* associations decreased as the distance of the sentinel variant from the TSS increased (Fig. 1b).

The minor allele frequency of independent pQTL variants was inversely associated with effect size (Fig. 1c). Notably, this association may be, in part, due to ascertainment bias as rarer variants (with lower minor allele frequencies) must have large effect sizes to attain the same level of power as more common variants. Independent pQTLs explained between 5.1% (rs12139487; DRAXIN; $P = 4.38 \times 10^{-10}$) and 52.5% (rs6938061; MDGA1; $P = 1.39 \times 10^{-87}$) of the phenotypic variance in plasma protein levels (Supplementary Data 2; Fig. 1d). The majority of pQTL variants were located in intergenic and intronic regions (Supplementary Data 2; Fig. 1e). The number of independent loci associated per protein is shown in Fig. 1f. One *trans* conditionally significant variant (rs4857414) was shared between Siglec-9 and CD200R. This variant was annotated to the *ST3GAL6-AS1* gene. Figure 2 demonstrates the effect of genetic variation at the most significant *cis* pQTL (rs6938061; MDGA1) and *trans* pQTL (rs4857414; Siglec-9) on protein levels.

We also used an alternative method, FUMA (FUnctional Mapping and Annotation) to find independent pQTLs. This approach identified 62 significant pQTLs associated with the levels of 37 proteins (90.3% *cis* and 9.7% *trans* effects; Bonferroni-corrected level of significance: $P < 5.4 \times 10^{-10}$) (Supplementary Data 3). In contrast to GCTA-COJO, FUMA retains the most significant pQTL to identify independent signals through linkage disequilibrium (LD)-based pruning; therefore, variants were identified for all 37 proteins. Seven independent pQTLs associated with the levels of 6 proteins were found using both approaches, whereas the remaining SNPs identified by COJO for a given protein were located within the same locus as corresponding SNPs identified by FUMA (overlapping SNPs highlighted in Supplementary Data 2). In addition, we calculated a measure of LD (r^2) between SNPs which were discordant between COJO and FUMA. As 7 independent pQTLs were identified by both methods, this left 34 (41–7) discordant SNPs from COJO and 55 (62–7) discordant SNPs from FUMA. Furthermore, as some proteins contained multiple QTLs, this resulted in 74 SNP-SNP comparisons between COJO and FUMA. SNPs which exhibited an r^2 coefficient > 0.75 were considered to show evidence of replication (through LD) between both methods. In total, 27 COJO SNP-FUMA SNP comparisons exhibited an r^2 > 0.75. This consisted of 26 unique SNPs identified by COJO and encompassed 24 proteins (Supplementary Table 1).

Colocalisation of *cis* pQTLs with *cis* eQTLs. Of the 30 sentinel *cis* pQTL variants, 12 (40.0%) were *cis* eQTLs for the same gene in blood tissue. For 3/12 proteins (DRAXIN, KYNU and MDGA1), there was strong evidence (posterior probability (PP) > 0.95) for colocalisation of *cis* pQTLs and *cis* eQTLs and for 2 proteins, LAIR-2 and SIGLEC9, there was weaker evidence (PP > 0.75) for colocalisation. For 5/12 proteins, there was strong evidence

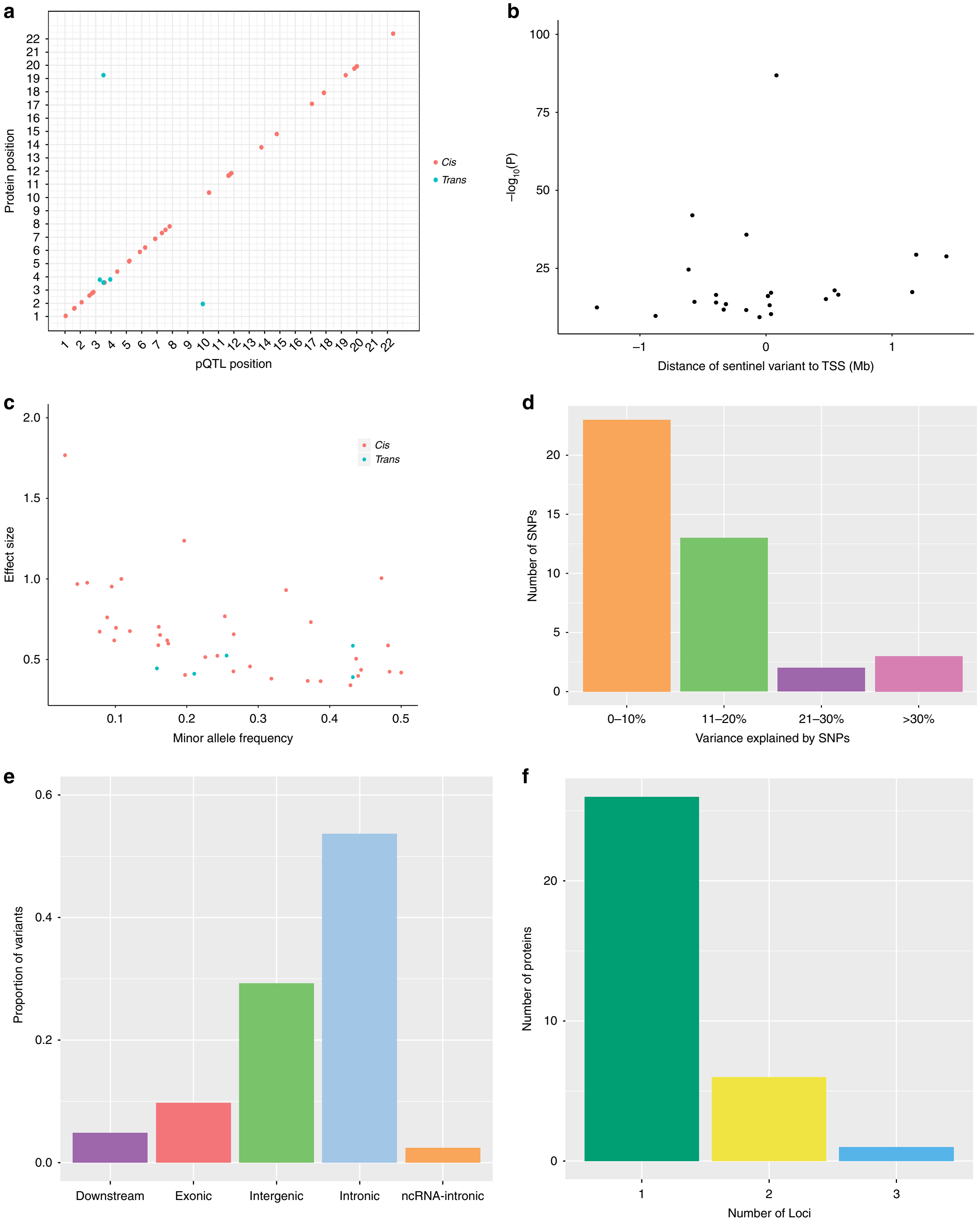

Fig. 1 Genome-wide association study of neurological protein biomarkers. **a** Chromosomal locations of pQTLs. The *x*-axis represents the chromosomal location of conditionally significant *cis* and *trans* SNPs associated with the levels of Olink® neurology proteins. The *y*-axis represents the position of the gene encoding the associated protein. *Cis* (red circles); *trans* (blue circles). **b** Significance of sentinel *cis* variants versus distance of variants from the gene transcription start site. **c** Absolute effect size (per standard deviation of difference in protein level per effect allele) of conditionally significant pQTLs versus minor allele frequency. *Cis* (red circles); *trans* (blue circles). **d** Variance in protein levels explained by conditionally significant pQTLs. It is important to note that these estimates may be inflated owing to winner's curse or over-fitting in the discovery GWAS. **e** Classification of pQTL variants by function as defined by functional enrichment analysis in FUMA. **f** Number of conditionally significant pQTL variants per Olink® neurology protein

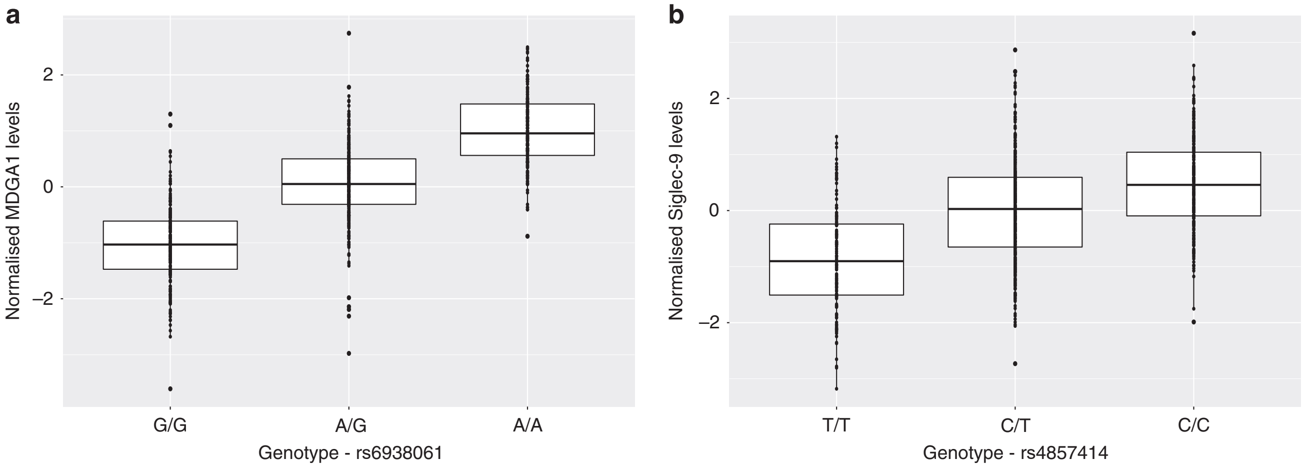

Fig. 2 Effect of genetic variation on neurological protein levels. **a** Box plot of MDGA1 levels as a function of genotype (rs6938061, effect allele: A, other allele: G, beta = 1.00, se = 0.05). **b** Box plot of Siglec-9 levels as a function of genotype (rs4857414, effect allele: T, other allele: C, beta = −0.58, se = 0.05). Centre line of boxplot: median, bounds of box: first and third quartiles and tips of whiskers: minimum and maximum

(PP > 0.95) for two distinct causal variants affecting transcript and protein levels in the locus. For CTSC, there was weaker evidence (PP > 0.75) for two separate causal signals affecting gene expression and plasma protein levels within the locus. Finally, for CLM-6, there was weak evidence (PP > 0.75) for a causal variant affecting gene expression, but not protein levels, within the locus (Supplementary Table 2).

For the 3 proteins with strong evidence in favour of a shared causal variant for gene expression and plasma protein levels, two-sample MR was performed to test for a causal association between perturbations in gene expression (using data from eQTLGen Consortium) and plasma protein levels (using our GWAS data). Pruned *cis* protein and expression QTL variants (LD $r^2 < 0.1$) were used as instrumental variables for the bidirectional MR analyses. For each trait, the intercept from MR Egger regression was non-significant, which does not suggest strong evidence for directional pleiotropy (DRAXIN: $P = 0.82$; MDGA1: $P = 0.38$; KYNU: $P = 0.36$). For 2 proteins, variation in gene expression was causally associated with plasma protein levels (Inverse variance-weighted method; MDGA1: beta = 0.99, se = 0.49, $P = 0.02$; KYNU: beta = 1.05, se = 0.22, $P = 2.2 \times 10^{-6}$). We did not observe a causal relationship between gene expression of DRAXIN and altered plasma protein levels (Inverse variance-weighted method; beta = −0.98, se = 0.62, $P = 0.10$); however, we did observe a causal relationship between DRAXIN plasma protein levels and changes in gene expression (beta = −0.72, se = 0.07, $P = 1.2 \times 10^{-23}$).

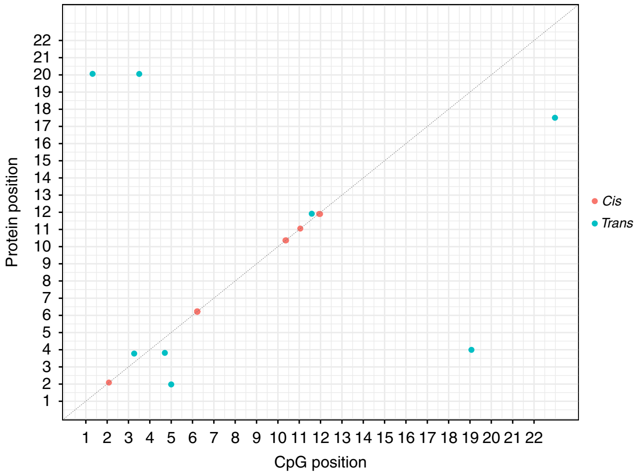

Fig. 3 Genomic locations of CpG sites associated with differential neurological protein levels. The *x*-axis represents the chromosomal location of CpG sites associated with the levels of Olink® neurology biomarkers. The *y*-axis represents the position of the gene encoding the associated protein. Notably, *cis* CpG sites ($n = 17$) identified by our EWAS on protein levels lay within the same cluster for a given protein. Some of these CpG sites lay too close to discriminate, resulting in the appearance of 5 *cis* CpG clusters in this figure

Epigenome wide study of neurological protein biomarkers. For the EWAS, a Bonferroni *P* value threshold of 3.9×10^{-10} (genome-wide significance level: 3.6×10^{-8}/92 proteins) was set[10] and analyses were performed using limma, a linear-model-based method. We identified 26 genome-wide significant CpG sites associated with the levels of 9 neurological proteins ($P < 3.9 \times 10^{-10}$). Of these associations, 17 were *cis* effects (65.4%) and 9 associations were *trans* effects (35.6%; with 6 *trans* variants located on chromosomes distinct from their respective Olink® gene) (Fig. 3; Supplementary Table 3). As an additional analysis, we performed a mixed-linear-model approach termed OSCA (OmicS-data-based Complex trait Analysis)-MOMENT. OSCA has been recently shown to identify fewer spurious signals than other methods (including linear regression) (Zhang et al.[41]). Of the 9 proteins with genome-wide significant CpG sites identified using limma ($n = 26$ CpG sites), 8 proteins were also shown to have genome-wide significant associations using OSCA ($n = 23$ CpG sites; 14 *cis* (60.9%) and 9 *trans* (39.1%) associations). Indeed, only CRTAM failed to show a Bonferroni-corrected significant association using OSCA when compared to limma. Furthermore, of the 23 CpG sites identified using OSCA, 19/23 CpGs (82.6%) were also identified by EWAS performed using limma showing a strong overlap between both methods (Supplementary Table 4).

Three proteins exhibited both genome-wide significant SNP and CpG site associations: MATN3, MDGA1, and NEP (Fig. 4). For MATN3, the *cis* pQTL identified in this study (rs3731663) has previously been identified as a methylation QTL (mQTL) for the single *cis* CpG site associated with MATN3 levels identified by our EWAS (cg24416238)[11]. Similarly, the 2 *cis* pQTLs for differential blood MDGA1 concentrations in our study have been significantly associated with methylation levels of *cis* CpG

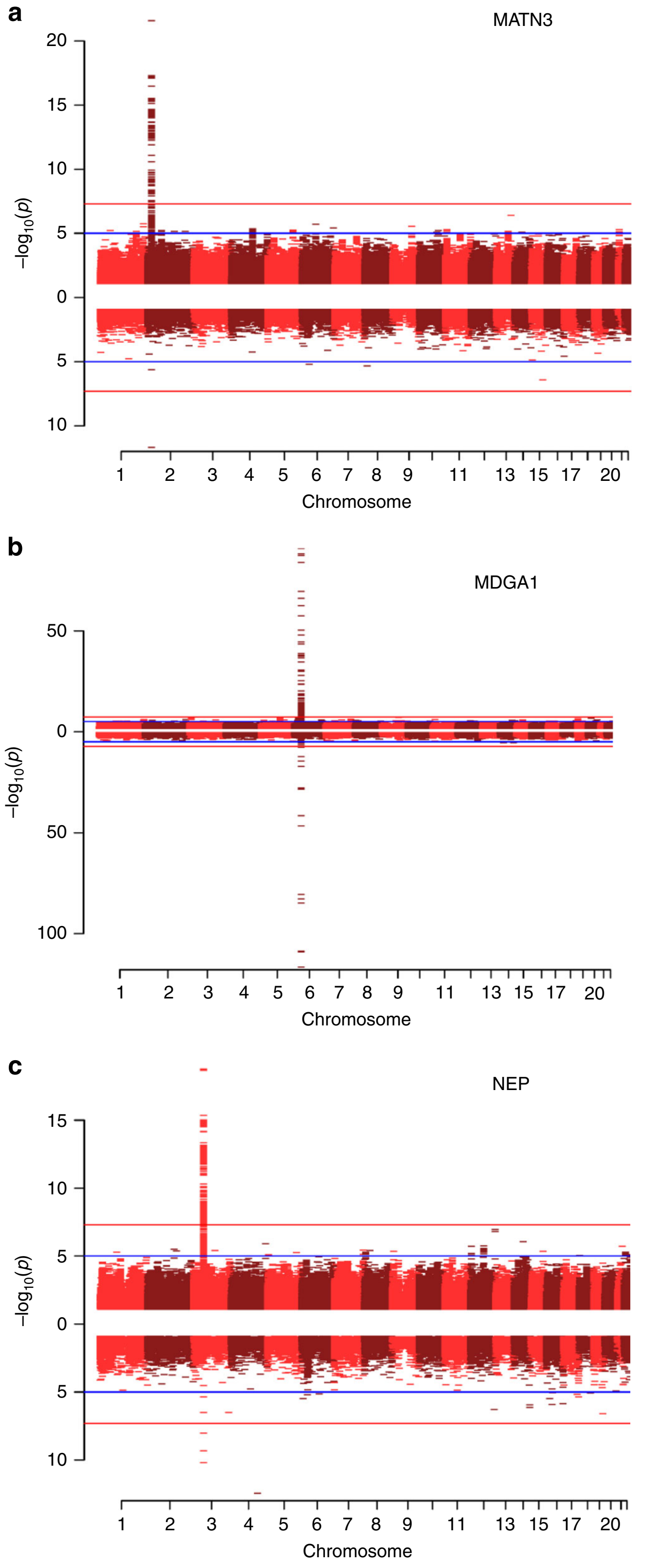

Fig. 4 Miami plots of three neurological proteins with both genome-wide significant SNP and genome-wide significant CpG associations. The top half of the plot (skyline) shows the results from the GWAS on protein levels, whereas the bottom half (waterfront) shows the results from the EWAS. Blue lines indicate suggestive associations; red lines indicate epigenome-wide significant associations. **a** Miami plot for MATN3 (chromosome 2: 20,191,813–20,212,455). **b** Miami plot for MDGA1 (chromosome 6: 37,600,284–37,667,082). **c** Miami plot for NEP (chromosome 3: 154,741,913–154,901,518)

sites identified by our EWAS on MDGA1 levels[11]. Finally, for NEP, we identified a sole independent *trans* pQTL (rs4687657) annotated to the *ITIH4* gene (beta: 0.53; effect allele: T), as well as three *trans* genome-wide significant CpG sites (cg11645453, cg18404041 and cg06690548 annotated to *ITIH1*, *ITIH4* and *SLC7A11*, respectively). In addition to higher circulating levels of NEP, this SNP has previously been associated with lower methylation levels of cg18404041 (*ITIH4*; beta: −0.93; effect allele: T; $P = 4.20 \times 10^{-17}$) and higher methylation levels of cg11645453 (*ITIH1*; beta: 0.83; effect allele: T; $P = 1.28 \times 10^{-87}$)[12]. We performed bidirectional MR analyses to formally test whether there was a causal relationship between DNA methylation at these sites and Olink® protein levels (see methods). For each protein, MR analyses suggested that differential DNA methylation was causally associated with changes in protein levels. Conversely, altered protein levels of MATN3, MDGA1 and NEP were also causally associated with differential methylation levels at CpG sites identified by our EWAS (Supplementary Table 5).

We conducted tissue specificity and pathway enrichment analyses (KEGG and GO—see methods for details) based on genes identified by methylation for each of the 9 proteins with genome-wide significant CpG associations. Tissue-specific patterns of expression were observed for 5/9 proteins (Supplementary Data 4). Neural tissue was the most common tissue type in which genes were differentially expressed ($n = 4/5$ proteins), followed by cardiac and splenic tissue ($n = 3/5$ proteins). Gene ontology analyses revealed that genes annotated to CpG sites associated with circulating SIGLEC1 and G-CSF levels are over-represented in immune system processes, viral response and cytokine response pathways (Supplementary Data 5–6; FDR-adjusted P value < 0.05). Furthermore, genes incorporating CpG sites associated with NEP levels are over-represented in metabolic pathways involving extracellular matrix components (Supplementary Data 7; FDR-adjusted P value < 0.05). For CRTAM, MDGA1, MATN3, NC-Dase, SMPD1 and TN-R, there were no significant results following multiple testing correction.

Causal evaluation of biomarkers in neurological disease. From our GWAS, we identified a conditionally significant *cis* pQTL for plasma poliovirus receptor (PVR) levels. Furthermore, variation in the *PVR* gene has been implicated in AD[13]. Therefore, colocalisation analysis was performed to test if the same SNP variant might be driving both associations. A 200 kb region surrounding the sentinel *cis* pQTL for PVR was extracted from GWAS summary statistics for PVR levels, as well as AD[14]. Default priors were applied. There was evidence to suggest that there are two distinct causal variants for altered protein levels and AD risk within the region (PP > 0.99).

In addition to the colocalisation analysis, two-sample MR was used to test for putatively causal associations between plasma PVR levels and AD[14]. After LD pruning, only one independent SNP remained (rs7255066). Therefore, causal effect estimates were determined using the Wald ratio test, i.e., a ratio of effect per risk allele on AD to effect per risk allele on PVR levels. MR analyses indicated that PVR levels were causally associated with AD (beta = 0.17, se = 0.02, $P = 5.2 \times 10^{-10}$; Wald ratio test). Testing for horizontal pleiotropy was not possible owing to an insufficient number of instruments. Conversely, AD risk was not causally associated with PVR levels (number of SNPs: 5; Inverse variance-weighted method: beta = 0.38, se = 0.29, $P = 0.34$). The intercept from MR Egger regression was −0.08 (se: 0.08; $P = 0.42$), which does not provide strong evidence for directional pleiotropy.

Replication of previous pQTL studies. Replication of the pQTL findings was carried out via lookup of genotype-protein summary statistics from existing pQTL studies[4,5,15,16]. Of the 33 proteins with a conditionally significant pQTL in the present study, 15 (with 18 QTLs) were available for lookup. In total, 6/18 (33.3%) pQTLs replicated at $P < 1.25 \times 10^{-7}$ (denoting the least conservative threshold across all studies) (Supplementary Data 8). We tested the correlation of beta values for these six significant pQTLs from our study versus those reported in the literature. Notably, beta values were only available for 3/6 pQTLs in the literature. However, for these remaining 3 pQTLs, there was strong agreement between our observed values and previously reported beta statistics (rs2075803: 0.50 vs. 0.55; rs481076: 0.44 vs. 0.46 and rs1448903: 0.76 vs. 0.65, respectively). In addition, in relation to the 15 proteins from the Olink® panel which were available for look-up, we extracted beta values for all significant pQTLs associated with the levels of these proteins reported in the literature. Notably, many of these pQTLs were non-significant in our study; indeed, in this case, we wished only to determine the correlation of betas for those pQTLs reported as significant in the literature with betas from our GWAS. Beta statistics were reported for 13/15 proteins (totalling 38 pQTLs). There was a strong correlation between betas for previously reported significant pQTLs and pQTLs from our study ($r^2 = 0.89$, Supplementary Fig. 1). Finally, of the 23 pQTLs identified by FUMA which were available for look-up, 9/23 (39.1%) replicated at $P < 1.25 \times 10^{-7}$ (Supplementary Data 9).

Discussion

Using a multi-omics approach, we identified 41 independent genome-wide significant pQTLs and 26 genome-wide significant CpG sites associated with circulating neurological protein levels. To probe the molecular mechanisms which modulate plasma protein levels, we integrated pQTL and eQTL data allowing for the examination of whether pQTLs affect gene expression. For three proteins, we found strong evidence that a common causal variant underpinned changes in transcript and protein levels. Mendelian randomisation analyses suggested that variants for two of these proteins (MDGA1 and KYNU) influence protein levels by altering gene expression. However, for one protein (DRAXIN), the converse may be true as our data suggested that altered plasma protein levels of this neurodevelopmental protein may affect gene expression, perhaps through a feedback mechanism. Genotype-protein associations for other proteins may exert their influence on protein levels through modulation of protein clearance, degradation, binding or secretion. Finally, methylation data revealed that neurological proteins were also implicated in immune, developmental and metabolic pathways.

In addition to leveraging methylation data to identify pathway enrichment for plasma proteins, identification of *trans* pQTLs may highlight previously unidentified pathways relevant to disease processes. For instance, we found that genetic variation at the inter-alpha-trypsin inhibitor heavy chain family member 4 locus (*ITIH4)* is associated with differential NEP levels (*trans* pQTL: rs4687657). In addition, two CpG sites annotated to *ITIH4* and *ITHI1* (cg18404041 and cg11645453, respectively) were associated with NEP levels. Methylation QTL analyses revealed that the SNP rs4687657 has been previously associated with lower methylation levels of cg18404041 (*ITIH4*) and higher DNA methylation levels of cg11645453 (*ITIH1*)[12]. Similarly, this SNP has been associated with higher gene expression of *ITIH4*[17] and lower protein levels of *ITIH1*[4]. Together, these data suggest that the expression of NEP, ITIH4 and ITIH1 may be co-regulated, involving inverse relationships between NEP and ITIH4 with ITIH1. Given that mutations in *NEP* have been linked to Alzheimer's pathology and that upregulation of ITIH4 has been demonstrated in sera of AD patients[18], mechanistic studies relating to co-expression of these proteins are merited in pathological states.

In this study, a single *trans* variant (rs4857414) was associated with the circulating levels of two proteins—CD200R1 and Siglec-9. This polymorphism mapped to the *ST3GAL6-AS1* gene. ST3GAL6-AS1 is a long non-coding RNA which is associated with increased expression of ST3GAL6, an enzyme responsible for catalysing the addition of sialic acid to cell surfaces[19]. Upregulation of ST3GAL6 has been reported in multiple myeloma[20,21]; this permits evasion of immune responses against cancer cells through binding of sialic acid to Siglec receptor proteins, such as Siglec-9. The recognition of sialic acid by Siglec proteins ignites signalling cascades which promotes immune inhibitory responses[22,23]. Furthermore, CD200-CD200R interaction results in the inhibition of immune responses against multiple myeloma cells[24]. Therefore, as polymorphisms in *ST3GAL6-AS1* are associated with altered expression of Siglec-9 and CD200R, this may provide further evidence for co-regulation of these proteins in pathological milieux, such as tumorigenesis in cancers including multiple myeloma. Polymorphisms in such *trans* pQTLs may also be used to predict disease risk, progression and provide pharmacogenomic information in predicting individual patient responses to inhibition of these co-regulated proteins.

By using *cis* pQTLs as instruments for MR analyses, it is possible to test whether plasma proteins are causally associated with disease states[25]. *PVR* is a component of the AD risk-associated *APOE*/*TOMM40* cluster on chromosome 19 and has been hypothesised to influence risk of AD through susceptibility to viral infections[13]. However, it is unknown whether PVR is causally linked to the disease. MR analyses suggested that circulating PVR levels may be causally associated with AD and not vice versa. However, an insufficient number of instruments were available to permit testing for potential pleiotropic effects. Furthermore, colocalisation analysis revealed that independent variants in the *PVR* locus are likely causally associated with altered plasma PVR levels and AD risk. While this does not support the argument for a single causal SNP underlying both altered plasma PVR levels and AD risk, it may nevertheless suggest that genetic variation in the PVR locus is causally associated with development of AD.

The discrepancy in replication of pQTLs reported in previous studies may be due to a number of factors. First, the sample sizes of these studies ($n < 100$[15,16]; $n > 1000$[4,5]) are different from that of the present study ($n = 750$) leading to differences in statistical power. Second, diverse proteomic platforms may result in the detection of different genotype-protein associations. Our study is the first to characterise the genetic variants associated with the Olink® neurology panel and thus, the protein list and measurement technology do not overlap with platforms employed in earlier studies. Depending on platform technology, susceptibility to cross-reactive events and detection of proteins in their free, versus complexed, forms can result in inappropriate readouts. SOMAmer technology, employed in the previous pQTL studies, is a highly sensitive, aptamer-based platform which overcomes limitations associated with antibody-based methods, such as cross-reactivity[26]. Moreover, Olink® technology is particularly effective in limiting the reporting of cross-reactive events. However, when compared to these platforms, other technologies such as mass-spectrometry can produce highly accurate measurements but with low sensitivity[27]. Lack of standardisation amongst proteomic platforms, insufficient power to detect associations and differences in study demographics may all contribute to variability in the detection of pQTLs for a given protein. In addition, we performed both FUMA (LD-based method) and COJO (stepwise conditional regression) to identify independent pQTL-protein associations and found a small overlap (17%) between SNPs identified by both methods. However, SNPs which were differentially identified by COJO and FUMA for a given protein were located within the same region. Indeed, the maximum distance between discordant SNPs for a given protein was 3 Mb.

We acknowledge several limitations in the present study. First, analyses were restricted to individuals of European descent, complicating the generalisability of our findings to individuals of other ethnic backgrounds. Second, functional enrichment analyses indicated that a number of *cis* pQTL variants may alter the amino acid sequence of the coded protein. This may lead to altered structural properties of the protein product, resulting in impaired antibody–antigen binding and consequently, the ability of assays to accurately detect protein levels. Notably, as the LBC1936 cohort consists of relatively healthy older adults, it is possible that levels of putative neurological-disease related proteins may differ in the general elderly population. Therefore, this may complicate the generalisability of our findings to other age ranges and other elderly cohorts with higher incidences of neurological and psychiatric conditions. Finally, as our findings pertain to whole blood samples, studies examining the genetic and epigenetic regulation of neurological proteins in *post-mortem* brain tissue are warranted.

In conclusion, we have identified genetic and epigenetic factors associated with neurological proteins in an older-age population. We have shown that use of a multi-omics approach can help define whether such proteins are causal in disease processes. We have shown that PVR may be causally associated with AD. Furthermore, we have provided a platform upon which future studies can interrogate pathophysiological mechanisms underlying neurological conditions. Together, this information may help inform disease biology, as well as aid in the prediction of disease risk and progression in clinical settings.

Methods

The Lothian Birth Cohort 1936. The Lothian Birth Cohort 1936 (LBC1936) comprises Scottish individuals born in 1936, most of whom took part in the Scottish Mental Survey 1947 at age 11. Participants who were living within Edinburgh and the Lothians were re-contacted approximately 60 years later, 1091 consented and joined the LBC1936. Upon recruitment, participants were approximately 70 years of age (mean age: 69.6 ± 0.8 years). Participants subsequently attended four additional waves of clinical examinations every three years. Detailed genetic, epigenetic, physical, psychosocial, cognitive, health and lifestyle data are available for members of the LBC1936. Recruitment and testing of the LBC1936 cohort have been described previously[28,29]. LBC1936 participants were 49.8% female. Key inclusion/exclusion criteria for the present study are highlighted in Supplementary Fig. 2.

Ethical approval. Ethical permission for the LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics Committee (LREC/2003/2/29). Written informed consent was obtained from all participants.

Protein measurements in the Lothian Birth Cohort 1936. Plasma was extracted from 816 blood samples collected in citrate tubes at mean age 72.5 ± 0.7 years (Wave 2). Plasma samples were analysed using a 92-plex proximity extension assay (Olink® Bioscience, Uppsala Sweden). The proteins assayed constitute the Olink® neurology biomarker panel. This panel represents proteins with established links to neuropathology, as well as exploratory proteins with roles in processes including cellular communication and immunology. In brief, 1 µL of sample was incubated in the presence of proximity antibody pairs linked to DNA reporter molecules. Upon binding of an antibody pair to their corresponding antigen, the respective DNA tails form an amplicon by proximity extension, which can be quantified by high-throughput real-time PCR. This method limits the reporting of cross-reactive events. The data were pre-processed by Olink® using NPX Manager software. Protein levels were transformed by rank-based inverse normalisation. Normalised plasma protein levels were then regressed onto age, sex, four genetic principal components of ancestry derived from the Illumina 610-Quadv1 genotype array (to control for population structure) and Olink® array plate. To obtain an estimate of population structure, multidimensional scaling (MDS) was performed on LBC1936 genotyping data and the first four MDS components were used to control for genetic ancestry in the analytic models. Standardised residuals from these linear regression models were used in our genome-wide and epigenome-wide association studies. Pre-adjusted (raw) and transformed (rank-based inverse normalised levels regressed on age, sex, population structure and array plate) protein levels are presented in Supplementary Data 10 and 11, respectively. The associations of pre-adjusted protein levels with biological and technical covariates are presented in Supplementary Data 12.

Methylation preparation in the Lothian Birth Cohort 1936. DNA from whole blood was assessed using the Illumina 450 K methylation array at the Edinburgh Clinical Research Facility (Wave 2; $n = 895$; mean age: 72.5 ± 0.7 years). Details of quality control procedures have been described in detail elsewhere[30]. Briefly, raw intensity data were background-corrected and normalised using internal controls. Following background correction, manual inspection permitted removal of low quality samples presenting issues relating to bisulphite conversion, staining signal, inadequate hybridisation or nucleotide extension. Quality control analyses were performed to remove probes with low detection rate <95% at $P < 0.01$. Samples with a low call rate (samples with <450,000 probes detected at p-values of less than 0.01) were also eliminated. Furthermore, samples were removed if they had a poor match between genotype and SNP control probes, or incorrect DNA methylation-predicted sex.

Genotyping in the Lothian Birth Cohort 1936. LBC1936 DNA samples were genotyped at the Edinburgh Clinical Research Facility using the Illumina 610-Quadv1 array (Wave 1; $n = 1005$; mean age: 69.6 ± 0.8 years; San Diego). Preparation and quality control steps have been reported previously[31]. SNPs were imputed to the 1000 G reference panel (phase 1, version 3). Individuals were excluded on the basis of sex discrepancies, relatedness, SNP call rate of less than 0.95, and evidence of non-Caucasian descent. SNPs with a call rate of greater than 0.98, minor allele frequency in excess of 0.01, and Hardy-Weinberg equilibrium test with $P \geq 0.001$ were included in analyses.

Genome-wide association studies. Genome-wide association analyses were conducted on 8,683,751 autosomal variants against protein residuals in 750 individuals from the Lothian Birth Cohort 1936. Linear regression was used to assess the effect of each genetic variant on the protein residuals using mach2qtl[32,33].

GWAS model: Olink® protein residuals~SNP

Epigenome-wide association studies. Epigenome-wide association analyses were conducted by regressing each of 459,309 CpG sites (as dependent variables) on transformed protein levels using linear regression with adjustments for age, sex, estimated white blood cell proportions (CD4$^+$ T cells, CD8$^+$ T cells, B cells, Natural Killer Cells and granulocytes) and technical covariates (plate, position, array, hybridisation, date). White blood cell proportions were estimated from methylation data using the Houseman method[34]. Outliers for white blood cell proportions ($n = 22$) were excluded prior to analyses. Complete methylation and proteomic data were available for 692 individuals. Genome-wide significant CpG associations mapping to sites with underlying polymorphisms were excluded, as well as those predicted to cross-hybridise based on findings by Chen et al.[35]. Analyses were performed using the limma package in R[36].

EWAS model: CpG site~Olink® protein residuals + age + sex + estimated white blood cell proportions + array + plate + date + set + position

Pathway enrichment was assessed among KEGG pathways and Gene Ontology (GO) terms via hypergeometric tests using the *phyper* function in R. All gene symbols from the 450 K array annotation (null set of sites) were converted to Entrez IDs using biomaRt[37,38]. GO terms and their corresponding gene sets were obtained from the Molecular Signatures Database (MSigDB)-C5[39] while KEGG pathways were downloaded from the KEGG REST server[40]. Furthermore, tissue specificity analyses were conducted using the GENE2FUNC function in FUnctional Mapping and Annotation (FUMA). Differentially expressed gene sets with Bonferroni-corrected P values of <0.05 and an absolute log-fold change of ≥0.58 (default settings) were considered to be enriched in a given tissue type (GTEx v7).

OSCA. We also performed EWAS analyses of Olink® protein levels using OmicS-data-based Complex trait Analysis software (OSCA). We carried out OSCA as an additional EWAS analysis as it has recently been shown to identify less spurious associations when compared to other methods (including linear regression)[41]. CpG site was the independent variable whereas Olink® protein levels were input as dependent variables. Models were adjusted for age, sex, estimated white blood cell proportions (CD4$^+$ T cells, CD8$^+$ T cells, B cells, Natural Killer Cells and granulocytes) and technical covariates (plate, position, array, hybridisation, date) as in the previous section. The MOMENT method was used to test for associations between traits of interest and DNAm at individual probes. MOMENT is a mixed-linear-model-based method that can account for unobserved confounders and the correlation between distal probes which may be introduced by such confounders.

Conditional and joint analysis. We performed approximate genome-wide stepwise conditional analysis through GCTA-COJO using the 'cojo-slct' option as the primary means to identify independent genetic-protein associations[42]. Individual level genotype data were used with default settings of the software.

Functional mapping and annotation of pQTLs. In addition to GCTA-COJO, the identification of independent pQTL variants from the GWAS which yielded significant genotype-protein associations, and their subsequent functional annotation, were performed using the independent SNP algorithm implemented in FUMA analysis[43]. Initial independent significant SNPs were identified using the SNP2GENE function. These were defined as variants with a P value of $<5 \times 10^{-8}$ that were independent of other genome-wide significant SNPs at $r^2 < 0.6$. Lead independent SNPs were further defined as the initial independent significant SNPs that were independent from each other at $r^2 < 0.1$. Independent significant SNPs were functionally annotated using ANNOVAR[44] and Ensembl genes (build 85).

Characterisation of *cis* and *trans* effects. Genome-wide significant pQTLs and CpG sites were categorised into *cis* and *trans* effects. *Cis* associations were defined as loci which reside within 10 Mb of the TSS of the gene encoding the protein of interest. *Trans* effects were defined as those loci which lay outside of this region or were located on a chromosome distinct from that which harboured the gene TSS. TSS positions were defined using the biomaRt package in R[37,38] and Ensembl v83.

Identification of overlap between *cis* pQTLs and eQTLs. We cross-referenced sentinel *cis* pQTLs with publicly available *cis* eQTL data from the eQTLGen consortium[45]. *Cis* eQTLs were filtered to retain only variants with $P < 5.4 \times 10^{-10}$. Furthermore, only *cis* eQTLs for the same gene as the *cis* pQTL protein were retained. These associations were then tested for colocalisation.

Colocalisation analysis. To test the hypothesis that a single causal variant might underlie both an eQTL and pQTL, resulting in modulation of transcript and protein levels, we conducted Bayesian tests of colocalisation. Colocalisation analyses were performed using the coloc package in R[46]. For each pQTL variant, a 200 kb region (upstream and downstream) was extracted from our GWAS summary statistics for each protein of interest. This window previously has been recommended in order to capture *cis* eQTLs, which often lie within 100 kb of their target gene[47]. Expression QTLs for genes within this region were extracted from eQTLGen consortium summary statistics and subset to the gene encoding the protein of interest[45]. All SNPs shared by transcripts and proteins were used to determine the posterior probability for five distinct hypothesis. Default priors were applied. Posterior probabilities (PP) > 0.95 provided strong evidence in favour of a given hypothesis. Hypothesis 4 states that two association signals were attributable to the same causal variant. Associations with PP4 > 0.95 were deemed highly likely to colocalise. Associations with PP3 > 0.95 provided strong evidence for hypothesis 3 that there were independent causal variants for protein levels and gene expression. In this study, hypothesis 2 referred to a causal variant for condition 2 (gene expression only) whereas hypothesis 1 represented a causal variant for protein levels only. Associations with PP0 > 0.95 (for hypothesis 0) indicated that it is highly likely there were no causal variants for either trait in the region.

Mendelian randomisation. Two-sample bidirectional Mendelian randomisation was used to test for putatively causal relationships between (i) PVR, a cell-surface glycoprotein, and AD risk, (ii) gene expression and plasma protein levels and (iii) DNA methylation and plasma protein levels. Pruned variants (LD $r^2 < 0.1$) were used as instrumental variables (IV) in MR analyses. In cases where only one independent SNP remained after LD pruning, causal effect estimates were determined using the Wald ratio test. When multiple independent variants were present, and if no evidence of directional pleiotropy was present (non-significant MR-Egger intercept), multi-SNP MR was conducted using inverse variance-weighted estimates. All MR analyses were conducted using MRbase[48].

(i) While 72 genome-wide significant *cis* pQTLs were identified for PVR levels, only one SNP (rs7255066) remained after LD pruning. Five independent SNPs were identified and used as IV to test for a causal relationship between AD risk and altered plasma PVR levels.
(ii) Expression QTLs obtained from eQTLGen consortium were used as IV to test whether changes in gene expression were causally associated with protein levels[45]. Protein QTLs identified by our GWAS were used as IV to test whether protein levels were causally associated with altered gene expression.
(iii) For the three proteins with GWAS and EWAS associations (MATN3, MDGA1 and NEP), we wished to test whether methylation affected protein levels and/or whether protein levels affected methylation. We queried Phenoscanner to examine whether pQTLs for protein levels of MATN3, MDGA1 and NEP, identified in this study, were previously identified as methylation QTLs (mQTLs) for corresponding genome-wide significant CpG sites[49]. Methylation QTLs were used as IV to test whether changes in DNA methylation were causally associated with Olink® protein levels. Conversely, pQTLs were used as IV to determine whether altered protein levels were causally linked to differential methylation levels. Of note, as methylation of the 11 *cis* CpG sites associated with differential MDGA1 levels in our study are highly inter-correlated (Supplementary Fig. 3), we considered only the most significant *cis* CpG site (cg20053110) for MR analyses.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Full and openly accessible summary statistics from the association studies on Olink® neurology protein levels are available on the University of Edinburgh Datashare site (https://datashare.is.ed.ac.uk/). For GWAS data, see: https://datashare.is.ed.ac.uk/handle/10283/3366; https://doi.org/10.7488/ds/2580. For EWAS data, see: https://datashare.is.ed.ac.uk/handle/10283/3367; https://doi.org/10.7488/ds/2581.

Code availability

Code will be available from the authors on request.

Received: 6 March 2019 Accepted: 24 June 2019
Published online: 18 July 2019

References

1. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13**, 942–942 (2017).
2. Polivka, J., Polivka, J. Jr., Krakorova, K., Peterka, M. & Topolcan, O. Current status of biomarker research in neurology. *EPMA J.* **7**, 14–14 (2016).
3. Yao, C. et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).
4. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
5. Suhre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).
6. Kim, S. et al. Genome-wide association study of CSF biomarkers Abeta1-42, t-tau, and p-tau181p in the ADNI cohort. *Neurology* **76**, 69–79 (2011).
7. Kauwe, J. S. et al. Genome-wide association study of CSF levels of 59 alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. *PLoS Genet.* **10**, e1004758 (2014).
8. Sasayama, D. et al. Genome-wide quantitative trait loci mapping of the human cerebrospinal fluid proteome. *Hum. Mol. Genet.* **26**, 44–51 (2017).
9. Ahsan, M. et al. The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. *PLoS Genet.* **13**, e1007005–e1007005 (2017).
10. Saffari, A. et al. Estimation of a significance threshold for epigenome-wide association studies. *Genet. Epidemiol.* **42**, 20–33 (2018).
11. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
12. Gaunt, T. R. et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
13. Porcellini, E., Carbone, I., Ianni, M. & Licastro, F. Alzheimer's disease gene signature says: beware of brain viral infections. *Immun. Ageing* **7**, 16 (2010).
14. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0311-9 (2019).
15. Di Narzo, A. F. et al. High-throughput characterization of blood serum proteomics of IBD patients with respect to aging and genetic factors. *PLoS Genet.* **13**, e1006565–e1006565 (2017).
16. Lourdusamy, A. et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum. Mol. Genet.* **21**, 3719–3726 (2012).
17. Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
18. Yang, M. H. et al. Activity-dependent neuroprotector homeobox protein: A candidate protein identified in serum as diagnostic biomarker for Alzheimer's disease. *J. Proteom.* **75**, 3617–3629 (2012).
19. Bull, C., den Brok, M. H. & Adema, G. J. Sweet escape: sialic acids in tumor immune evasion. *Biochim. et. Biophys. Acta* **1846**, 238–246 (2014).
20. Shen, Y. et al. Focusing on long non-coding RNA dysregulation in newly diagnosed multiple myeloma. *Life Sci.* **196**, 133–142 (2018).
21. Glavey, S. V. et al. The sialyltransferase ST3GAL6 influences homing and survival in multiple myeloma. *Blood* **124**, 1765–1776 (2014).

22. Jandus, C. et al. Interactions between Siglec-7/9 receptors and ligands influence NK cell-dependent tumor immunosurveillance. *J. Clin. Investig.* **124**, 1810–1820 (2014).
23. Adams, O. J., Stanczak, M. A., von Gunten, S. & Laubli, H. Targeting sialic acid-Siglec interactions to reverse immune suppression in cancer. *Glycobiology* **28**, 640–647 (2018).
24. Conticello, C. et al. CD200 expression in patients with multiple myeloma: another piece of the puzzle. *Leuk. Res.* **37**, 1616–1621 (2013).
25. Zheng, J. et al. Recent developments in Mendelian randomization studies. *Curr. Epidemiol. Rep.* **4**, 330–345 (2017).
26. Gold, L. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE* **5**, e15004 (2010).
27. Hathout, Y. Proteomic methods for biomarker discovery and validation. Are we there yet? *Expert Rev. Proteom.* **12**, 329–331 (2015).
28. Deary, I. J. et al. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr.* **7**, 28–28 (2007).
29. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort profile update: the Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **47**, 1042–1042r (2018).
30. Shah, S. et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* **24**, 1725–1733 (2014).
31. Davies, G. et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996–1005 (2011).
32. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genom. Hum. Genet.* **10**, 387–406 (2009).
33. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
34. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinforma.* **13**, 86 (2012).
35. Chen, Y.-a et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
36. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
37. Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformarmatics* **21**, 3439–3440 (2005).
38. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protocol* **4**, 1184 (2009).
39. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
40. Tenenbaum, D. KEGGREST: Client-side REST access to KEGG. *R package version* **1** (2016).
41. Zhang, F. et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* **20**, 107 (2019).
42. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, s361–s363 (2012).
43. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
44. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
45. Võsa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 447367, https://doi.org/10.1101/447367 (2018).
46. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
47. Guo, H. et al. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum. Mol. Genet.* **24**, 3305–3313 (2015).
48. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, https://doi.org/10.7554/eLife.34408 (2018).
49. Staley, J. R. et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).

Acknowledgements

The authors thank all LBC1936 study participants and research team members who have contributed, and continue to contribute, to ongoing LBC1936 studies. The LBC1936 is supported by Age UK (Disconnected Mind program) and the Medical Research Council (MR/M01311/1). Methylation typing was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. This work was conducted in the Centre for Cognitive Ageing and Cognitive Epidemiology, which is supported by the Medical Research Council and Biotechnology and Biological Sciences Research Council (MR/K026992/1), and which supports I.J.D. We acknowledge NIH Grants R01AG054628 and R01AG05462802S1 for supporting this research and Grant P2CHD042849 for supporting the Population Research Center at the University of Texas. R.F.H. and A.J.S. 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]. A.S. is supported by a Medical Research Council PhD Studentship in Precision Medicine with funding by the Medical Research Council Doctoral Training Programme and the University of Edinburgh College of Medicine and Veterinary Medicine. D.L.Mc.C. and R.E.M. are supported by Alzheimer's Research UK major project grant ARUK-PG2017B−10. This research was supported by Australian National Health and Medical Research Council (grants 1010374, 1046880 and 1113400) and by the Australian Research Council (DP160102400). P.M.V., N.R.W. and A.F.M. are supported by the NHMRC Fellowship Scheme (1078037, 1078901 and 1083656).

Author contributions

Conception and design: R.F.H. and R.E.M. Data analysis: R.F.H., D.L.Mc.C., D.C.L. and R.E.M. Drafting the article: R.F.H. and R.E.M. Revision of the article: all authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11177-x.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.3 Conclusion

GWAS and EWAS prioritised possibly co-regulated networks of proteins for follow-up mechanistic studies to examine their role in cancer (multiple myeloma) or AD. The EWAS stage implicated neurology-related proteins in developmental, metabolic and inflammatory pathways. Integrating multiple lines of omics data identified molecular mechanisms through which pQTLs affect the plasma levels of five proteins.

MR analyses suggested that PVR levels causally associate with AD risk, but not vice versa. There was a high probability (>99%) that distinct causal variants within the *PVR* locus affect blood PVR levels and AD risk. Knockout animal models of PVR are required to examine its role in neurological phenotypes. Furthermore, the mechanisms that link pQTLs to the expression of PVR and other neurological protein biomarkers should be examined in *in vitro* systems.

In this chapter, I performed the first GWAS and EWAS on a panel of proteins enriched for their relevance to neurological conditions. GWAS and EWAS were conducted separately. In the next chapter, I model genetic and epigenetic data alone and together to probe the molecular correlates of 70 inflammation-associated proteins.

6 Genome-wide and epigenome-wide studies on inflammatory proteins

6.1 Introduction

Inflammation is a key feature of neurodegenerative disease states. Microglia or endothelial cells within the brain recruit peripheral immune cells, which execute detrimental and protective effects in the CNS (412). Peripheral inflammatory proteins associate with dementia and cognitive decline in epidemiological studies. For instance, higher blood levels of inflammatory proteins during mid-life were associated with cognitive decline over 20 years (413). Elevations in inflammatory proteins in middle age associated with brain lesions in older age (414). In a recent meta-analysis of 175 studies, the blood levels of 16 proteins were altered in AD patients when compared to controls. IL6 levels negatively correlated with MMSE scores (415). GWAS on AD risk implicate genes that encode proteins involved in the innate immune system including *CR1*, *CLU*, *CD33* and *TREM2* (82, 84, 85). Many lines of evidence implicate peripheral inflammation as a possible causal mechanism in cognitive decline and dementia. GWAS and EWAS on inflammatory proteins can help to elucidate whether associations between individual proteins and dementia risk are causal or might reflect confounding due to lifestyle factors.

In this chapter, I perform an integrated GWAS and EWAS on plasma levels of proteins present on the Olink Inflammation panel (n = 876, LBC1936). I use a novel Bayesian penalised regression framework termed BayesR+ that accounts for unknown confounders and intercorrelations between genetic and epigenetic data. I estimate the separate and combined contributions of common genetic and epigenetic variation towards inter-individual variability in inflammatory protein levels. In Chapter 4, I outline that BayesR+ outperforms mixed-effects models in estimating the variance in complex traits explained by genetic and methylation data. BayesR+ also showed reduced mean squared errors between true and simulated coefficients in single probe-regression when compared to linear and penalised regression models (133). Here, I assess the replication of pQTL and CpG associations across BayesR+, linear regression and mixed-effects models. I also apply MR analyses to probe causal relationships between 13 inflammatory proteins and the risk of AD.

This study was published in *Genome Medicine* (416) in July 2020 and is included in full in Section 6.2.

6.2 Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults

Genome Medicine

RESEARCH Open Access

Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults

Robert F. Hillary[1], Daniel Trejo-Banos[2], Athanasios Kousathanas[2], Daniel L. McCartney[1], Sarah E. Harris[3,4], Anna J. Stevenson[1], Marion Patxot[2], Sven Erik Ojavee[2], Qian Zhang[5], David C. Liewald[3], Craig W. Ritchie[6], Kathryn L. Evans[1], Elliot M. Tucker-Drob[7,8], Naomi R. Wray[5], Allan F. McRae[5], Peter M. Visscher[5], Ian J. Deary[3,4], Matthew R. Robinson[9*] and Riccardo E. Marioni[1*]

Abstract

Background: The molecular factors which control circulating levels of inflammatory proteins are not well understood. Furthermore, association studies between molecular probes and human traits are often performed by linear model-based methods which may fail to account for complex structure and interrelationships within molecular datasets.

Methods: In this study, we perform genome- and epigenome-wide association studies (GWAS/EWAS) on the levels of 70 plasma-derived inflammatory protein biomarkers in healthy older adults (Lothian Birth Cohort 1936; *n* = 876; Olink® inflammation panel). We employ a Bayesian framework (BayesR+) which can account for issues pertaining to data structure and unknown confounding variables (with sensitivity analyses using ordinary least squares- (OLS) and mixed model-based approaches).

Results: We identified 13 SNPs associated with 13 proteins (*n* = 1 SNP each) concordant across OLS and Bayesian methods. We identified 3 CpG sites spread across 3 proteins (*n* = 1 CpG each) that were concordant across OLS, mixed-model and Bayesian analyses. Tagged genetic variants accounted for up to 45% of variance in protein levels (for MCP2, 36% of variance alone attributable to 1 polymorphism). Methylation data accounted for up to 46% of variation in protein levels (for CXCL10). Up to 66% of variation in protein levels (for VEGFA) was explained using genetic and epigenetic data combined. We demonstrated putative causal relationships between CD6 and IL18R1 with inflammatory bowel disease and between IL12B and Crohn's disease.

Conclusions: Our data may aid understanding of the molecular regulation of the circulating inflammatory proteome as well as causal relationships between inflammatory mediators and disease.

* Correspondence: matthew.robinson@ist.ac.at; riccardo.marioni@ed.ac.uk
[9]Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria
[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK
Full list of author information is available at the end of the article

Background

Inflammation represents a concerted cascade of molecular and cellular events to combat infectious pathogens and endogenous insults. Inflammatory proteins are key mediators of defence and repair responses, and tight spatiotemporal regulation of their plasma concentrations permits effective immune activation and resolution [1]. Whereas acute inflammatory states may prompt severe illness and death, absence of resolution precipitates transition from acute to deleterious chronic inflammatory states [2]. Chronic inflammation facilitates the pathogenesis of various disease states, including diabetes, heart disease, stroke and allergic conditions [3]. Furthermore, inflammatory lesions in brain tissue are often associated with, and may contribute to, neurodegeneration and cognitive decline [4]. Globally, 60% of individuals will die as a consequence of a chronic inflammation-associated disease state [5]. Therefore, identifying biological factors which govern inter-individual variation in circulating inflammatory protein levels may allow for better prediction of individual disease risk and prognosis, and inform disease biology.

To date, a number of studies have aimed to characterise genetic factors associated with the levels of single inflammatory proteins or a small number of such proteins, including C-reactive protein, fibrinogen and interleukin-6 [6–25]. These genetic factors are also known as protein quantitative trait loci or pQTLs. Additionally, studies have examined the genetic architecture of panels of proteins, including inflammatory mediators, and have investigated co-regulatory pathways and associations with disease states [26–35]. Instead of using imputed genotype data, Höglund et al. used whole genome sequencing data to carry out genome-wide association studies (GWAS) on the levels of 72 inflammatory proteins. This led to the identification of 18 novel loci that were not identified using genotyped or imputed SNPs [36]. A number of studies have also carried out epigenome-wide association studies (EWAS) on the levels of a small set of inflammatory proteins, including C-reactive protein, interleukins-(1β, 4, 6, 9 and 10), interferon-gamma, transforming growth factor-beta and tumour necrosis factor [37–42]. Zaghlool et al. performed an EWAS of 1123 proteins, which pointed towards networks of chronic low-grade inflammatory biomarkers ($n = 944$ individuals) [43]. In an integrative approach, Ahsan et al. aimed to identify genetic and epigenetic markers associated with protein biomarkers including inflammatory mediators ($n \leq 1033$ individuals) [44]. No study has modelled GWAS and EWAS both as stand-alone association studies and in a combined analysis in the context of proteomic data. This would allow for the identification of genetic and epigenetic correlates of inflammatory protein levels and for the estimation of variance in protein levels explained by genetic and epigenetic data, considered in isolation but also conditioned on one another to reflect reciprocal influences of these molecular data types. Here, we triangulate results from multiple statistical approaches to provide a robust set of genetic and epigenetic correlates of inflammatory protein levels.

Notably, most studies examining the molecular architecture of human traits have relied on linear model-based methods which examine marker or probe effects marginally [45, 46]. A number of issues may arise when using linear regression-based methods and if these are not addressed in the study design, it may lead to model overfitting and biased estimation of effect sizes. These potential issues include correlation structure within molecular datasets, data structure (i.e. cellular heterogeneity, batch effects) and omitted variable bias [47]. Several approaches have been proposed to address these issues [47–53] and these encompass strategies which permit the joint and conditional estimation of effect sizes whilst accounting for correlations among markers and confounding variables. Here, we consider a Bayesian penalised regression framework termed BayesR+ which was developed to assess genetic and epigenetic architectures of complex traits [54]. In BayesR+, marker effects (SNP or CpG site) can be estimated jointly whilst controlling for data structure and correlations among molecular markers of different types. Indeed, this method permits the estimation of variance explained in the trait by all methylation probes or genetic markers, either separately or together. BayesR+ has been shown to outperform single-probe linear regression and penalised regression approaches, such as ridge and LASSO, in relation to the correlation of estimated effects with true simulated values as well as mean squared errors between true and estimated coefficients for single-probe regression. Additionally, BayesR+ shows a higher correlation between estimated effects for variance explained by genetic and epigenetic markers in phenotypic traits and true simulated values when compared to a mixed model strategy in both sparse and non-sparse marker settings [54].

In the present study, we use the BayesR+ method (and sensitivity analyses using ordinary least squares (OLS) [55, 56] and mixed model methods [57]) to examine both the genetic and epigenetic architectures of 70 blood inflammatory proteins in 876 relatively healthy older adults from the Lothian Birth Cohort 1936 study (mean age 69.8 ± 0.8 years; levels adjusted for age, sex, population structure and array plate). Hereinafter, we refer to the adjusted inflammatory protein levels as protein levels. These proteins are present on the Olink® inflammation panel and comprise a mixture of proteins with defined functions pertinent to human inflammatory pathways as well as putative roles in inflammation-related disease states. We use priors guided by results from previous genome-wide and epigenome-wide studies [54, 58] for the expected variance explained in circulating protein levels by genetic and epigenetic factors. Applying a stringent approach, we only consider markers or probes that were identified across all methods employed as being associated with a given protein (concordantly identified) and integrate multiple levels of 'omics' data to investigate mechanisms by which genetic variants may influence protein levels. Finally, we use our GWAS summary data to test for putatively causal relationships between inflammatory protein biomarkers and neurological or inflammatory disease states. Thus, this paper has two major aims. The first aim is to provide robust and novel estimates for the contribution of genetic and epigenetic factors towards inter-individual variation in circulating inflammatory protein concentrations. The relationships between genetic and epigenetic factors with inflammatory proteins levels are modelled both alone and together. The second aim is to provide the first use of multiple statistical methods in performing genome-wide and epigenome-wide association studies of human proteomic data.

Methods

The Lothian Birth Cohort 1936

The Lothian Birth Cohort 1936 (LBC1936) study is a longitudinal study of ageing. Cohort members were all born in 1936 and most took part in the Scottish Mental Survey 1947 at age 11 years. Participants who were living mostly within the Edinburgh area were re-contacted approximately 60 years later ($n = 1091$, recruited at mean age 70 years). Recruitment and testing of the LBC1936 cohort have been described previously [59, 60].

Protein measurements in the Lothian Birth Cohort 1936

Plasma was extracted from 1047 blood samples and collected in lithium heparin tubes at mean age 69.8 ± 0.8 years. Following quality control, 1017 samples remained. Plasma samples were analysed using a 92-plex proximity extension assay (Olink® Bioscience, Uppsala Sweden). One protein from the panel, BDNF, failed quality control and was removed from the study. For a further 21 proteins, over 40% of samples fell below the lowest limit of detection. These proteins were removed from analyses leaving a final set of 70 proteins. The proteins assayed comprise the Olink® inflammatory biomarker panel. Briefly, 1 μL of sample was incubated in the presence of proximity antibody pairs linked to DNA reporter molecules. Upon appropriate antigen-antibody recognition, the DNA tails form an amplicon by proximity extension which is quantified by real-time PCR. Data pre-processing was performed by Olink® using NPX Manager software. Protein levels were transformed by rank-based inverse normalisation and regressed onto age, sex, four genetic principal components of ancestry and array plate. Standardised residuals from these regression models were brought forward for all genetic-protein and epigenetic-protein analyses. Pre-adjusted protein level distributions are presented in Additional file 1. Associations between pre-adjusted protein levels and biological as well as technical covariates are detailed in Additional file 2: Table S1.

Genome-wide association studies

LBC1936 DNA samples were genotyped at the Edinburgh Clinical Research Facility using the Illumina 610-Quadv1 array ($n = 1005$; mean age 69.6 ± 0.8 years; San Diego). Quality control procedures for genetic data are detailed in Additional file 3.

BayesR+ is a software implemented in C++ for performing Bayesian penalised regression on complex traits [54]. The joint and conditional effects of typed SNPs ($n = 521,523$ variants) on transformed protein levels were examined. The prior distribution is specified as a mixture of Gaussian distributions, corresponding to effect sizes of different magnitude, and a discrete spike at zero which enables the omission of probes and markers with negligible effect on the phenotype. Informed by data from our previous pQTL study [58], mixture variances for genetic data were set to 0.01 and 0.1 for the stand-alone BayesR+ GWAS. In the combined analysis with epigenetic data, owing to the need for the same number of mixture variances for genetic and epigenetic data in the BayesR+ software, mixture variances were set to 0.01, 0.1 and 0.2. Input data were scaled to mean zero and unit variance, and adjusted for age and sex. To obtain estimates of effect sizes, Gibbs sampling was used to sample over the posterior distribution conditional on the input data. The Gibbs algorithm consisted of 10000 samples and 5000 samples of burn-in after which a thinning of 5 samples was utilised to reduce autocorrelation. Genetic markers which exhibited a posterior inclusion probability of ≥ 95% were deemed to be significant.

Details for the OLS regression model approach are outlined in Additional file 3. In the linear method, markers which surpassed a Bonferroni-corrected conditional significance threshold of 7.14×10^{-10} (= genome-wide significance $5.0 \times 10^{-8}/70$ phenotypes) were considered. The genome-wide significance level of 5.0×10^{-8} was selected as per convention in GWAS studies.

Epigenome-wide association studies

DNA from whole blood was assessed using the Infinium 450 K methylation array at the Edinburgh Clinical Research Facility ($n = 876$; mean age 69.8 ± 0.8 years). Quality control procedures for methylation data are detailed in Additional file 3.

Using BayesR+, prior mixture variances for methylation data ($n = 459{,}309$ CpG sites) were set to 0.001, 0.01 and 0.1. Age, sex and Houseman-estimated white blood cell proportions [61] were incorporated as fixed effect covariates. The same settings as in the genetic analyses were applied. Methylation probes which had a posterior inclusion probability of ≥ 95% were deemed to be significant.

Details for the OLS and mixed linear model approaches are outlined in Additional file 3. For these methods, probes which surpassed a Bonferroni-corrected significance threshold of 5.14×10^{-10} (= genome-wide significance 3.6×10^{-8}/70 phenotypes) were deemed to be significant. The genome-wide significance level of 3.6×10^{-8} was selected as per the recommendations of Safarri et al. [62].

Functional annotation of genetic and epigenetic loci

Genetic markers that were independently associated with protein levels were functionally annotated using ANNOVAR [63] and Ensembl genes (build 85) in FUMA (*FU*nctional *M*apping and *A*nnotation) [64]. Epigenetic probes associated with protein levels were annotated using the *IlluminaHumanMethylation450kanno.ilmn12.hg19* package [65].

Identification of overlap between *cis* pQTLs and *cis* eQTLs

To determine whether pQTL variants may affect protein levels through modulation of gene expression, we cross-referenced *cis* pQTLs with publicly available (and FDR-corrected significant) *cis* expression QTL (eQTL) data from the eQTLGen consortium. Expression QTL data were derived from blood tissue, 85% of samples were derived from whole blood and 15% of samples were derived from peripheral blood mononuclear cell data [66]. For each protein, expression QTLs were also subset to the gene (messenger RNA) encoding the protein of interest.

Colocalisation

To test whether a sole causal variant might underlie both an eQTL and pQTL association, we performed Bayesian tests of colocalisation using the *coloc* package in R [67]. For each protein of interest, a 200-kb region (upstream and downstream—recommended default setting) surrounding the appropriate pQTL was extracted from our GWAS summary statistics [68]. For each respective protein, the same region was also extracted from eQTLGen summary statistics. Default priors were applied. Summary statistics for all SNPs within these regions were used to determine the posterior probability for five distinct hypotheses: a single causal variant for both traits, no causal variant for either trait, a causal variant for one of the traits (encompassing two hypotheses), or distinct causal variants for the two traits. Posterior probabilities (PP) ≥ 0.95 provided strong evidence in favour of a given hypothesis.

Pathway enrichment and tissue specificity analyses

Using methylation data, pathway enrichment was assessed among KEGG pathways and Gene Ontology (GO) terms through hypergeometric tests using the *phyper* function in R. All gene symbols from the 450 K array annotation (null set of sites) were converted to Entrez IDs using *biomaRt* [69, 70]. GO terms and their corresponding gene sets were retrieved from the Molecular Signatures Database (MSigDB)-C5 [71]. KEGG pathways were downloaded from the KEGG REST server [72]. Tissue specificity analyses were performed using the GENE2FUNC function in FUMA. Differentially expressed gene sets with Bonferroni-corrected P values < 0.05 and an absolute log-fold change of ≥ 0.58 (default settings) were considered to be enriched in a given tissue type (GTEx v7).

Mendelian randomisation

Two-sample Mendelian randomisation was used to test for putatively causal relationships between (i) the 4 proteins whose pQTLs were previously shown to be associated with human traits, as identified through GWAS Catalog, and the respective traits [73, 74] (http://www.nealelab.is/uk-biobank/); (ii) the 13 proteins which harboured significant pQTLs and Alzheimer's disease risk [75]; (iii) gene expression and inflammatory protein levels; and (iv) DNA methylation and inflammatory protein levels. Pruned variants (LD $r^2 < 0.1$) were used as instrumental variables (IV) in MR analyses. In tests where only one independent SNP remained after LD pruning, causal effect estimates were assessed using the Wald ratio test, i.e. a ratio of effect per risk allele on trait to effect per risk allele on protein levels. In tests where multiple independent variants were identified, and if no evidence of directional pleiotropy was present (non-significant MR-Egger intercept), multi-SNP MR was carried out using inverse variance-weighted estimates. Analyses were conducted using MRbase [76]. Further details are provided in Additional file 3.

Results

Genome-wide studies of inflammatory protein levels

In a Bayesian penalised regression model (BayesR+), 16 pQTLs were identified for 14 proteins (Additional file 2: Table S2). Thirteen of these 16 pQTLs ($n = 13$ proteins) directly, or through variants in high linkage disequilibrium (LD) $r^2 > 0.75$, replicated conditionally significant pQTLs from the OLS regression model (Additional file 2: Tables S3-S5; Additional file 3). The correlation

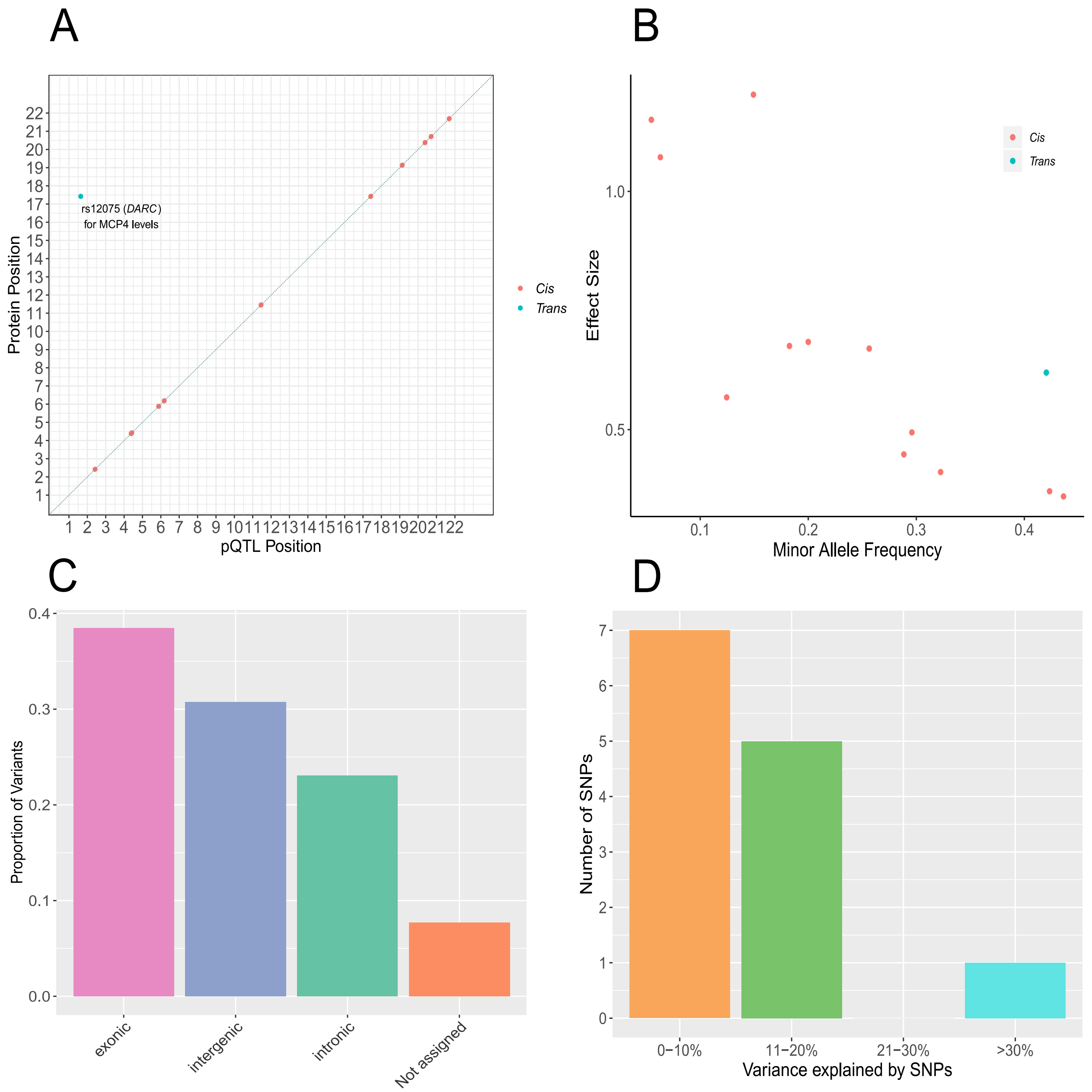

Fig. 1 Genetic architecture of inflammatory protein biomarkers in the Lothian Birth Cohort 1936. **a** Chromosomal locations of pQTLs concordant between Bayesian penalised and ordinary least squares regression models for genome-wide association studies ($n = 13$ pQTLs). The *x*-axis represents the chromosomal location of concordantly identified *cis* and *trans* SNPs associated with the levels of Olink® inflammatory proteins. The *y*-axis represents the position of the gene encoding the associated protein. The sole conditionally significant concordant *trans* association is annotated. *Cis* (red circles); *trans* (blue circles). **b** Absolute effect size (per standard deviation of difference in protein level per effect allele) of pQTLs versus minor allele frequency. *Cis* (red circles); *trans* (blue circles). **c** Classification of 13 pQTLs by function as defined by functional enrichment analysis in FUMA. **d** Variance in protein levels explained by pQTLs (estimates from Bayesian penalised regression are displayed)

structure among these 13 proteins is shown in Additional file 4: Fig. S1.

Twelve (92.3%) of the concordant SNPs were *cis* pQTLs (SNP within 10 Mb of the transcription start site (TSS) of a given gene [69, 70]) and 1 pQTL (7.7%) was a *trans*-associated variant (Fig. 1a; Additional file 2: Table S6). There was an inverse relationship between the minor allele frequency of variants and their effect size (Fig. 1b). The functional category to which the greatest proportion of variants was assigned was exonic variants (38.5%), as identified by FUMA (*FU*nctional *M*apping and *A*nnotation analysis) (Fig. 1c). Four of the five SNPs

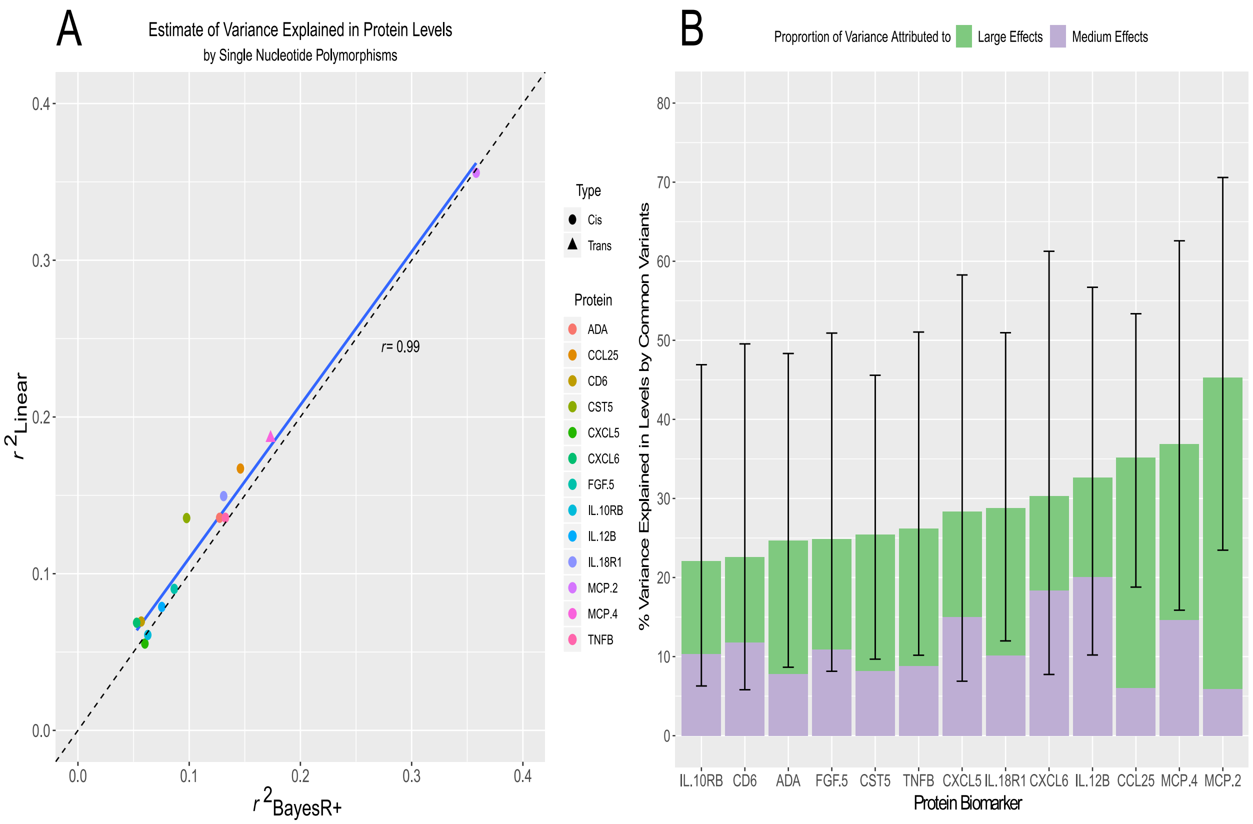

Fig. 2 Variance in circulating inflammatory protein levels explained by common genetic variation. **a** In this panel, the variance explained (r^2) by consensus SNPs ($n = 13$ SNP, 1 per protein) in the ordinary least squares regression model was compared against the variance explained by the same SNP set identified in the Bayesian penalised regression approach. **b** The proportion of variance explained in Olink® inflammatory protein levels by common genetic variants genotyped in the LBC1936 participants is shown. Only those proteins which had significant pQTL associations in both the ordinary least squares and Bayesian methods are presented ($n = 13$). Additionally, the proportion of variance explained attributable to medium effects (prior: variance of 1% explained) and large effects (prior: variance of 10% explained) are demonstrated in purple and green, respectively. Error bars represent 95% credible intervals

annotated to exonic regions produce missense mutations. From the Bayesian model, pQTLs explained between 5.28% (rs10005565; CXCL6) and 35.80% (rs3138036; MCP2) of inter-individual variation in protein levels (Fig. 1d). The estimates for variance accounted for in protein levels by single SNPs were correlated 99% between the BayesR+ and OLS regression models (Fig. 2a; Additional file 2: Table S6). The BayesR+ common (minor allele frequency > 1%) SNP-based heritability estimates ranged from 11.4% (CXCL9; 95% credible interval [0%, 43.5%]) to 45.3% (MCP2; 95% credible interval: [23.5%, 70.6%]), with a mean estimate of 20.2% across the 70 proteins (Additional file 2: Table S7). Figure 2b shows heritability estimates for the 13 proteins exhibiting concordantly identified pQTLs across OLS regression and Bayesian approaches. Figure 3 demonstrates the effect of genetic variation at the most significant *cis* pQTL (rs3138036; MCP2) and the sole *trans* pQTL (rs12075; MCP4) on protein levels.

There was a strong correlation between our SNP-based heritability estimates and those from a previous study of 961 individuals [44]: 29 overlapping proteins, *r* 0.71, 95% CI [0.43, 0.84] (Additional file 2: Table S8 and Additional file 4: Fig. S2).

Molecular mechanisms underlying pQTLs: colocalisation analysis

Of the 12 *cis* pQTLs which were identified across OLS regression and BayesR+, 8 SNPs (66.67%) previously have been identified as *cis*-acting expression QTLs (eQTLs) in blood (Additional file 2: Table S9). Using *coloc* [67], we tested the hypothesis that one causal variant might underlie both a pQTL and eQTL for each protein. For 4/8 proteins, there was strong evidence (posterior probability (PP) > 0.95) for colocalisation of *cis* pQTLs and *cis* eQTLs (Additional file 2: Table S10). These proteins were CCL25, CD6, CXCL5 and CXCL6.

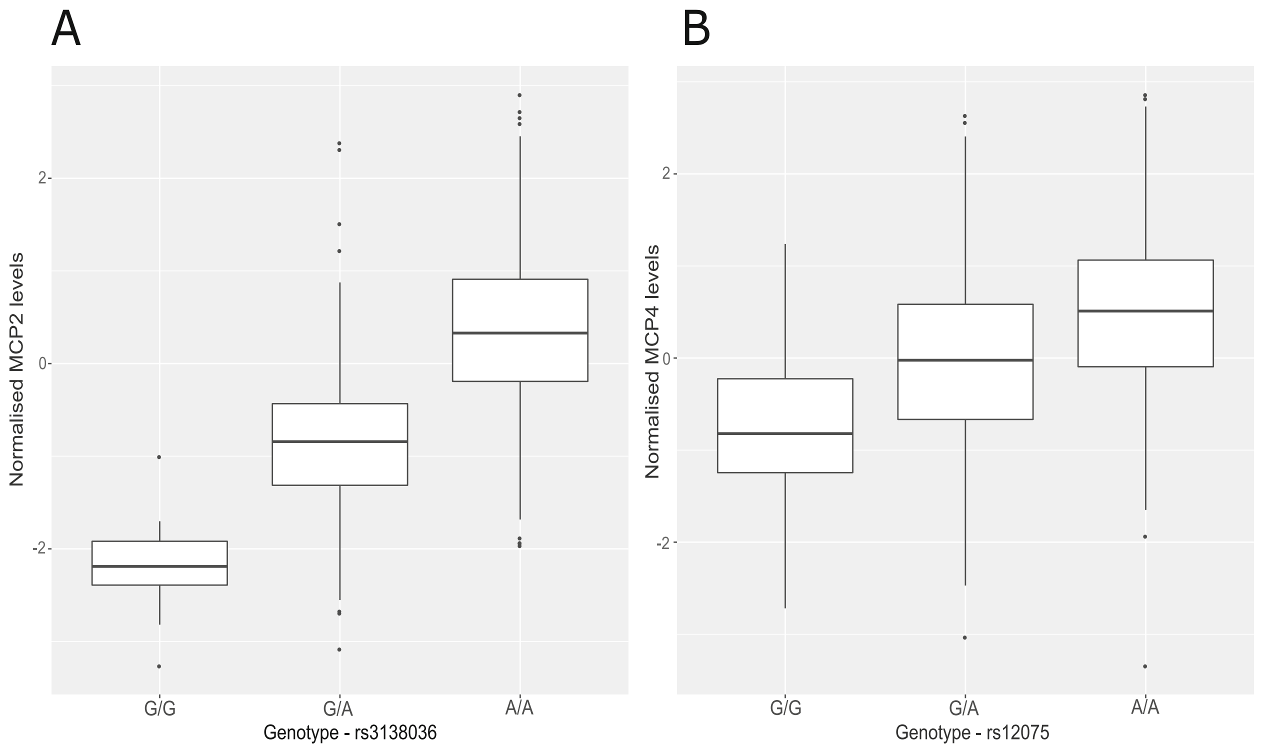

Fig. 3 Effect of genetic variation on inflammatory protein levels. **a** Box plot of MCP2 levels as a function of genotype (rs3138036, effect allele: G, other allele: A, beta = – 1.20, se = 0.06). **b** Box plot of MCP4 levels as a function of genotype (rs14075, effect allele: G, other allele: A, beta = – 0.62, se = 0.05). Centre line of boxplot: median, bounds of box: first and third quartiles

Mendelian randomisation analyses (MR; see the 'Methods' section) indicated that altered gene expression was causally associated with changes in protein levels for each of the four aforementioned proteins (CCL25, CD6, CXCL5 and CXCL6; range of beta [0.68, 12.25], se [0.09, 1.12], P [9.54×10^{-7}, 1.05×10^{-37}]). However, a second colocalisation approach termed Sherlock [77] suggested that, from the 13 proteins with concordantly identified pQTLs, only expression of *ADA*, *CXCL5* and *IL18R1* were associated with levels of their respective protein products (Additional file 2: Table S11; Additional file 3).

Epigenome-wide studies of inflammatory protein levels

In the Bayesian model, 8 CpG-protein associations ($n = 8$ proteins) had a posterior inclusion probability of more than 95% (Additional file 2: Table S12). Five of these associations overlapped with those identified by the OLS regression model ($P < 5.14 \times 10^{-10}$; Additional file 2: Table S13); three of which were also identified in the mixed model approach ($P < 5.14 \times 10^{-10}$; Additional file 2: Table S14). These were the smoking-associated probe cg05575921 for CCL11 levels (*trans* association at *AHRR*; mixed model—beta – 1.97, se 0.32, P 4.86×10^{-10}), cg07839457 for CXCL9 levels (*trans* association at *NLRC5*; beta – 2.91, se 0.39, P 8.03×10^{-14}) and cg03938978 for IL18R1 levels (*cis* association at *IL18RAP*; beta – 1.37, se 0.16, P 5.86×10^{-17}) (Additional file 2: Table S14). Adjustment for smoking attenuated the association between CCL11 levels and the cg05575921 probe (linear model—before adjustment: beta – 1.74, P 2.68×10^{-10}, after adjustment: beta – 1.20, P 0.03; % attenuation 31.03%). GWAS and EWAS of CCL11 levels were repeated adjusting for smoking status, the results of the association studies are detailed in Additional file 3. Figure 4 depicts an epigenetic map of CpG-protein associations within this study and demonstrates the degree of overlap between methodologies. The correlation among the three proteins with concordantly identified CpG associations is shown in Additional file 4: Fig. S3. Look-up analyses of the top GWAS and EWAS findings with those reported in the literature are detailed in Additional file 3. For the GWAS, 11/13 pQTLs (84.62%) from the present study were previously reported in the literature. The two loci which represent novel pQTLs are rs11700291 (ADA) and rs1458038 (FGF-5). Beta coefficients displayed a correlation coefficient of 0.88 between those in the present study and those reported in previous studies. For the EWAS, only one of the three concordantly identified

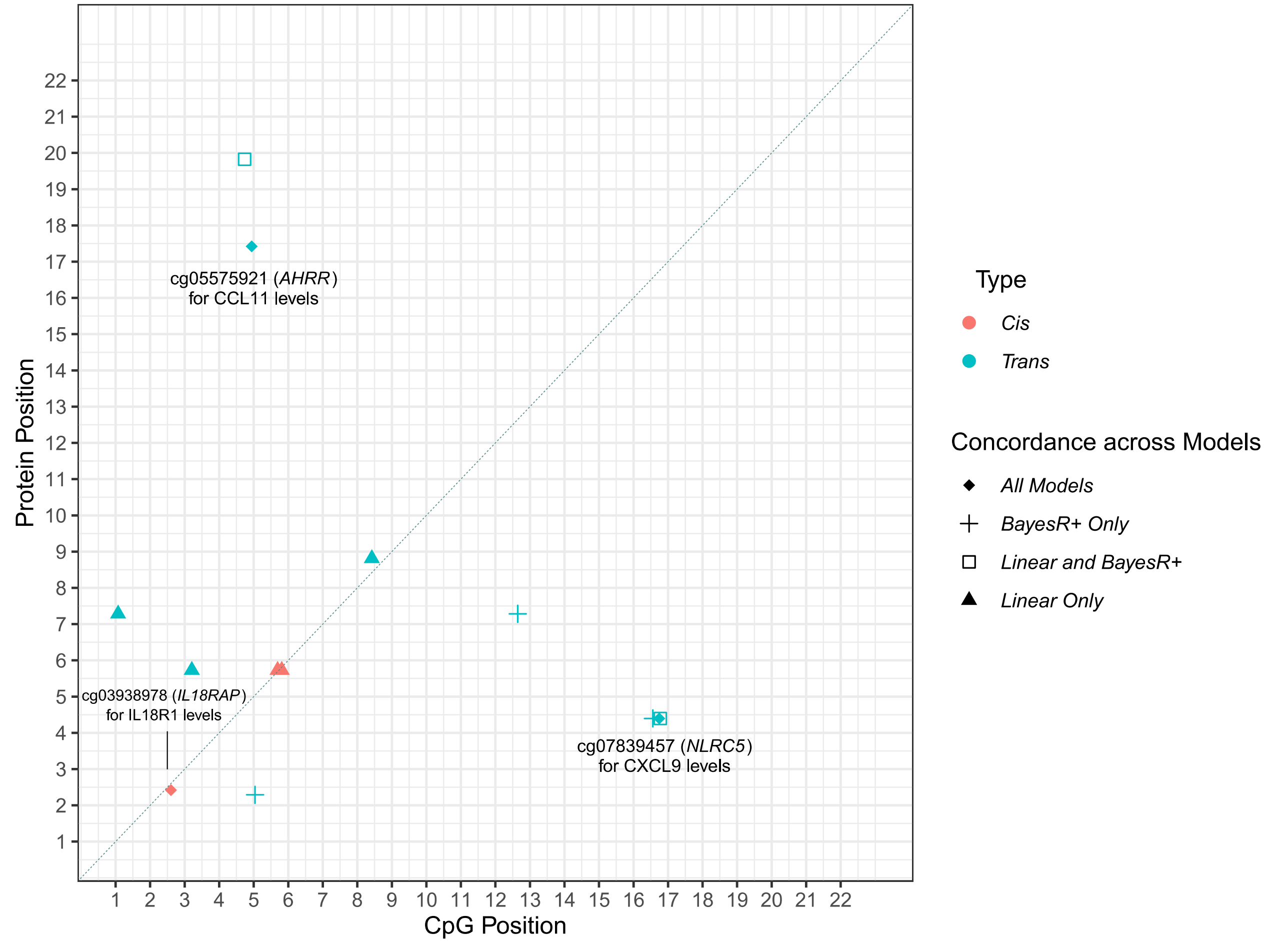

Fig. 4 Genomic locations of CpG sites associated with differential inflammatory protein levels. The *x*-axis represents the chromosomal location of CpG sites associated with the levels of Olink® inflammation biomarkers. The *y*-axis represents the position of the gene encoding the associated protein. The level of concordance across three models used to perform epigenome-wide association studies is represented by different shape patterns. Those CpG sites (*n* = 3) which were identified by linear (ordinary least squares), mixed model and Bayesian penalised regression models, and passed a Bonferroni-corrected significance threshold are represented by diamonds and annotated. Three proteins (CXCL9, CXCL10 and CXCL11) were associated with differential methylation levels at the cg07839457 site in the *NLRC5* transcription factor locus. Additionally, two proteins (CCL11 and TGF-alpha) were associated with the smoking-associated cg05575921 site in the *AHRR* locus. *Cis* (red); *trans* (blue)

CpG-protein associations was previously reported in the literature by Ahsan et al. [44]. This association was between the cg07839457 probe (*NLRC5*) and CXCL9 levels (beta_{LBC} – 2.91 vs. $\text{beta}_{\text{Ahsan}}$ – 3.26).

We conducted tissue specificity and pathway enrichment analyses based on genes identified by EWAS for each of the 3 proteins with significant CpG associations. Tissue-specific patterns of expression were observed for 2/3 proteins (Additional file 4: Fig. S4-S6). For CCL11, differential expression was observed in breast, adipose and kidney tissue. For IL18R1, differential expression of associated genes was observed in pancreatic tissue. Furthermore, down-regulation of genes associated with IL18R1 was observed in the hippocampus and substantia nigra. There was no significant enrichment of pathways incorporating genes annotated to CXCL9, CCL11 or IL18R1 following multiple testing correction.

One protein, IL18R1, harboured both a significant *cis* pQTL and *cis* CpG site in our study (Additional file 4: Fig. S7). This SNP (rs917997) previously has been identified as a methylation QTL (mQTL) for the single *cis* CpG site associated with IL18R1 levels identified by our epigenome-wide studies (cg03938978) [78]. Using bidirectional MR analysis (Wald ratio test; see methods), we show evidence that DNA methylation at this locus may be causally associated with circulating IL18R1 levels (beta – 0.81, se 0.17, P 2.14 × 10^{-33}). Conversely, IL18R1 levels may also be causally associated with altered DNA methylation (beta – 1.22, se 0.16, P 3.4 × 10^{-14}).

The methylation data explained an average of 18.2% of variance in protein levels using BayesR+; estimates ranged from 6.3% (IL15RA, 95% credible interval [0.0%, 27.3%]) to 46.1% (CXCL10, 95% credible interval [24.1%,

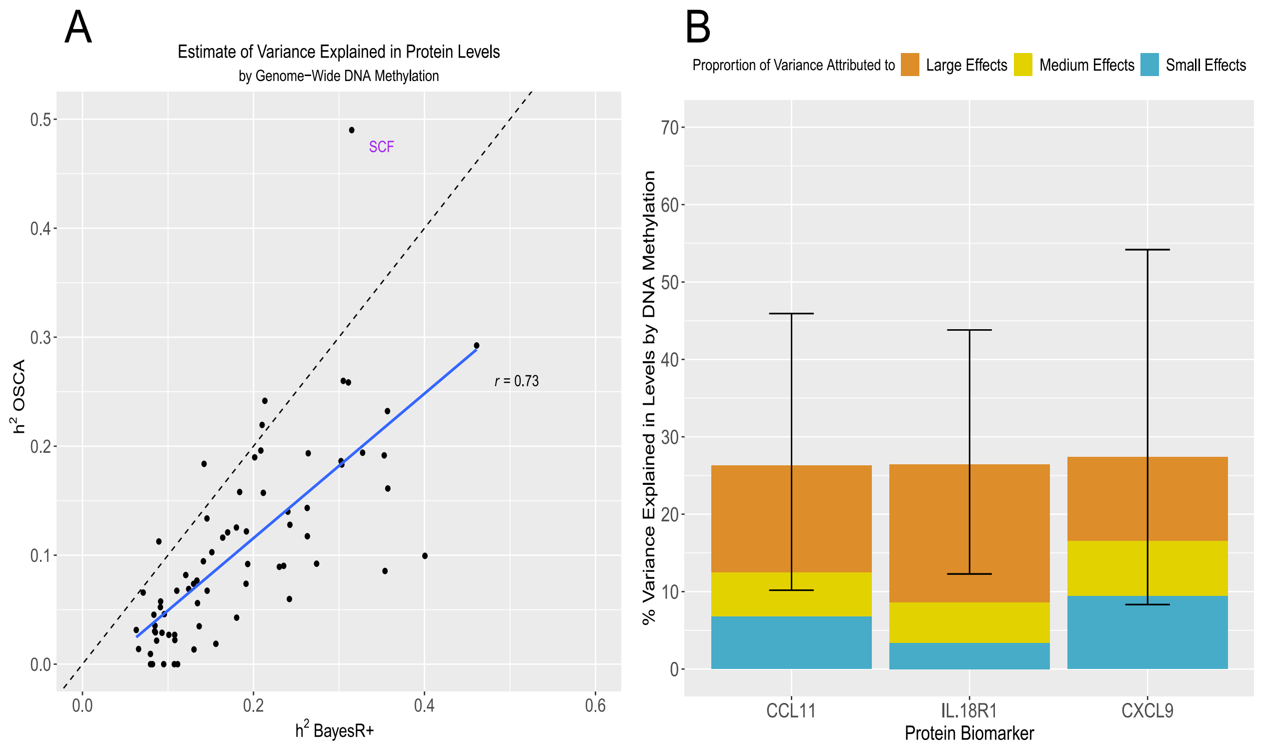

Fig. 5 Variance in circulating inflammatory protein levels explained by DNA methylation. **a** In this panel, the variance explained in circulating protein levels by complete methylation data from sites present on the Infinium 450 K methylation array was examined. A comparison between variance explained (h^2) by a mixed model approach (OSCA) and a Bayesian penalised regression approach (BayesR+) is shown. **b** The proportion of variance explained in Olink® inflammatory protein levels by DNA methylation, as estimated by BayesR+ is shown. Only those proteins (*n* = 3) which had significant CpG associations in ordinary least squares, mixed model and Bayesian methods are presented. Additionally, the proportion of variance explained attributable to small effects (prior: variance of 0.1% explained), medium effects (prior: variance of 1.0% explained) and large effects (prior: variance of 10% explained) are demonstrated in blue, gold and dark orange, respectively. Error bars represent 95% credible intervals

67.1%]) (Additional file 2: Table S15). There was strong concordance with estimates from the mixed model sensitivity analysis (Additional file 2: Table S16 and Fig. 5a). Figure 5b shows the variance explained by methylation data for the 3 proteins exhibiting concordantly identified CpGs across OLS regression, mixed-model and Bayesian approaches.

Variation in inflammatory protein levels explained by genetics and DNA methylation

When accounting for genetic data, the estimates for variance explained by methylation data were largely unchanged for most proteins (Additional file 2: Table S17; *n* = 9 proteins with change > 5%, 1 with change < − 5% (VEGFA)). The mean absolute change was 2.6% (minimum 0.01% for TNFRSF9 and maximum 15.0% for IL18R1). Similarly, estimates from genetic data were largely unchanged in the combined analysis (*n* = 2 proteins with change > 5%). The mean absolute change was 1.8% (minimum 0.02% for CD244 and maximum 6.7% for CCL28). For 22 proteins, the variance explained by methylation data was greater than that explained by genetic data (Additional file 5).

For each protein, we performed *t*-tests to determine whether the variance explained by methylation or genetic data alone was significantly different from the estimate for variance explained in the combined analysis. For methylation data, 40 proteins showed a significant difference between the estimates for variance in protein levels explained by methylation data alone and methylation data conditional on SNPs ($P < 0.05$). For genetic data, 50 proteins showed a significant difference ($P < 0.05$) (Additional file 2: Table S17).

The combined estimate for variance explained by genetic and methylation data ranged from 23.4% for CXCL1 to 66.4% for VEGFA. The mean and median estimates were 37.7 and 36.0%, respectively. Details of which SNPs and CpGs were identified as being associated with protein levels in the combined BayesR+ analyses, accounting for all genetic and epigenetic factors together, is outlined in Additional file 2: Table S18 and Additional file 3.

Evaluating causal associations between inflammatory biomarkers and human traits

The 13 independent pQTL associations were queried against GWAS Catalog to identify existing associations between these pQTLs and phenotypes [73]. We investigated whether these associations represented causal relationships. Using two-sample MR, we showed that CD6 levels were causally associated with inflammatory bowel disease (IBD) (beta 0.20, se 0.04, P 2.59×10^{-6}). Furthermore, FGF-5 levels were causally associated with systolic and diastolic blood pressure (beta 0.07 and 0.07, se 0.01 and 0.01, P 1.04×10^{-34} and 4.29×10^{-42}, respectively). IL12B levels were associated with Crohn's disease (beta 0.42, se 0.05, P 2.76×10^{-15}). Circulating IL18R1 levels showed a causal relationship with IBD (beta 0.17, se 0.03, P 1.63×10^{-9}).

Peripheral inflammatory processes and proteins have been linked to risk of late-onset Alzheimer's disease (AD) [79, 80]. We tested whether the 13 proteins with significant genetic correlates in our study were causally associated with AD risk (Additional file 3). One protein, IL18R1, showed a nominally significant, unidirectional relationship with AD risk (beta 0.02, se 0.01, P 0.04) (Additional file 2: Table S19).

Discussion

Using a Bayesian framework and sensitivity analyses with OLS regression and mixed linear models, we robustly identified 13 independent genetic and 3 epigenetic correlates of circulating inflammatory protein levels. Two of these pQTLs and two CpG sites have not been previously reported as genome-wide significant in the literature. This is the first study to have integrated genetic and epigenetic data together using multiple methods to identify molecular correlates of, and estimate the contribution of these molecular factors towards inter-individual variability in, the circulating proteome. Our results also provide an important and novel demonstration of the overlap between disparate methodologies for performing genome-wide and epigenome-wide association studies on proteomic data. Using integrative causal frameworks, we identified mechanisms through which genetic variation may perturb plasma protein levels. Additionally, we demonstrated causal relationships between prioritised circulating inflammatory proteins and blood pressure as well as inflammatory bowel diseases.

For genome-wide association studies, there is a necessity to perform secondary analyses in order to identify independent loci from association studies. This is often carried out through employing conditional and joint analyses (GCTA-COJO) or LD clumping-based methods, such as those implemented in FUMA [54, 64]. BayesR+ negates the need for such secondary analyses; it allows for the modelling of single marker or probe effects whilst controlling for all other markers or probes. Indeed, BayesR+ can outperform OLS regression or mixed model methods in providing single probe or marker coefficient estimates whilst controlling for all other input SNP and/or CpG sites, as well as known and unknown confounding variables. However, identifying true molecular correlates of protein data over false positive associations is challenging. By relying on careful corrections for multiple testing and triangulation of evidence across disparate methods, our stringent approach was well-equipped to identify likely true biological signal as opposed to false positives.

The issue of identifying true biological signals over false positive associations is particularly pertinent in relation to *trans* associations which show poor replication and often have smaller effect sizes than *cis* associations [81]. We identified one *trans* pQTL (rs12075) associated with levels of the chemokine MCP4 (encoded for by *CCL13* gene on chromosome 17). This SNP represents a nonsynonymous polymorphism (Asp42Gly) annotated to the *Duffy antigen/chemokine receptor* (*DARC*) gene on chromosome 1. Previously, this SNP has been associated with lower MCP1 levels and evidence shows that the base-change results in altered chemokine-receptor binding [10, 20, 82]. Additionally, this polymorphism has been shown to explain approximately 20% of variation in MCP1 levels, similar to our estimate of 18.66% in MCP4 levels [82]. The Duffy antigen receptor is expressed on erythrocytes and acts as a reservoir for circulating chemokines resulting in reduced distribution of chemokines to extravascular tissue and dampened pro-inflammatory effects [83]. Our findings suggest that this polymorphism may also lead to reduced MCP4 levels, possibly through augmented chemokine-receptor interaction.

In the EWAS analyses, the probe cg05575921, located in the *AHRR* locus, was associated with CCL11 levels. This probe is strongly associated with smoking status [84–91] and the association was attenuated after adjustment for smoking. Furthermore, higher levels of CCL11 have been associated with tobacco smoking and cannabis use [92–94]. We also found altered methylation at the *NLRC5* locus (*NOD-like receptor family CARD domain containing 5*) is associated with circulating CXCL9 levels. NLRC5 acts as a potent regulator of the inflammasome [44, 95]. Zaghlool et al. showed that altered methylation at the *NLRC5* locus associates with several inflammatory markers, including CXCL10 and CXCL11, with pathway analyses linking it to disease states in which NLRC5 dysfunction is implicated such as cancer and cardiovascular disease [43].

Using our database of genotype-protein associations, we tested for causal relationships between inflammatory protein biomarkers and human phenotypes. However, in each case, only one variant was available to test for such associations which does not allow for the testing of pleiotropic effects. CD6 was associated with clinically diagnosed IBD. Expression of the CD6 receptor and its ligand, ALCAM, are overexpressed in the intestinal mucosa of IBD patients where it may promote $CD4^+$ T cell proliferation and differentiation into pro-inflammatory Th1/Th17 cells [96]. FGF-5 levels were associated with automated readings of systolic and diastolic pressure; previously, FGF-5 levels have been significantly correlated with blood pressure [97]. Variation in the *IL12B* gene has been linked strongly to the pathogenesis of Crohn's disease and an antibody targeted towards the p40 subunit of IL12 demonstrated efficacy in the treatment of moderate-to-severe Crohn's disease [98]. In our study, we showed that circulating IL12B levels may be causally linked to this disease. Lastly, IL18R1 levels may also be causally associated with IBD. A number of studies have demonstrated that increased IL18 signalling confers detrimental effects in the context of gastrointestinal inflammatory processes [99].

Our study has a number of caveats. First, proteins with high sequence homology and structural similarities to a targeted protein of interest may be inappropriately captured by assay probes resulting in quantification errors. Olink®'s Proximity Extension Assay technology uses a matched pair of antibodies, coupled to unique, partially complementary oligonucleotides resulting in exceptional readout specificity and greatly reducing this problem compared to other immunoassays. Second, there was a strong correlation structure among the inflammatory protein panel. However, given that inflammatory proteins are often co-expressed and synergistic, overlapping loci may reveal biologically important foci or nodes of co-regulation [100]. Third, functional enrichment analyses indicated that four robustly identified pQTL signals reflect missense mutations in their protein products, three of which were *cis* associations with proteins present on the Olink® inflammation panel. This may lead to altered structural properties of the protein target, thereby affecting antibody-antigen recognition and the ability of assays to accurately quantify protein levels. It is possible that the variants identified may not reflect variants causally associated with blood protein levels, and instead capture a causal variant in the locus. Nevertheless, the identification of such potential protein-altering variants is an important technical consideration in studies aiming to determine the molecular architecture of the human proteome. Furthermore, these variants reflect important candidates for functional characterisation in in vitro studies which aim to dissect their influence on protein abundance in cellular systems. Fourth, our Scottish cohort contains individuals from a homogenous genetic background limiting the generalisability of our findings to individuals of other ethnic backgrounds. Fifth, ageing is closely linked to chronic low-grade inflammation. Therefore, the distributions of, and correlation structure among, inflammatory protein biomarkers may differ in our cohort of healthy older ageing when compared to other age ranges and the general older adult population. Sixth, the sample size within our study resulted in large confidence and credible intervals in the reported estimates for heritabilities in inflammatory protein levels.

Conclusions

Our integrative and multi-method approach has identified high-confidence genetic and epigenetic loci associated with inflammatory protein biomarker levels. Furthermore, we have provided novel estimates for the contribution of common genetic and epigenetic variation towards differences in circulating inflammatory biomarker levels, considered alone and together. Together, our data may have important implications for informing the molecular regulation of the human proteome. Our data provides a platform upon which other researchers may investigate relationships between inflammatory biomarkers and disease, and a resource to further inform biological insights into immunological and inflammatory processes.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13073-020-00754-1.

Additional file 1. Distribution of raw values for inflammatory protein levels across individuals in Lothian Birth Cohort 1936.

Additional file 2: Supplementary Tables. The association of pre-adjusted protein levels with biological and technical covariates. Protein levels were adjusted for age, sex, array plate and four genetic principal components (population structure) prior to analyses. Significant associations are emboldened. (**Table S1**). pQTLs associated with inflammatory biomarker levels from Bayesian penalised regression model (Posterior Inclusion Probability > 95%). (**Table S2**). All pQTLs associated with inflammatory biomarker levels from ordinary least squares regression model ($P < 7.14 \times 10^{-10}$). (**Table S3**). Summary of lambda values relating to ordinary least squares GWAS and EWAS performed on inflammatory protein levels ($n = 70$) in Lothian Birth Cohort 1936 study. (**Table S4**). Conditionally significant pQTLs associated with inflammatory biomarker levels from ordinary least squares regression model ($P < 7.14 \times 10^{-10}$). (**Table S5**). Comparison of variance explained by ordinary least squares and Bayesian penalised regression models for concordantly identified SNPs. (**Table S6**). Estimate of heritability for blood protein levels as well as proportion of variance explained attributable to different prior mixtures. (**Table S7**). Comparison of heritability estimates from Ahsan et al. (maximum likelihood) and Hillary et al. (Bayesian penalised regression). (**Table S8**). List of concordant SNPs identified by linear model and Bayesian penalised regression and whether they have been previously identified as eQTLs. (Table S9). Bayesian tests of colocalisation for *cis* pQTLs and *cis* eQTLs. (**Table S10**). Sherlock algorithm: Genes whose expression are putatively associated with circulating inflammatory proteins that harbour pQTLs. (**Table S11**). CpGs associated with inflammatory protein biomarkers as identified by Bayesian model (Bayesian model; Posterior Inclusion Probability > 95%). (**Table S12**). CpGs associated with inflammatory protein biomarkers as identified by linear model (*limma*) at $P < 5.14 \times 10^{-10}$. (Table S13). CpGs associated with inflammatory protein biomarkers as identified by mixed linear model (OSCA) at $P < 5.14 \times 10^{-10}$. (**Table S14**). Estimate of variance explained for blood protein levels by DNA methylation as well as proportion of explained attributable to different prior mixtures - BayesR+. (**Table S15**). Comparison of variance in protein levels explained by genome-wide DNA methylation data by mixed linear model (OSCA) and Bayesian penalised regression model (BayesR+). (**Table S16**). Variance in circulating inflammatory protein biomarker levels explained by common genetic and methylation data (joint and conditional estimates from BayesR+). Ordered by combined variance explained by genetic and epigenetic data - smallest to largest. Significant results from t-tests comparing distributions for variance explained by methylation or genetics alone versus combined estimate are emboldened. (**Table S17**). Genetic and epigenetic factors identified by BayesR+ when conditioning on all SNPs and CpGs together. (**Table S18**). Mendelian Randomisation analyses to assess whether proteins with concordantly identified genetic signals are causally associated with Alzheimer's disease risk. (**Table S19**).

Additional file 3. Details of Supplementary Methods. Contains information for the following data: Conditional and joint analysis from ordinary least squares GWAS on protein levels; Sherlock: identifying genes whose expression associates with inflammatory biomarkers; GWAS and EWAS of CCL11 levels – incorporating smoking status as a covariate; Replication of previous pQTLs and protein associated-CpG sites; BayesR+ combined analysis – GWAS and EWAS modelled together; Evaluating causal associations between blood inflammatory proteins and Alzheimer's risk.

Additional file 4: Supplementary Figures. Correlation between the 13 proteins with significant pQTLs as identified by ordinary least squares and Bayesian penalised regression. (**Figure S1**). Correlation between heritability estimates for circulating inflammatory protein biomarkers from present study and that of Ahsan et al. The protein with the greatest discordance between studies (MMP-1) is annotated. (**Figure S2**). Correlation between the 3 proteins with significant CpG associations as identified across ordinary least squares model, mixed model and Bayesian penalised regression approaches. (**Figure S3**). Tissue-specific expression of genes annotated to CpGs associated with CCL11 levels at $P < 1 \times 10^{-5}$. Differential expression was observed in kidney, adipose and breast tissue. (**Figure S4**). Tissue-specific expression of genes annotated to CpGs associated with IL18R1 levels at $P < 1 \times 10^{-5}$. Differential expression was observed in pancreatic, hippocampal and substantia nigra tissue. (**Figure S5**). Tissue-specific expression of genes annotated to CpGs associated with CXCL9 levels at P < 1 × 10–5. No tissue-specific expression was observed. (**Figure S6**). Miami plot for IL18R1 which exhibited both genome-wide significant SNP and genome-wide significant CpG associations. The top half of the plot (skyline) shows the results from the GWAS on protein levels, whereas the bottom half (waterfront) shows the results from the EWAS. IL18R1 (chromosome 2: 102,311,529-102,398,775). (**Figure S7**).

Additional file 5. Variance in circulating protein levels explained by common genetic and methylation data together.

Acknowledgements
The authors thank LBC1936 study participants and research team members who have contributed, and continue to contribute, to ongoing LBC1936 studies.

Authors' contributions
R.F.H, M.R.R. and R.E.M were responsible for the conception and design of the study. R.F.H carried out the data analyses. R.F.H, M.R.R and R.E.M drafted the article. D.T.B, A.K, D.L.Mc.C., Q.Z, D.C.L and S.E.H contributed to the data preparation. S.E.H, N.R.W, A.F.M, P.M.V and I.J.D were responsible for the data collection. All authors read and approved the final manuscript.

Funding
The LBC1936 is supported by Age UK (Disconnected Mind program, which supports S.E.H), the Medical Research Council (MR/M01311/1), and the University of Edinburgh. Genotyping was supported by the Biotechnology and Biological Sciences Research Council (BB/F019394/1). Methylation typing was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. Proteomic analyses were supported for by the LBC1936 Age UK grant. This work was conducted in the Centre for Cognitive Ageing and Cognitive Epidemiology, which was supported by the Medical Research Council and Biotechnology and Biological Sciences Research Council (MR/K026992/1), and which supported I.J.D. We acknowledge NIH Grants R01AG054628 and R01AG05462802S1 for supporting this research and Grant P2CHD042849 for supporting the Population Research Center at the University of Texas. R.F.H. and A.J.S. 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 [R.F.H: 108890/Z/15/Z; A.J.S: 203771/Z/16/Z]. D.L.Mc.C. and R.E.M. are supported by Alzheimer's Research UK major project grant ARUK-PG2017B–10. This research was supported by Australian National Health and Medical Research Council (grants 1010374, 1046880 and 1113400) and by the Australian Research Council (DP160102400). P.M.V., N.R.W. and A.F.M. are supported by the NHMRC Fellowship Scheme (1078037, 1078901 and 1083656). P.M.V was also funded by the Australian Research Council (DP160102400 and FL180100072).

Availability of data and materials
Lothian Birth Cohort 1936 data are available on request from the Lothian Birth Cohort Study, Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh. Lothian Birth Cohort 1936 data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

Full and openly accessible summary statistics from the association studies on Olink® inflammatory protein levels are available on the University of Edinburgh Datashare site (https://datashare.is.ed.ac.uk/). These data pertain to summary statistics for GWAS (performed by two methods) and EWAS (performed by three methods) on the levels of 70 inflammatory proteins measured in members of the Lothian Birth Cohort 1936. For OLS regression GWAS data, see https://datashare.is.ed.ac.uk/handle/10283/3624; https://doi.org/10.7488/ds/2814 [101]. For BayesR+ GWAS data, see https://datashare.is.ed.ac.uk/handle/10283/3673; https://doi.org/10.7488/ds/2854 [102]. For OLS regression EWAS data, see https://datashare.is.ed.ac.uk/handle/10283/3628, https://doi.org/10.7488/ds/2818 [103]. For OSCA EWAS data, see https://datashare.is.ed.ac.uk/handle/10283/3627, https://doi.org/10.7488/ds/2817 [104]. For BayesR+ EWAS data, see https://datashare.is.ed.ac.uk/handle/10283/3626; https://doi.org/10.7488/ds/2816 [105]. Summary statistics for the OLS GWAS data are also available at GWAS Catalog (https://www.ebi.ac.uk/gwas/; Study Accessions: GCST90000437-GCST90000506) [106].

Ethics approval and consent to participate
Ethical permission for the LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics Committee (LREC/2003/2/29). Written informed consent was obtained from all participants. This study was performed in accordance with the Helsinki declaration.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. [2]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland. [3]Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK. [4]Lothian Birth Cohorts, University of Edinburgh, Edinburgh EH8 9JZ, UK. [5]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia. [6]Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh EH16 4UX, UK. [7]Department of Psychology, The University of Texas at Austin, Austin, TX 78712, USA. [8]Population Research Center, The University of Texas at Austin, Austin, TX 78712, USA. [9]Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria.

Received: 11 March 2020 Accepted: 10 June 2020
Published online: 08 July 2020

References
1. Chen L, Deng H, Cui H, Fang J, Zuo Z, Deng J, et al. Inflammatory responses and inflammation-associated diseases in organs. Oncotarget. 2017;9(6):7204–18.
2. Murakami M, Hirano T. The molecular mechanisms of chronic inflammation development. Fronti immunol. 2012;3:323.

3. Furman D, Campisi J, Verdin E, Carrera-Bastos P, Targ S, Franceschi C, et al. Chronic inflammation in the etiology of disease across the life span. Nat Med. 2019;25(12):1822–32.
4. Amor S, Puentes F, Baker D, van der Valk P. Inflammation in neurodegenerative diseases. Immunology. 2010;129(2):154–69.
5. Pahwa R, Goyal A, Bansal P, Jialal I. Chronic Inflammation. Treasure Island (Florida): StatPearls Publishing; 2020.
6. Ligthart S, Vaez A, Vosa U, Stathopoulou MG, de Vries PS, Prins BP, et al. Genome analyses of >200,000 individuals identify 58 loci for chronic inflammation and highlight pathways that link inflammation and complex disorders. Am J Hum Genet. 2018;103(5):691–706.
7. Dehghan A, Dupuis J, Barbalic M, Bis JC, Eiriksdottir G, Lu C, et al. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation. 2011;123(7):731–8.
8. Ho JE, Chen WY, Chen MH, Larson MG, McCabe EL, Cheng S, et al. Common genetic variation at the IL1RL1 locus regulates IL-33/ST2 signaling. J Clin Invest. 2013;123(10):4208–18.
9. de Vries PS, Chasman DI, Sabater-Lleal M, Chen M-H, Huffman JE, Steri M, et al. A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. Hum Mol Genet. 2016;25(2):358–70.
10. Naitza S, Porcu E, Steri M, Taub DD, Mulas A, Xiao X, et al. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS Genet. 2012; 8(1):e1002480.
11. Durda P, Sabourin J, Lange EM, Nalls MA, Mychaleckyj JC, Jenny NS, et al. Plasma levels of soluble interleukin-2 receptor alpha: associations with clinical cardiovascular events and genome-wide association scan. Arterioscler Thromb Vasc Biol. 2015;35(10):2246–53.
12. Matteini AM, Li J, Lange EM, Tanaka T, Lange LA, Tracy RP, et al. Novel gene variants predict serum levels of the cytokines IL-18 and IL-1ra in older adults. Cytokine. 2014;65(1):10–6.
13. Tekola Ayele F, Doumatey A, Huang H, Zhou J, Charles B, Erdos M, et al. Genome-wide associated loci influencing interleukin (IL)-10, IL-1Ra, and IL-6 levels in African Americans. Immunogenetics. 2012;64(5):351–9.
14. Huang J, Sabater-Lleal M, Asselbergs FW, Tregouet D, Shin SY, Ding J, et al. Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. Blood. 2012;120(24):4873–81.
15. Levin AM, Mathias RA, Huang L, Roth LA, Daley D, Myers RA, et al. A meta-analysis of genome-wide association studies for serum total IgE in diverse study populations. J Allergy Clin Immunol. 2013;131(4):1176–84.
16. Viktorin A, Frankowiack M, Padyukov L, Chang Z, Melén E, Sääf A, et al. IgA measurements in over 12 000 Swedish twins reveal sex differential heritability and regulatory locus near CD30L. Hum Mol Genet. 2014;23(15): 4177–84.
17. Yang M, Wu Y, Lu Y, Liu C, Sun J, Liao M, et al. Genome-wide scan identifies variant in TNFSF13 associated with serum IgM in a healthy Chinese male population. PloS One. 2012;7(10):e47990-e.
18. Liao M, Ye F, Zhang B, Huang L, Xiao Q, Qin M, et al. Genome-wide association study identifies common variants at TNFRSF13B associated with IgG level in a healthy Chinese male population. Genes Immun. 2012;13(6): 509–13.
19. He M, Cornelis MC, Kraft P, van Dam RM, Sun Q, Laurie CC, et al. Genome-wide association study identifies variants at the IL18-BCO2 locus associated with interleukin-18 levels. Arterioscler Thromb Vasc Biol. 2010;30(4):885–90.
20. Voruganti VS, Laston S, Haack K, Mehta NR, Smith CW, Cole SA, et al. Genome-wide association replicates the association of Duffy antigen receptor for chemokines (DARC) polymorphisms with serum monocyte chemoattractant protein-1 (MCP-1) levels in Hispanic children. Cytokine. 2012;60(3):634–8.
21. Kwan JS, Hsu YH, Cheung CL, Dupuis J, Saint-Pierre A, Eriksson J, et al. Meta-analysis of genome-wide association studies identifies two loci associated with circulating osteoprotegerin levels. Hum Mol Genet. 2014;23(24):6684–93.
22. Huang J, Huffman JE, Yamakuchi M, Trompet S, Asselbergs FW, Sabater-Lleal M, et al. Genome-wide association study for circulating tissue plasminogen activator levels and functional follow-up implicates endothelial STXBP5 and STX2. Arterioscler Thromb Vasc Biol. 2014;34(5):1093–101.
23. Choi SH, Ruggiero D, Sorice R, Song C, Nutile T, Vernon Smith A, et al. Six novel loci associated with circulating VEGF levels identified by a meta-analysis of genome-wide association studies. PLoS Genet. 2016;12(2): e1005874.
24. Yang X, Sun J, Gao Y, Tan A, Zhang H, Hu Y, et al. Genome-wide association study for serum complement C3 and C4 levels in healthy Chinese subjects. PLoS Genet. 2012;8(9):e1002916-e.
25. Smith NL, Huffman JE, Strachan DP, Huang J, Dehghan A, Trompet S, et al. Genetic predictors of fibrin D-dimer levels in healthy adults. Circulation. 2011;123(17):1864–72.
26. Yao C, Chen G, Song C, Keefe J, Mendelson M, Huan T, et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat Commun. 2018;9(1):3268.
27. Folkersen L, Fauman E, Sabater-Lleal M, Strawbridge RJ, Frånberg M, Sennblad B, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet. 2017;13(4):e1006706.
28. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A. 2012;109(4): 1204–9.
29. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.
30. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73–9.
31. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science (New York). 2018;361(6404):769–73.
32. Enroth S, Maturi V, Berggrund M, Enroth SB, Moustakas A, Johansson A, et al. Systemic and specific effects of antihypertensive and lipid-lowering medication on plasma protein biomarkers for cardiovascular diseases. Sci Rep. 2018;8(1):5531.
33. Deming Y, Xia J, Cai Y, Lord J, Del-Aguila JL, Fernandez MV, et al. Genetic studies of plasma analytes identify novel potential biomarkers for several complex traits. Sci Rep. 2016;6:18092.
34. Di Narzo AF, Telesco SE, Brodmerkel C, Argmann C, Peters LA, Li K, et al. High-throughput characterization of blood serum proteomics of IBD patients with respect to aging and genetic factors. PLoS Genet. 2017;13(1):e1006565.
35. Sun W, Kechris K, Jacobson S, Drummond MB, Hawkins GA, Yang J, et al. Common Genetic Polymorphisms Influence Blood Biomarker Measurements in COPD. PLoS Genet. 2016;12(8):e1006011-e.
36. Hoglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, et al. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. Sci Rep. 2019; 9(1):16844.
37. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biol. 2016;17(1):255.
38. Liang L, Willis-Owen SAG, Laprise C, Wong KCC, Davies GA, Hudson TJ, et al. An epigenome-wide association study of total serum immunoglobulin E concentration. Nature. 2015;520(7549):670–4.
39. Verschoor CP, McEwen LM, Kobor MS, Loeb MB, Bowdish DME. DNA methylation patterns are related to co-morbidity status and circulating C-reactive protein levels in the nursing home elderly. Exp Gerontol. 2018;105: 47–52.
40. Verschoor CP, McEwen LM, Kohli V, Wolfson C, Bowdish DM, Raina P, et al. The relation between DNA methylation patterns and serum cytokine levels in community-dwelling adults: a preliminary study. BMC Genet. 2017;18(1):57.
41. Marzi C, Holdt LM, Fiorito G, Tsai PC, Kretschmer A, Wahl S, et al. Epigenetic signatures at AQP3 and SOCS3 engage in low-grade inflammation across different tissues. PLoS One. 2016;11(11):e0166015.
42. Sun YV, Lazarus A, Smith JA, Chuang YH, Zhao W, Turner ST, et al. Gene-specific DNA methylation association with serum levels of C-reactive protein in African Americans. PLoS One. 2013;8(8):e73480.
43. Zaghlool SB, Kühnel B, Elhadad MA, Kader S, Halama A, Thareja G, et al. Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. Nat Commun. 2020;11(1):15.
44. Ahsan M, Ek WE, Rask-Andersen M, Karlsson T, Lind-Thomsen A, Enroth S, et al. The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. PLoS Genet. 2017;13(9):e1007005.
45. Flanagan JM. Epigenome-wide association studies (EWAS): past, present, and future. Methods Mol Biol (Clifton). 2015;1238:51–63.
46. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9:29.

47. van Iterson M, van Zwet EW, Heijmans BT, the BC. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. Genome Biol. 2017;18(1):19.
48. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics (Oxford). 2012;13(3):539–52.
49. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.
50. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44(9):1066–71.
51. Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat Methods. 2016;13:443.
52. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nat Methods. 2014;11:309.
53. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics (Oxford). 2014;30(10):1431–9.
54. Trejo Banos D, McCartney DL, Patxot M, Anchieri L, Battram T, Christiansen C, et al. Bayesian reassessment of the epigenetic architecture of complex traits. Nature communications. 2020;11(1):2865.
55. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34.
56. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
57. Zhang F, Chen W, Zhu Z, Zhang Q, Nabais MF, Qi T, et al. OSCA: a tool for omic-data-based complex trait analysis. Genome Biol. 2019;20(1):107.
58. Hillary RF, McCartney DL, Harris SE, Stevenson AJ, Seeboth A, Zhang Q, et al. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. Nat Commun. 2019;10(1):3160.
59. Taylor AM, Pattie A, Deary IJ. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. Int J Epidemiol. 2018;47(4):1042-r.
60. Deary IJ, Gow AJ, Taylor MD, Corley J, Brett C, Wilson V, et al. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. BMC Geriatr. 2007;7:28.
61. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics. 2012;13(1):86.
62. Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, et al. Estimation of a significance threshold for epigenome-wide association studies. Genet Epidemiol. 2018;42(1):20–33.
63. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164.
64. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1): 1826.
65. Hansen KD. IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays. R package version 060; 2016.
66. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv. 2018; https://doi.org/10.1101/447367:447367.
67. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.
68. Guo H, Fortune MD, Burren OS, Schofield E, Todd JA, Wallace C. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. Hum Mol Genet. 2015;24(12):3305–13.
69. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184.
70. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics (Oxford). 2005;21(16):3439–40.
71. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
72. Tenenbaum D. KEGGREST: client-side REST access to KEGG. R package version; 2016. p. 1.
73. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896–901.
74. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979–86.
75. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry. 2018;8(1):99.
76. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-base platform supports systematic causal inference across the human phenome. eLife. 2018;7:e34408.
77. He X, Fuller Chris K, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am J Hum Genet. 2013;92(5):667–80.
78. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131–8.
79. Morgan AR, Touchard S, Leckey C, O'Hagan C, Nevado-Holgado AJ, Barkhof F, et al. Inflammatory biomarkers in Alzheimer's disease plasma. Alzheimers Dement. 2019;15(6):776–87.
80. Cao W, Zheng H. Peripheral immune system in aging and Alzheimer's disease. Mol Neurodegener. 2018;13(1):51.
81. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45(10):1238–43.
82. Schnabel RB, Baumert J, Barbalic M, Dupuis J, Ellinor PT, Durda P, et al. Duffy antigen receptor for chemokines (Darc) polymorphism regulates circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators. Blood. 2010;115(26):5289–99.
83. Rot A. Contribution of Duffy antigen to chemokine function. Cytokine Growth Factor Rev. 2005;16(6):687–94.
84. Tsaprouni LG, Yang T-P, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382–96.
85. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9(5):436–47.
86. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS One. 2013;8(5):e63812-e.
87. Zhang Y, Breitling LP, Balavarca Y, Holleczek B, Schottker B, Brenner H. Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality. Int J Cancer. 2016;139(11):2482–92.
88. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenet. 2014;6(1):4.
89. Philibert RA, Beach SRH, Brody GH. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. Epigenetics. 2012; 7(11):1331–8.
90. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.
91. Kodal JB, Kobylecki CJ, Vedel-Krogh S, Nordestgaard BG, Bojesen SE. AHRR hypomethylation, lung function, lung function decline and respiratory symptoms. The European respiratory journal. 2018;51(3):1701512.
92. Shiels MS, Katki HA, Freedman ND, Purdue MP, Wentzensen N, Trabert B, et al. Cigarette smoking and variations in systemic immune and inflammation markers. J Natl Cancer Inst. 2014;106(11):dju294.
93. Fernandez-Egea E, Scoriels L, Theegala S, Giro M, Ozanne SE, Burling K, et al. Cannabis use is associated with increased CCL11 plasma levels in young healthy volunteers. Prog Neuro-Psychopharmacol Biol Psychiatry. 2013;46:25–8.
94. Krisiukeniene A, Babusyte A, Stravinskaite K, Lotvall J, Sakalauskas R, Sitkauskiene B. Smoking affects eotaxin levels in asthma patients. J Asthma. 2009;46(5):470–6.
95. Davis BK, Roberts RA, Huang MT, Willingham SB, Conti BJ, Brickey WJ, et al. Cutting edge: NLRC5-dependent activation of the inflammasome. Journal Immunol. 2011;186(3):1333–7.
96. Ma C, Wu W, Lin R, Ge Y, Zhang C, Sun S, et al. Critical role of CD6highCD4+ T cells in driving Th1/Th17 cell immune responses and mucosal inflammation in IBD. J Crohns Colitis. 2019;13(4):510–24.

97. Ren Y, Jiao X, Zhang L. Expression level of fibroblast growth factor 5 (FGF5) in the peripheral blood of primary hypertension and its clinical significance. Saudi J Biol Sci. 2018;25(3):469–73.
98. Sandborn WJ, Feagan BG, Fedorak RN, Scherl E, Fleisher MR, Katz S, et al. A randomized trial of Ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with moderate-to-severe Crohn's disease. Gastroenterology. 2008;135(4):1130–41.
99. Williams MA, O'Callaghan A, Corr SC. IL-33 and IL-18 in inflammatory bowel disease etiology and microbial interactions. Front Immunol. 2019;10:1091.
100. Borish LC, Steinke JW. 2. Cytokines and chemokines. J Allergy Clin Immunol. 2003;111(2 Suppl):S460–75.
101. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. Linear Regression GWAS Proteins. Edinburgh Datashare. 2020; https://doi.org/10.7488/ds/2814.
102. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. BayesR+ GWAS Proteins. Edinburgh Datashare. 2020; https://doi.org/10.7488/ds/2854.
103. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. Linear Regression EWAS Proteins. Edinburgh Datashare. 2020; https://doi.org/10.7488/ds/2818.
104. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. OSCA EWAS Proteins. Edinburgh Datashare. 2020; https://doi.org/10.7488/ds/2817.
105. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. BayesR+ EWAS Proteins. Edinburgh Datashare. 2020; https://doi.org/10.7488/ds/2816.
106. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. GWAS Summary Statistics on 70 Inflammatory Proteins - OLS Regression GWAS. 2020. https://www.ebi.ac.uk/gwas/:GCST90000437-GCST90000506.

Publisher's Note

6.3 Conclusion

For the first time, I provided estimates for the proportion of inter-individual variability in blood protein levels explained by genome-wide genotype and methylation data. The 70 inflammatory proteins that passed quality control criteria showed a wide range of estimates for the variance in plasma levels explained by genetic and methylation data.

I provided further evidence for associations between methylation in the *NLRC5* locus and blood levels of cytokines (366). My analyses suggested that cigarette smoking might influence CCL11 levels by altering DNA methylation in the *AHRR* locus. A *cis*-acting pQTL may explain the association between *IL18R1* methylation and plasma IL18R1 levels. We did not observe strong evidence for causal associations between AD risk and the 13 inflammatory proteins that exhibited significant pQTLs in this study. However, MR analyses suggested that higher blood levels of CD6 and IL18R1 associate with an increased risk for inflammatory bowel disease. There was a positive association between circulating IL12B levels and Crohn’s disease risk.

GWAS using larger sample sizes are required to uncover further pQTL associations for inflammatory protein levels. This will refine our knowledge of the molecular mechanisms that regulate the circulating levels of proteins involved in inflammation. We could apply these data in more robust multi-SNP MR analyses to evaluate the relationships between inflammatory proteins and AD or other neurological disorders. In this integrative study, I provided insights into the molecular processes that are associated with plasma levels of inflammatory proteins. In the next chapter, I use BayesR+ to study the genetic and epigenetic factors that associate with 282 proteins linked to AD in the literature.

7 Genome-wide and epigenome-wide studies on Alzheimer’s disease-associated proteins

7.1 Introduction

Many blood proteins have been associated with AD and other causes of dementia in the literature. In this chapter, I perform a structured literature review to identify blood proteins that have been associated with AD diagnosis or features of the disease. I then use BayesR+ to conduct an integrated GWAS and EWAS on plasma levels of these proteins in GS participants ($n \leq 1,064$). Plasma protein levels were measured using the SOMAscan platform in GS. Therefore, I focus on existing studies that assayed plasma proteins using SOMAscan technology. I apply two-sample MR to investigate causal relationships between plasma protein levels and AD risk.

This study has been submitted for publication and is included in full from Section 7.2 to 7.4.

7.2 Genome wide studies of plasma protein biomarkers for Alzheimer's disease implicate TBCA and TREM2 in disease risk (submitted)

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

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

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

[3]Centre for Clinical Brain Sciences, Chancellor's Building, 49 Little France Crescent, University of Edinburgh, EH16 4SB, Edinburgh, UK.

[4]Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh, EH16 4UX, Edinburgh, UK.

[5]Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, EH8 9JZ, Edinburgh, UK.

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

[7]Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, EH16 4UX, Edinburgh, UK.

[8]Johnson and Johnson Medical Ltd, Wokingham, RG40 3EW, UK.

[9]Institute of Science and Technology Austria, Am Campus 1, Klosterneuburg, Austria.

Abstract

The levels of many blood proteins are associated with Alzheimer's disease or its pathological hallmarks. Elucidating the molecular factors that control circulating levels of these proteins may help to identify proteins associated with disease risk mechanisms. Genome-wide and epigenome-wide studies (nindividuals ≤ 1,064) were performed on plasma levels of 282 Alzheimer's disease-associated proteins, identified by a structured literature review. Bayesian penalised regression estimated contributions of genetic and epigenetic variation towards inter-individual differences in plasma protein levels. Mendelian randomisation and colocalisation tested associations between proteins and disease-related phenotypes. Sixty-four independent genetic and 26 epigenetic loci were associated with 45 proteins. Novel findings included an association between plasma TREM2 levels and a polymorphism and CpG site within the MS4A4A locus. Higher plasma TBCA and TREM2 levels were significantly associated with lower Alzheimer's disease risk. Our data inform the regulation of biomarker levels and their relationships with Alzheimer's disease.

7.3 Background

Alzheimer's disease (AD) is one of the leading causes of disease burden and death globally (417, 418). Blood-based methods for assessing disease risk are potentially more cost-effective and less-invasive than neuroimaging methods or lumbar punctures for collecting cerebrospinal fluid (CSF). Further, blood-based measures that reflect brain pathology including beta-amyloid and phosphorylated tau are highly promising markers for diagnosis, and for recruitment and patient stratification in preventative trials (419-422). Approaches that use genomics and untargeted proteomics have suggested that there are signals in blood that might supplement targeted assays, and contribute to the understanding and prediction of AD (82, 84, 181). However, the relevance of many candidate protein markers identified by untargeted approaches to AD remains unclear (181, 422-424). Understanding the molecular factors that regulate the levels of AD-associated proteins may identify proteins with likely causal roles in disease risk (204, 205, 339, 425).

Unlike genetic factors which remain largely stable over the life-course, differential DNA methylation (DNAm) profiles at individual CpG sites are influenced by genetic and non-genetic factors. These include dietary and lifestyle behaviours (116). DNAm data may capture independent information beyond genetic factors in explaining inter-individual variation in circulating protein levels. Several genome-wide association studies (GWAS) have catalogued polymorphisms associated with plasma protein levels and identified causal relationships between proteins and disease states including AD (204, 206, 331, 334-336, 339, 343, 351). Further, Zaghlool *et al.* (2020) performed the only large-scale epigenome-wide association study (EWAS) to date on plasma protein levels (>1,000 proteins) (366). Few studies have combined GWAS and EWAS data to quantify the independent and combined contributions of genetic and epigenetic factors towards differential protein biomarker levels (326, 411, 416).

We performed a structured literature review of studies that report associations between plasma proteins and AD diagnosis or related traits such as amyloid burden and cortical atrophy (155, 163, 170-172, 178, 179, 426-430). We focused on studies that measured plasma protein levels using the SOMAscan affinity proteomics platform (SomaLogic Inc.) as this matches the protocol used in our study, Generation Scotland. We identified 282 proteins that were also measured in our sample. Our first aim was to quantify the degree to which genome-wide genetic and DNA methylation factors explain inter-individual differences in plasma levels of 282 AD-associated proteins. Using these data, our second aim was to investigate whether plasma proteins have likely causal relationships with AD.

For our first aim, we performed a combined GWAS/EWAS on circulating levels of 282 proteins in up to 1,064 participants of the family-based Generation Scotland study. Using Bayesian penalised regression (through BayesR+

software), we estimated the proportion of inter-individual variability in plasma protein levels that can be accounted for by variation in genetic and DNA methylation factors. BayesR+ implicitly adjusts for probe intercorrelations and data structure, including relatedness (133). Results were then integrated with publicly-available methylation and expression quantitative trait loci (mQTL/eQTL) data to probe the molecular mechanisms that might regulate protein abundances in plasma. For our second aim, Mendelian randomisation (MR) and colocalisation analyses tested for possibly causal relationships between plasma protein levels and AD phenotypes.

7.4 Methods

7.4.1 Study Cohort

Analyses were performed using blood samples from participants of the **ST**ratifying **R**esilience **a**nd **D**epression **L**ongitudinally (STRADL) cohort, which comprises 1,188 individuals from the larger, family-structured Generation Scotland: the Scottish Family Health Study (GS). GS consists of 24,084 individuals from across Scotland, some of whom were members of the Walker Cohort in Dundee (431) and the Aberdeen Children of the 1950s study (432). Recruitment for GS took place between 2006 and 2011. Members of the STRADL cohort partook in follow-up data collection 4-13 years after baseline (388, 433). Of the original GS members, 5,649 were invited to take part in the STRADL cohort. There were 1,188 positive respondents. Participants were tested across two sites (n = 582 and 606 from Aberdeen and Dundee, respectively).

7.4.2 Search strategy and inclusion criteria

We searched MEDLINE (Ovid interface, Ovid MEDLINE in-process and other non-indexed citations and Ovid MEDLINE 1946 onwards), Embase (Ovid interface, 1980 onwards), Web of Science (core collection, Thomson Reuters) and medRxiv/bioRxiv to identify relevant articles indexed as of 28 May 2021. Search terms are outlined in Additional File 1. Twenty-five articles were identified and one further article was identified through a supplemental manual literature search. After removal of duplicates, 23 articles were assessed for inclusion criteria: (i) original research article, (ii) proteins were measured in plasma, (iii) proteins were measured using SOMAscan technology and (iv) proteins were associated with Alzheimer's disease or related phenotypes. Twelve articles met inclusion criteria.

7.4.3 Protein measurements in Generation Scotland

The 5k SOMAscan v4 array was used to quantify the levels of plasma proteins in GS participants (n = 1,065). This highly multiplexed platform uses chemically modified aptamers termed SOMAmers (**S**low **O**ff-rate **M**odified **A**pta**mers**) that recognise epitopes on their cognate protein targets with high specificity and high affinity in the nanomolar-to-picomolar range. The recognition signal is measured as relative fluorescence units (RFUs) on microarrays (195).

Plasma samples were collected in 150 µl aliquots and stored at -80°C. Samples were run in 96-well plates and reagents were spread across three dilution factors (0.005%, 0.5%, and 20%) to create distinct sets for high, medium, and low abundance proteins, respectively. Raw microarray data were normalised through a number of quality control steps, which are detailed in Additional File 1 (403). After quality control and the exclusion of non-human proteins, deprecated markers and spuriomers, 4,235 SOMAmers were retained for proteomic analyses.

Normalised RFUs (from SomaLogic) were first log-transformed and regressed onto the following covariates: age, sex, study site (Aberdeen/Dundee), time between sample being collected and sample being processed for proteomics (factor, 4 levels) and 20 genetic principal components (PCs) of ancestry from multidimensional scaling (to control for population structure). Relationships between covariates and SOMAmers are shown in Additional File 2: Table S1. Residualised RFUs were transformed by rank-based inverse normalisation. We refer to these as protein levels; however, they reflect RFUs which have undergone a number of quality control, transformation and pre-correction steps.

7.4.4 Genome-wide association studies

Generation Scotland samples were genotyped using the Illumina Human OmniExpressExome-8v1.0 Bead Chip and processed using the Illumina Genome Studio software v2011 (Illumina, San Diego, CA, USA) (389). Quality control steps are outlined in Additional File 1. After quality control, 561,125 SNPs remained for 1,064 individuals. In total, 1,064 individuals had both genotype and proteomic data available for analyses.

Bayesian penalised regression GWAS were performed using BayesR+ software in C++ (133). BayesR+ utilises a mixture of prior Gaussian distributions to allow for markers with effect sizes of different magnitudes. It also includes a discrete spike at zero that enables the exclusion of markers with non-identifiable effects on the trait of interest. Guided by data from our previous studies, mixture variances for the stand-alone GWAS were set to 0.01 and 0.1 to allow for markers that account for 1% or 10% of variation in circulating protein levels, respectively (411, 416). In the combined GWAS/EWAS analysis, genotype and DNAm data must have had the same number of prior variances (n = 3 each). Therefore, mixture variances for SNP data were set to 0.01, 0.1 and 0.2 in the combined analyses. Input data were scaled to mean zero and unit variance. Gibbs sampling was used to sample over the posterior distribution conditional on input data and 10,000 samples were used. The first 5,000 samples of burn-in were removed and a thinning of 5 samples was applied to reduce autocorrelation. SNPs which exhibited a posterior inclusion probability ≥ 95% were deemed significant.

7.4.5 Epigenome-wide association studies

Blood DNAm in Generation Scotland participants was assessed using the Illumina HumanMethylationEPIC BeadChip Array. Blood DNAm was assessed in two separate sets. After quality control, 793,706 and 773,860 CpG remained in sets 1 and 2, respectively. In total, 772,619 CpG sites were shared across sets. Each set was truncated to these overlapping probes.

In the stand-alone EWAS and combined GWAS/EWAS, mixture variances were set to 0.001, 0.01 and 0.1 (n = 778). Missing DNAm data were mean imputed separately within each set as BayesR+ cannot accept missing values. Both sets were then combined and adjusted for DNAm batch, set, age and sex. Each CpG site was scaled to mean zero and unit variance. Houseman-estimated white blood cell proportions were included as fixed-effect covariates in EWAS models (434). CpG sites that had a posterior inclusion probability ≥ 95% were deemed significant.

Linear mixed-effects models were performed in sensitivity EWAS analyses using the lmekin function from the *coxme* package in R (version 2.2-16) (407). DNAm data were pre-corrected for age, sex, batch and set. Houseman-estimated white blood cell proportions were incorporated as fixed-effect covariates and a kinship matrix was fitted to account for relatedness among individuals in the family-based STRADL cohort.

7.4.6 Colocalisation analyses

Formal Bayesian tests of colocalisation were used to determine whether a shared causal variant likely underpinned two traits of interest (234). For each protein, a 200 kilobase region (upstream and downstream, recommended default setting) surrounding the variant was extracted from our GWAS summary statistics.

Expression QTL data were extracted from eQTLGen summary statistics. Methylation QTL summary statistics were extracted from phenoscanner, GoDMC or our own mQTL analyses. Methylation QTL analyses were performed using additive linear regression models and by regressing CpG sites (beta values) on SNPs (0, 1, 2) while adjusting for age, sex, DNAm batch, set, Houseman-estimated white blood cell proportions and 20 genetic PCs (n = 778). In instances where an mQTL effect was identified in more than one database, summary statistics from the study with the largest sample size were used in *coloc* (199, 275, 435). For AD-related traits, summary statistics were extracted from the relevant GWAS (84, 436, 437). Default priors were applied. Summary statistics for all SNPs (± 200 kilobases from the queried SNP) were used to estimate the posterior probability for five separate hypotheses: a single causal variant for both traits, separate causal variants for both traits, a causal variant for just one trait (encompassing two hypotheses), or no causal variant for either trait. Posterior probabilities ≥ 95% provided strong evidence for a given hypothesis.

7.4.7 Mendelian randomisation

Bidirectional Mendelian randomisation was used to test for possibly causal relationships between (i) gene expression and plasma protein levels, (ii) DNAm and plasma protein levels and (iii) plasma protein levels and AD risk or related biomarkers. Pruned variants ($r^2 < 0.1$) were used as instrumental variables (IVs) in MR analyses. In tests where only one independent variant remained after pruning, effect size estimates were assessed using the Wald ratio test. In tests where two SNPs remained, analyses were performed using the inverse variance-weighted method. Analyses were conducted using MR-base (408). Further information on IVs used are provided in Additional File 1.

7.4.8 Ethics approval and consent to participate

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

7.5 Results

7.5.1 Identification of plasma proteins associated with Alzheimer's disease

Following a structured search of MEDLINE (Ovid), Embase (Ovid), Web of Science (Thomson Reuters) and preprint servers, twelve studies were identified that reported associations between SOMAscan plasma proteins and AD or related traits (Figure 7-1). In total, 359 unique proteins were identified and 22 (6.1%) were reported in more than one study (Additional File 2: Table S2-S4). In the Generation Scotland dataset, there were 308 SOMAmers (**S**low **O**ff-rate **M**odified **A**pta**mers**) that targeted 282/359 proteins of interest (Additional File 2: Table S5 and Additional File 3: Figure S1). The 282 unique proteins were brought forward for analyses (UniProt IDs and Seq-ids are shown in Additional 2: Table S6).

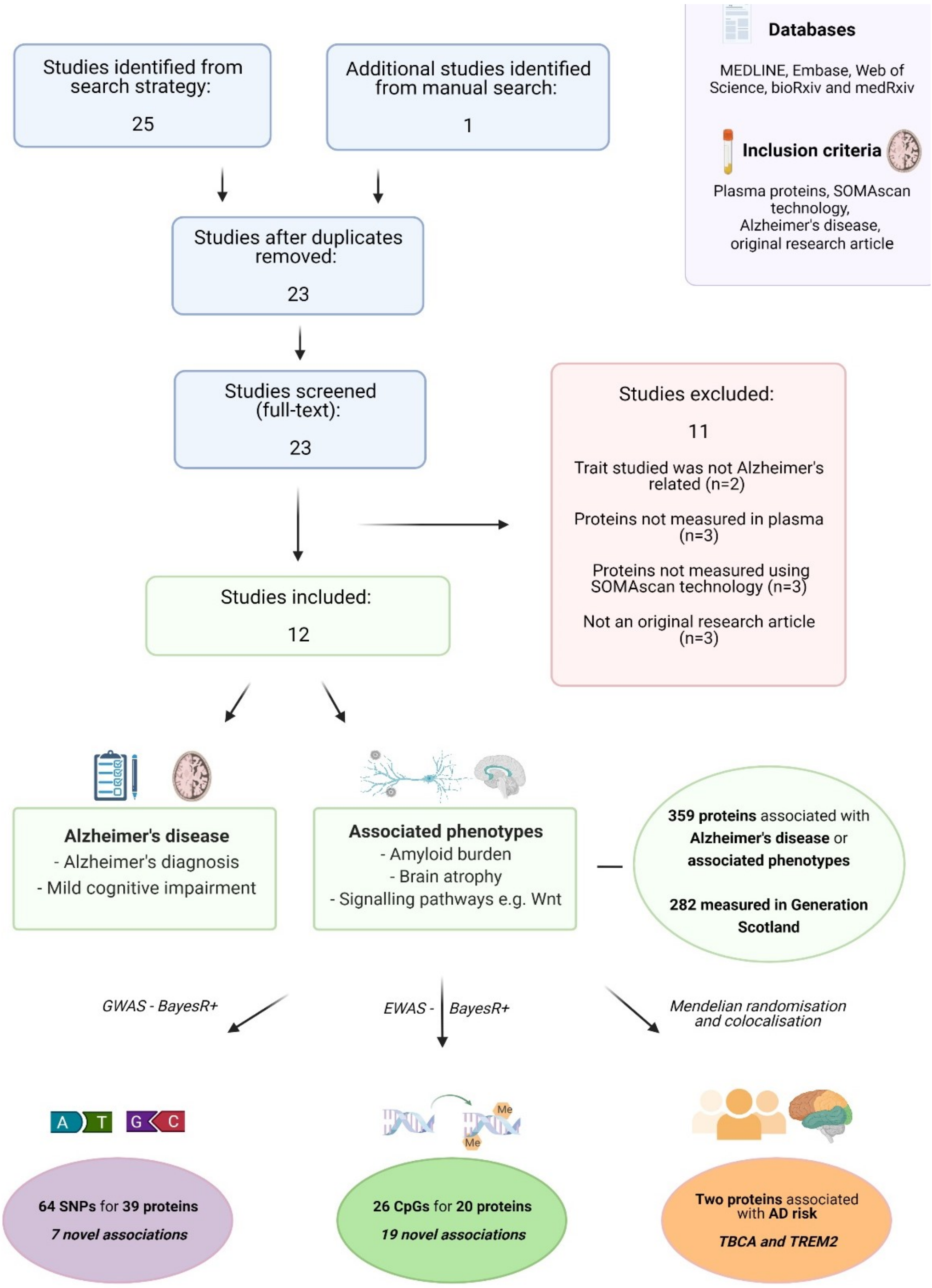
Databases
MEDLINE, Embase, Web of Science, bioRxiv and medRxiv
Inclusion criteria
Plasma proteins, SOMAscan technology, Alzheimer's disease, original research article
Studies identified from search strategy: 25
Additional studies identified from manual search: 1
Studies after duplicates removed: 23
Studies screened (full-text): 23
Studies excluded: 11
Trait studied was not Alzheimer's related (n=2)
Proteins not measured in plasma (n=3)
Proteins not measured using SOMAscan technology (n=3)
Not an original research article (n=3)
Studies included: 12
Alzheimer's disease
- Alzheimer's diagnosis
- Mild cognitive impairment
Associated phenotypes
- Amyloid burden
- Brain atrophy
- Signalling pathways e.g. Wnt
359 proteins associated with Alzheimer's disease or associated phenotypes
282 measured in Generation Scotland
GWAS - BayesR+
EWAS - BayesR+
Mendelian randomisation and colocalisation
A
T
G
C
Me
Me
64 SNPs for 39 proteins
7 novel associations
26 CpGs for 20 proteins
19 novel associations
Two proteins associated with AD risk
TBCA and TREM2

Figure 7-1 Structured literature review of SOMAscan plasma proteins that were associated with Alzheimer's disease in the literature, and assessment of their molecular architectures and relationships with Alzheimer's disease in the present study. The MEDLINE, Embase, Web of Science databases and preprint servers were queried to identify studies that reported associations between SOMAscan-measured plasma proteins and Alzheimer's disease. GWAS, EWAS and causal inference analyses were performed to identify molecular correlates of 282 AD-associated plasma protein levels and to probe their causal relationships with Alzheimer's disease and related traits. AD, Alzheimer's disease; EWAS, epigenome-wide association studies; GWAS, genome-wide association studies. Figure created using Biorender.com.

7.5.2 Genome-wide studies on plasma protein correlates of Alzheimer's disease

There were 1,064 individuals with genotype and proteomic data in Generation Scotland. The mean age of the sample was 59.9 (SD = 5.9) years and 59.1% of the sample was female. In the BayesR+ GWAS, 65 independent variants (or protein quantitative trait Loci, pQTLs) were associated with 41 SOMAmers that mapped to 39 unique protein targets (posterior inclusion probability (PIP) ≥ 95%; Additional File 2: Table S7). The phenotypic correlation structure of these 41 SOMAmers is presented in Additional File 3: Figure S2. The median correlation coefficient between SOMAmer levels was 0.18. Thirty-six pQTLs represented *cis* associations (pQTLs within 10 Mb of transcription start site (TSS) for a given gene) and 28 pQTLs were *trans*-chromosomal effects (Figure 7-2). The majority of variants were located in intronic regions using annotations from the ENSEMBL variant effect predictor (46.9%, Additional File 3: Figure S3) (438).

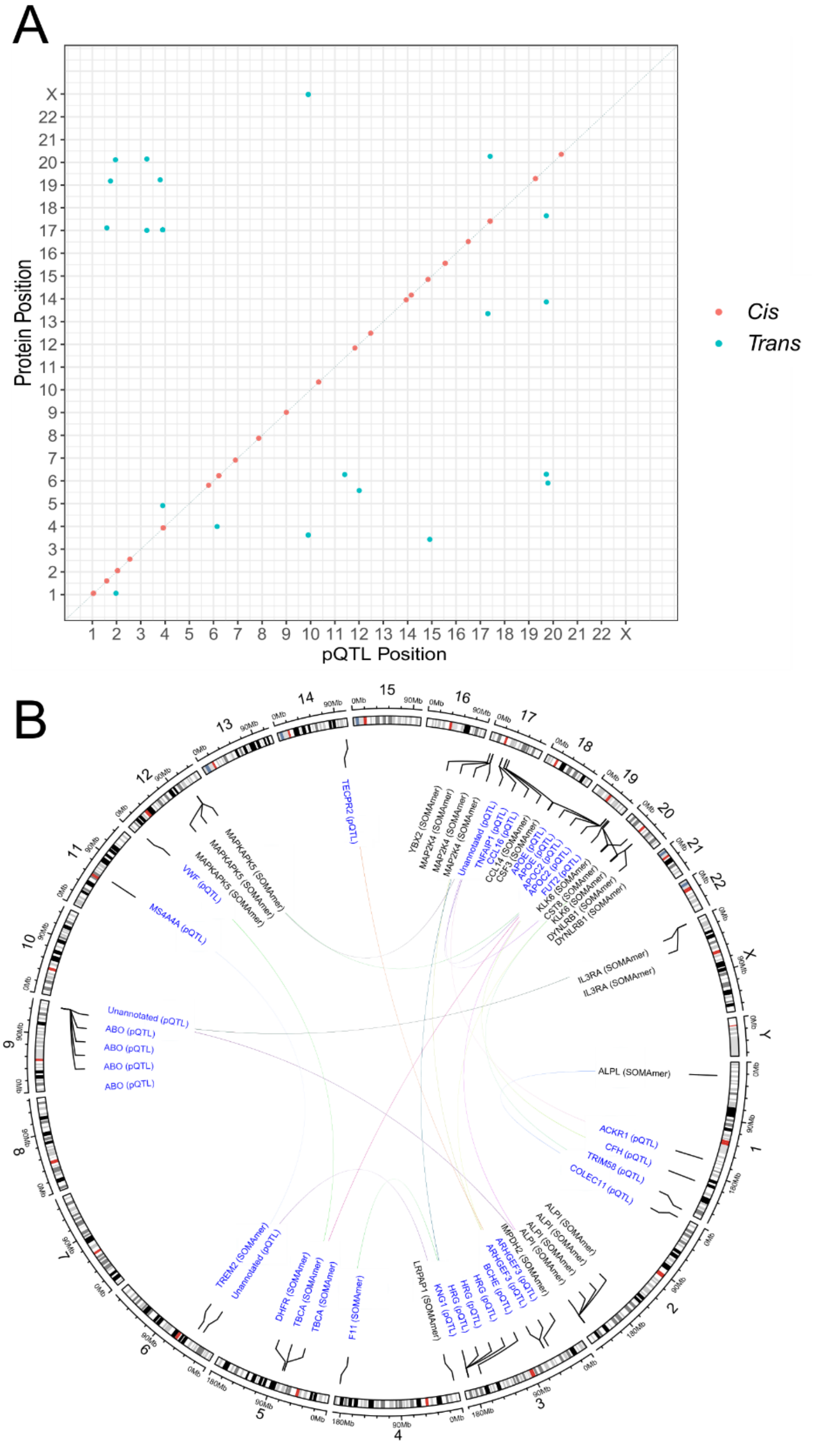

Figure 7-2. Genome-wide association studies on plasma protein levels previously associated with Alzheimer's disease and disease-related phenotypes. (A) Chromosomal locations of pQTLs identified through Bayesian penalised regression GWAS. The *x*-axis shows the chromosomal location of pQTLs associated with the levels of SOMAmers that correlate with Alzheimer's disease status or related pathways. The *y*-axis represents the position of the gene encoding the target protein. *Cis* (red circles); *trans* (blue circles). (B) A circos plot for the 28 *trans*-associated pQTLs from (A). Lines indicate an association between a pQTL and SOMAmer. GWAS, genome-wide association studies; pQTL, protein quantitative trait locus.

Fifty-seven pQTLs were previously reported in GWAS of blood protein levels (Additional File 2: Table S8) (204-206, 328, 331, 333, 339, 343, 345, 346, 351, 360, 363, 411, 416, 439, 440). Variants either directly replicated known associations or showed high linkage disequilibrium (LD, $r^2 > 0.75$) with known pQTLs for queried proteins. Relative effect sizes reported in the literature correlated strongly with those in our study ($r = 0.77$, 95% confidence interval (CI) = [0.66, 0.84]). We identified seven novel pQTLs associated with seven unique proteins. Three pQTLs were in *cis* (for GM2A, MATN3 and IL1RAP). Four pQTLs represented *trans*-chromosomal effects: rs1126680 (*BCHE* for KLK6), rs7867739 (near *ABO* for ALPI), rs3820897 (*COLEC11* for ALPL) and rs1530914 (*MS4A4A* for TREM2).

Thirty-three pQTLs were associated with at least one trait in the GWAS Catalog at $P < 5 \times 10^{-8}$ (range = 1-96 associations, Additional File 2: Table S9) (62). In relation to AD traits, the trans pQTL in *MS4A4A* (rs1530914) for TREM2 levels is in high LD with a TREM2 variant (rs1582763, $r^2 \sim 0.9$) associated with AD in APOE ε4 carriers and family history of AD (84, 441). Further, the *trans* pQTL in *APOE* (rs769449) for TBCA levels was associated with 15 AD-related traits including genetic predisposition to AD and CSF biomarkers of the disease (442, 443).

BayesR+ was used to estimate the proportions of inter-individual variation in individual plasma protein levels that were attributable to common SNPs (minor allele frequency > 1%). Estimates ranged from 5.3% (PRL; 95% credible interval (CrI) = [0%, 24.4%]) to 73.0% (IL1RAP; 95% CrI = [56.0%, 83.0%]), with a median estimate of 13.0% across all 308 SOMAmers (Additional File 2: Table S10).

7.5.3 Colocalisation of protein QTLs with expression QTLs

The 36 *cis* pQTLs identified in BayesR+ were annotated to 23 unique proteins. For 12/23 proteins, at least one pQTL was previously reported to be an expression QTL for the respective gene in blood tissue (eQTL consortium database) (199). The R package *coloc* (234) was used to test the hypothesis that one causal variant might underpin differences in gene expression (eQTL) and protein levels (pQTL) for each gene of interest. For two proteins (PCSK7 and F7), there was strong evidence (posterior probability (PP) > 95%) for a shared causal variant underlying gene expression and protein levels (Additional File 2: Table S11). MR analyses provided evidence for reciprocal associations between changes in gene expression and circulating levels of these proteins (Additional File 2: Table S12). Three proteins had weaker evidence for colocalisation (PP ≥ 75% for GM2A, LYZ and PDCD1LG2) and seven proteins had strong evidence for separate variants underlying gene expression and protein levels.

7.5.4 Epigenome-wide studies on plasma protein correlates of Alzheimer's disease

There were 778 individuals with DNA methylation and proteomic data in the Generation Scotland sample. The mean age of the sample was 60.2 (SD = 8.8) years and 56.4% of the sample were female. Twenty-six CpGs were associated with the levels of 20 unique proteins (PIP > 95%, Additional File 2: Table S13 and Additional File 3: Figure S4). The median correlation coefficient between measured protein levels was 0.16. The associations consisted of 10 *cis* CpG sites and 16 *trans* CpG loci (Figure 7-3). The cg07839457 probe in the *NLRC5* locus was associated with IL18BP and CSF1R levels, and the smoking-associated probe cg05575921 in *AHRR* (368, 369, 444) was associated with PIGR, GHR and WFDC2 levels.

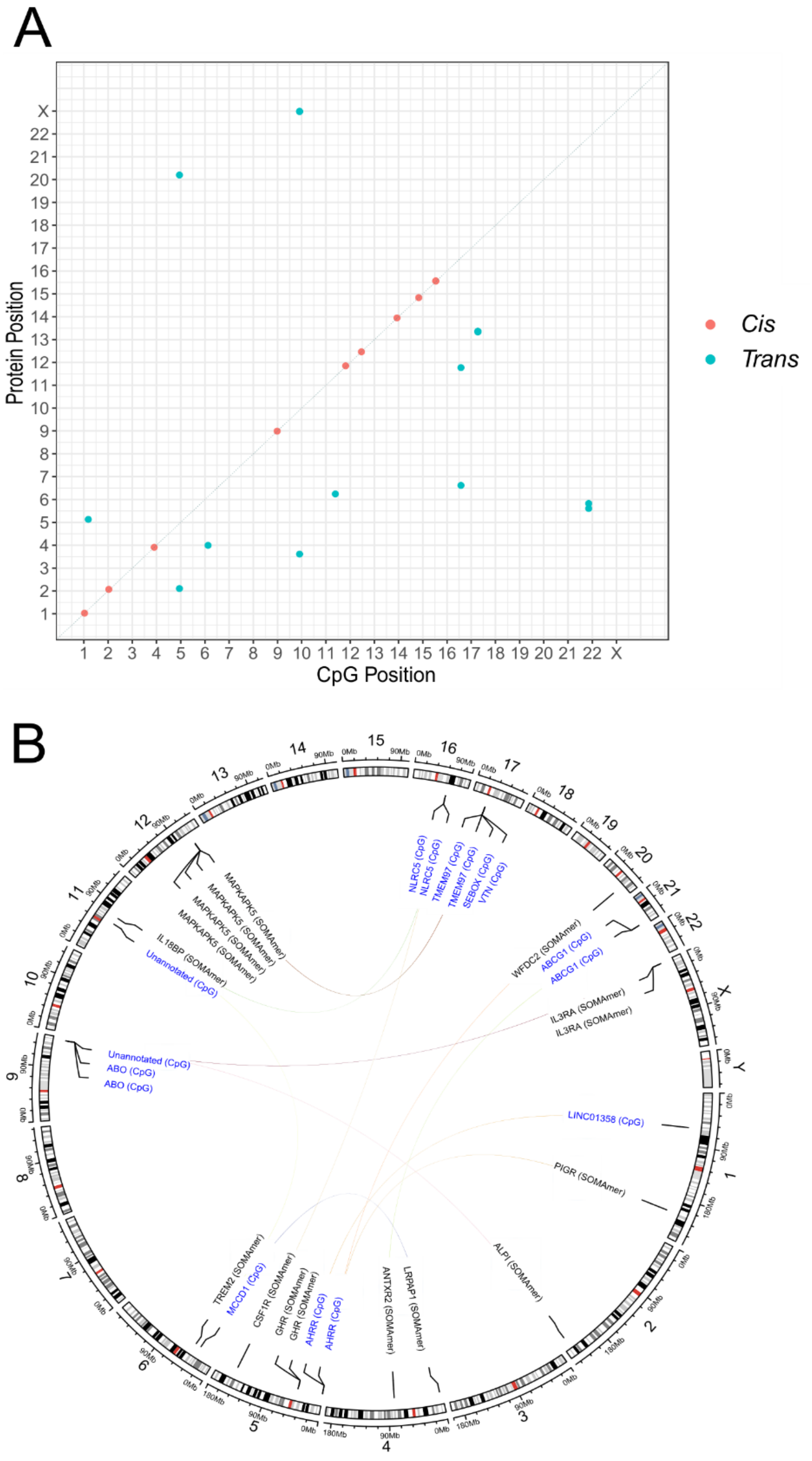

Figure 7-3. Epigenome-wide association studies on plasma protein levels previously associated with Alzheimer's disease and disease-related phenotypes. (A) Chromosomal locations of CpGs identified through Bayesian penalised regression EWAS. The *x*-axis shows the chromosomal location of CpG sites and the *y*-axis represents the position of the gene encoding the target protein. *Cis* (red circles); *trans* (blue circles). (B) A circos plot for the 16 *trans*-associated CpGs from (A). Lines indicate an association between a CpG site and SOMAmer. EWAS, epigenome-wide association studies.

We used linear mixed-effects models that accounted for relatedness to perform sensitivity analyses for the 26 CpG associations identified in BayesR+ (Additional File 2: Table S14) (407). Effect sizes were highly correlated with those from BayesR+ and showed full directional concordance (r = 0.95, 95% CI = [0.90, 0.98], Additional File 3: Figure S5). Twenty-one associations were replicated at a genome-wide significance threshold of P < 3.6 x 10^{-8} (125) and the remaining five associations were replicated with P < 2.0 x 10^{-3}. Further, 7/26 CpG associations were previously reported in the literature and effect sizes correlated strongly with those in our study (r = 0.98, 95% CI = [0.87, 1.0]). The 19 novel CpG sites were associated with levels of 14 unique proteins.

In BayesR+, estimates for the proportions of variability in SOMAmer levels that could be accounted for by DNA methylation measured on the EPIC BeadChip array ranged from 7.1% (EEA1; 95% CrI = [0%, 27.7%]) to 33.8% (MAPKAPK5; 95% CrI = [22.6%, 47.0%]), with a median estimate of 10.0% (Additional File 2: Table S15).

Estimates for variance in SOMAmer levels accounted for by genetic and methylation data together, while conditioned on each other, ranged from 21.8% for ENTPD1 (95% CrI = [0.0%, 59.1%]) to 93.3% for GHR (95% CrI = 80.1%, 100%]) (Additional File 2: Table S16 and Additional File 4). The mean and median estimates were 48.7% and 46.8%, respectively.

7.5.5 Colocalisation of protein QTLs with methylation QTLs

Fourteen proteins had both genome-wide significant pQTL and CpG associations in our study. There were 39 possible SNP-CpG pairs across these proteins. For each pair, we used linear regression to test whether the SNP was associated with CpG methylation at P < 5 x 10^{-8}, thereby representing an mQTL effect (Additional File 2: Table S17). We also performed look-up analyses of mQTL databases including GoDMC and phenoscanner (267, 275, 435). In instances where an mQTL effect was identified in more than one database, coefficients from the study with the largest sample size were brought forward for colocalisation analyses. Further, in instances where two or more mQTLs were associated with the same CpG site in a given locus, only the most significant mQTL was brought forward for colocalisation analyses (n = 19 mQTLs, 13 proteins, Additional File 2: Table S18).

For six proteins, we observed strong evidence in *coloc* that a single causal variant might underpin differential DNA methylation levels and protein abundances (PP > 95%, Additional File 2: Table S19). The six proteins were ANXA2, F7, MATN3, PLA2G2A, PCSK7 and SERPINA3. MR analyses provided evidence that relationships between methylation and protein levels were bidirectional (Additional File 2: Table S20).

7.5.6 Causal associations between plasma proteins and Alzheimer's disease risk

Bidirectional MR was applied to test for relationships between the 41 SOMAmers with pQTL associations in BayesR+ and 20 AD-related traits (Additional File 2: Table S21). A Bonferroni-corrected threshold of P < 6.10 x 10^{-5} (<0.05/41 x 20) was set. Plasma levels of three proteins had a unidirectional association with AD risk: TREM2 (Table 7-1, Wald ratio test, beta = -0.13, SE = 0.05, *P* = 8.4 x 10^{-17}), CSF3 (Wald ratio test, beta = 0.10, SE = 0.02, *P* = 5.9 x 10^{-6}) and TBCA (inverse variance-weighted method, beta = -0.50, SE = 0.12, *P* = 1.2 x 10^{-5}). Conversely, AD risk was not associated with plasma levels of these proteins. Colocalisation analyses provided evidence for one causal variant underlying TREM2 or TBCA levels and AD risk, and two separate causal variants underlying CSF3 levels and AD risk (Additional File 2: Table S22).

Table 7-1. Mendelian randomisation analyses of plasma protein levels and Alzheimer's disease-associated traits (Bonferroni-corrected $P < 6.10 \times 10^{-5}$).

Protein	Trait	Method	Beta	SE	P	Reference
Protein levels affecting Alzheimer's disease-associated traits						
TBCA	Log-transformed CSF Aβ42	IVW	-0.09	0.01	2.5×10^{-17}	(437)
TREM2	Alzheimer's disease risk	Wald ratio	-0.13	0.02	8.4×10^{-17}	(84)
TBCA	CSF APOE	Wald ratio	0.75	0.10	7.3×10^{-14}	(445)
TBCA	CSF Aβ (Z-scores)	IVW	-0.45	0.06	2.1×10^{-13}	(437)
TBCA	Log-transformed CSF Aβ42/Aβ40	IVW	-0.08	0.01	6.9×10^{-10}	(437)
CSF3	Alzheimer's disease risk	Wald ratio	0.10	0.02	5.9×10^{-6}	(84)
TBCA	Alzheimer's disease risk	IVW	-0.50	0.12	1.2×10^{-5}	(84)
Alzheimer's disease-associated traits affecting protein levels						
TBCA	Log-transformed CSF Aβ42	Wald ratio	-11.14	0.53	4.4×10^{-98}	(437)
TBCA	CSF Aβ (Z-scores)	Wald ratio	-2.13	0.10	5.7×10^{-98}	(437)
TBCA	Log-transformed CSF Aβ42/Aβ40	Wald ratio	-11.13	0.53	5.7×10^{-98}	(437)
TBCA	CSF Aβ	Wald ratio	12.21	0.63	3.7×10^{-84}	(436)

CSF, cerebrospinal fluid; IVW, inverse variance-weighted method; SE, standard error.

7.6 Discussion

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

For the first time, we show that the *trans* pQTL (rs1530914) in the *MS4A4A* locus associates with higher plasma TREM2 levels. It is in strong LD (r^2 ~ 0.9) with the variant rs1582763, which has been associated with higher CSF TREM2 levels and lower AD risk (84, 446). Furthermore, it is in moderate LD (r^2 = 0.6) with a variant in the 3'UTR region of *MS4A6A* (rs610932) that was associated with serum TREM2 levels in a sample of 35,559 Icelanders (447). Polymorphisms in *MS4A4A* was shown to alter *MS4A4A* expression and subsequently modulate TREM2 concentration in human macrophages (448). We also identified a novel blood CpG correlate of plasma TREM2 levels (cg02521229) located near *MS4A4A* that previously associated with dementia risk in Generation Scotland participants (449). Our data suggest that risk mechanisms arising from *MS4A4A* polymorphisms and TREM2 levels can be captured in plasma assays and that these mechanisms may involve differential methylation in the *MS4A4A* locus.

We observed associations between plasma levels of three proteins (CSF3, MAPKAPK5 and TBCA) and *trans* pQTLs in the *TOMM40-APOE-APOC2* locus. Further, we identified two pQTLs and three CpG correlates of plasma MAPKAPK5 levels in the *TMEM97* locus. MAPKAPK5 correlated with cognitive decline in the Twins UK cohort, however its relationship with neuropathology is unknown (426). TMEM97 acts a synaptic receptor for beta-amyloid and mediates its cellular update via APOE-dependent and APOE-independent mechanisms (450, 451). Given that *TMEM97* polymorphisms may influence MAPKAPK5 levels, our data prioritise MAPKAPK5 for follow-up studies as a potential downstream effector or correlate of TMEM97 in amyloid clearance. TBCA correlates with beta-amyloid burden (172). TBCA levels are higher in individuals with the protective *APOE* ε2/ε2 genotype and lower in carriers of the risk ε4 polymorphism (452). These data are consistent with our GWAS and MR analyses. Future studies should examine whether TBCA dysregulation is a cause or consequence of disease risk mechanisms in carriers of *APOE* ε4 polymorphisms.

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

7.7 Conclusion

In this study, the integration of multiple omics measures has clarified associations between blood proteins and disease risk mechanisms in AD. The findings in this chapter implicate TBCA and TREM2 as candidate blood-based markers of AD risk.

Across Chapters 5-7, I have performed GWAS and EWAS on plasma levels of 422 unique proteins. My analyses suggested that protein QTLs might influence blood levels of eleven proteins through altered DNA methylation. These proteins were: ANXA2, DRAXIN, F7, IL18R1, KYNU, MATN3 (Chapters 5 and 7), MDGA1, NEP, PLA2G2A, PCSK7 and SERPINA3. I used BayesR+ to quantify genetic and epigenetic variance component estimates for 342 unique proteins across Chapters 6 and 7. I have shown that combining genetic, epigenetic and proteomic data can help to elucidate relationships between individual blood proteins and disease risk.

In the next chapter, I consider an existing composite biomarker termed DNAm GrimAge, which combines epigenetic and blood proteomic data to predict biological ageing. I assess cross-sectional associations between DNAm GrimAge and several cognitive and neurology-related traits.

8 DNAm GrimAge and measures of brain health

8.1 Introduction

In Section 2.6, I introduce epigenetic measures of ageing and their potential utility in understanding the mechanisms that underlie age-related frailty and disease. The first generation of epigenetic clocks included Horvath Age and Hannum Age and were trained to predict chronological age (314, 315). There is conflicting evidence for their cross-sectional associations with cognitive function (453-456). However, accelerated biological ageing as indexed by Horvath Age and Hannum Age is associated with brain lesions (453, 455), longitudinal measures of cognitive decline (457, 458) and established risk factors for AD (309). Lu *et al.* (2019) demonstrated that a novel epigenetic predictor of mortality termed DNAm GrimAge outperformed existing epigenetic clocks in predicting the incidence of common diseases including coronary artery disease and cancer. Furthermore, an accelerated DNAm GrimAge cross-sectionally associated with type 2 diabetes, high blood pressure and excess visceral fat (313). However, the authors did not test for associations between DNAm GrimAge and measures of cognitive ability or decline.

DNAm GrimAge is of particular interest in this thesis as it incorporates methylation-based signatures of seven plasma protein levels. Therefore, in this chapter, I investigate cross-sectional associations between age-adjusted DNAm GrimAge and measures of physical and cognitive fitness, lesions on brain MRI scans and neurological protein biomarkers. I carry out these analyses using blood, cognitive and neuroimaging data from 709 LBC1936 participants (mean age = 73 years). I also examine whether an accelerated DNAm GrimAge measured at age 70 predicts all-cause mortality and cognitive decline up to age 79 in the LBC1936 cohort (n = 906).

This study was published in *Molecular Psychiatry* (459) in December 2019 and is included in full in Section 8.2.

8.2 An epigenetic predictor of death captures multi-modal measures of brain health

Molecular Psychiatry
https://doi.org/10.1038/s41380-019-0616-9

ARTICLE

An epigenetic predictor of death captures multi-modal measures of brain health

Robert F. Hillary[1] · Anna J. Stevenson[1] · Simon R. Cox[2,3] · Daniel L. McCartney[1] · Sarah E. Harris[2,3] · Anne Seeboth[1] · Jon Higham[4] · Duncan Sproul[4,5] · Adele M. Taylor[2,3] · Paul Redmond[2,3] · Janie Corley[2,3] · Alison Pattie[2,3] · Maria del. C. Valdés Hernández[2,6] · Susana Muñoz-Maniega[2,6] · Mark E. Bastin[2,6] · Joanna M. Wardlaw[2,6,7] · Steve Horvath[8,9] · Craig W. Ritchie[10] · Tara L. Spires-Jones[7,11] · Andrew M. McIntosh[2,12] · Kathryn L. Evans[1,2] · Ian J. Deary[2,3] · Riccardo E. Marioni[1,2]

Received: 30 July 2019 / Revised: 14 November 2019 / Accepted: 20 November 2019

Abstract

Individuals of the same chronological age exhibit disparate rates of biological ageing. Consequently, a number of methodologies have been proposed to determine biological age and primarily exploit variation at the level of DNA methylation (DNAm). A novel epigenetic clock, termed 'DNAm GrimAge' has outperformed its predecessors in predicting the risk of mortality as well as many age-related morbidities. However, the association between DNAm GrimAge and cognitive or neuroimaging phenotypes remains unknown. We explore these associations in the Lothian Birth Cohort 1936 ($n = 709$, mean age 73 years). Higher DNAm GrimAge was strongly associated with all-cause mortality over the eighth decade (Hazard Ratio per standard deviation increase in GrimAge: 1.81, $P < 2.0 \times 10^{-16}$). Higher DNAm GrimAge was associated with lower age 11 IQ ($\beta = -0.11$), lower age 73 general cognitive ability ($\beta = -0.18$), decreased brain volume ($\beta = -0.25$) and increased brain white matter hyperintensities ($\beta = 0.17$). There was tentative evidence for a longitudinal association between DNAm GrimAge and cognitive decline from age 70 to 79. Sixty-nine of 137 health- and brain-related phenotypes tested were significantly associated with GrimAge. Adjusting all models for childhood intelligence attenuated to non-significance a small number of associations (12/69 associations; 6 of which were cognitive traits), but not the association with general cognitive ability (33.9% attenuation). Higher DNAm GrimAge associates with lower cognitive ability and brain vascular lesions in older age, independently of early-life cognitive ability. This epigenetic predictor of mortality associates with different measures of brain health and may aid in the prediction of age-related cognitive decline.

Introduction

The rapid ageing of the global population has resulted in an increase in the personal and societal burden of age-associated disease and disability [1]. Consequently, there is an urgent need to identify those individuals at high risk of age-related morbidities and mortality. Recently, a number of methods for determining biological age have been developed which leverage inter-individual variation in physiological and molecular characteristics [2–6]. Primarily, these measures of biological age have focussed on variation at the level of DNA methylation (DNAm). DNAm is a commonly-studied epigenetic mechanism typically characterised by the addition of a methyl group to a cytosine-phosphate-guanine (CpG) nucleotide base pairing, thereby permitting regulation of gene activity [7]. Crucially, these biological age predictors, also referred to as 'epigenetic clocks', correlate strongly with chronological age; furthermore, for a given chronological age, an advanced epigenetic age is associated with increased mortality risk and many age-related morbidities [8–12].

A novel epigenetic clock, termed 'DNAm GrimAge' has been developed to predict mortality [13]. To derive DNAm GrimAge, an elastic net Cox regression model was used to

These authors contributed equally: Robert F. Hillary, Anna J. Stevenson

Supplementary information The online version of this article (https://doi.org/10.1038/s41380-019-0616-9) contains supplementary material, which is available to authorized users.

✉ Riccardo E. Marioni
riccardo.marioni@ed.ac.uk

Extended author information available on the last page of the article.

Published online: 03 December 2019

regress time-to-death due to all-cause mortality on chronological age, sex and DNAm-based surrogates for smoking pack years and 12 plasma proteins. The model selected chronological age, sex and methylation-based surrogates for smoking pack years and for 7/12 plasma proteins. The linear combination of these variables allows for an estimation of DNAm GrimAge. As with other epigenetic clocks, if an individual's DNAm GrimAge is higher than their chronological age, then this provides a measure of accelerated biological ageing. Lu et al. [13] comprehensively demonstrated that an accelerated DNAm GrimAge (also known as AgeAccelGrim) is associated with a number of peripheral, lifestyle and cardiometabolic traits and outperforms predecessor clocks in predicting death. However, the relationship between an accelerated GrimAge and cognitive as well as neuroimaging phenotypes remains unexplored. As brain structure and cognitive function show mean declines with age, and associate with disability and disease burden, the discovery of molecular correlates of neurological and neurostructural aberrations may be of particular benefit in gerontology [14, 15]. In this study we test the hypothesis that, in a large narrow age-range population cohort of older adults (Lothian Birth Cohort 1936 (LBC1936)), an accelerated DNAm GrimAge is cross-sectionally associated with poorer cognitive performance, structural neuroimaging measures and neurology-related proteins.

In addition, higher childhood intelligence (as defined by age 11 IQ) is associated with a lower risk of mortality across the life course [16–18]. Furthermore, childhood intelligence associates with a healthier lifestyle and less morbidity in middle age, as well as a lower allostatic load in older age [19–21]. Intelligence in early life is related to variability in cortical thickness, white matter macro- and micro-structure, as well as cognitive ability, fewer vascular lesions and lower risk of stroke in later life [22–27]. Notably, adjustment for age 11 IQ was recently shown to attenuate associations between another epigenetic clock measure, DNAm PhenoAge, and a wide range of phenotypes including cognitive traits in LBC1936 [28]. Therefore, we also test the hypothesis that controlling for childhood intelligence attenuates associations between DNAm GrimAge and mortality, cognitive and neuroimaging measures, as well as neurology-related proteins in older age.

Materials and methods

The Lothian Birth Cohort 1936

The LBC1936 comprises Scottish individuals born in 1936, most of whom took part in the Scottish Mental Survey 1947 at age 11. Participants who were living within Edinburgh and the Lothians were re-contacted ~60 years later. Of these participants, 1091 consented and joined the LBC1936. Upon recruitment, participants were ~70 years of age (mean age: 69.6 ± 0.8 years) and subsequently attended four additional waves of clinical examinations about every 3 years. Detailed genetic, epigenetic, physical, psychosocial, cognitive, neuroimaging, health and lifestyle data are available for members of the LBC1936. Recruitment and testing of the LBC1936 have been described previously [29, 30].

Methylation preparation in the Lothian Birth Cohort 1936

DNA from whole blood was assessed using the Illumina 450 K methylation array at the Edinburgh Clinical Research Facility. Details of quality control procedures have been described elsewhere (see Supplementary Methods) [31, 32].

Derivation of DNAm GrimAge

DNAm GrimAge was calculated using the online age calculator (https://dnamage.genetics.ucla.edu/) developed by Horvath [33]. LBC1936 methylation data were used as input for the algorithm and data underwent a further round of normalisation by the age calculator. The DNAm GrimAge biomarker was calculated using a method developed by Lu et al. [13] and is based on a linear combination of age, sex, DNAm-based surrogates for smoking, and seven proteins (adrenomedulin (DNAm ADM), beta-2-microglobulin (DNAm B2M), cystatin C (DNAm cystatin C), growth differentiation factor 15 (DNAM GDF15), leptin (DNAm leptin), plasminogen activation inhibitor 1 (DNAm PAI1), and tissue inhibitor metalloproteinaise (DNAm TIMP1)). Supplementary Fig. 1 shows the correlation between all methylation-based surrogates. All predictors, with the exception of DNAm Leptin ($r^2 = -0.29$), were positively correlated with DNAm GrimAge (absolute range = [0.24: 0.82], median = 0.25 and mean of correlation coefficients = 0.25). The difference between DNAm GrimAge and chronological age (an accelerated DNAm GrimAge) provides a measure of biological ageing. In a previous study, for a given chronological age, individuals with higher DNAm GrimAge had a higher risk for mortality than individuals of the same chronological age with a lower DNAm GrimAge [13].

Phenotypic data

Our phenotypic analyses were divided into four sections. Firstly, we examined the association between age-adjusted DNAm GrimAge and mortality in the LBC1936 over 9 years of follow-up. For our survival models (and later longitudinal cognitive analyses), we aimed to determine whether DNAm GrimAge at Wave 1 of the LBC1936 study ($n = 906$; age: 70 years) could predict mortality (or cognitive decline) over all four waves of available data (to age 79 years). For all other phenotypic analyses, we examined cross-sectional associations with age-adjusted DNAm GrimAge at Wave 2 (age: 73 years). This is because complete proteomic, brain imaging, DNAm and phenotypic data were available at this time point only ($n = 709$ individuals). For cross-sectional analyses, we did not wish to determine whether Wave 1 (age: 70 years) epigenetic data associated with Wave 2 (age: 73 years) phenotypic data in order to limit the potential issue of retrocausality. In this first section, we also investigated the cross-sectional association of an accelerated DNAm GrimAge with a number of physical (body mass index, height, grip strength, lung function and weight) and blood traits (albumin, C-reactive protein, cholesterol, creatinine, ferritin, interleukin-6 and iron; at Wave 2; age 73 years) that have been related to mortality and frailty in older age [34–42].

Secondly, we tested the association between an accelerated DNAm GrimAge and cognitive traits ($n = 18$ phenotypes). Cognitive tests taken at Wave 2 (age: 73 years) included six Wechsler Adult Intelligence Scale-III UK (WAIS-III) non-verbal subtests (matrix reasoning, letter number sequencing, block design, symbol search, digit symbol, and digit span backward). Principal component analysis (PCA) was performed using these cognitive tests and scores on the first un-rotated principal component (general cognitive ability, *g*) were extracted which explained 51% of variance. Individual test loadings ranged from 0.65 to 0.75. Wechsler Memory Scale-III items as well as measures of crystallised intelligence and reaction time were also examined in relation to DNAm GrimAge. In addition, we examined whether an accelerated DNAm GrimAge associated with *APOE* ε4 carrier status. Similar to our survival analyses, we used Wave 1 epigenetic data to determine whether DNAm GrimAge (at age 70 years) could predict decline in general cognitive ability across all four waves of the LBC1936 study. For this analysis, we used the lmerTest package in R to fit mixed-effects models to regress general cognitive ability onto sex and an interaction term between DNAm GrimAge at Wave 1 and chronological age, all as fixed effects [43]. In addition, participant ID was fitted as a random effect on the intercept.

Thirdly, we tested the association between an accelerated DNAm GrimAge and neuroimaging phenotypes at Wave 2 (age: 73 years, see Supplementary Methods). The brain MRI acquisition and processing pipeline has been made available in an open access protocol paper [44]. Total brain, normal-appearing white matter, grey matter and white matter hyperintensity volumes were segmented using a semi-automated multi-spectral technique [45]. These volumes were then expressed as a proportion of intracranial volume (ICV), which controls for the confounding effect of head size. The resultant ratios were tested for associations with age-adjusted DNAm GrimAge. Diffusion-tensor imaging-derived measures of fractional anisotropy (FA) and mean diffusivity (MD) were obtained for participants at Wave 2 (age: 73 years). Prior to conducting region-specific analyses, general factors of FA (gFA) and MD (gMD) were derived by entering the left and right FA and MD values of each tract separately into a PCA. Scores from the first un-rotated principal component were extracted and labelled as gFA (variance explained: 52%, loadings: 0.46–0.95) or gMD (variance explained: 48%, loadings: 0.47–0.88), respectively. These general factors reflect common microstructural properties across main white matter pathways and capture the common variance in white matter integrity [46].

Fourthly, we tested the association between an accelerated DNAm GrimAge and the levels of 92 neurological protein biomarkers (Olink® neurology panel). The neurology panel represents proteins with established links to neuropathology as well as exploratory proteins with roles in processes including cellular communication and immunology. Plasma was extracted from 816 blood samples collected in citrate tubes at mean age 72.5 ± 0.7 years (Wave 2; Supplementary Methods). Protein levels were transformed by rank-based inverse normalisation. Normalised protein levels were regressed onto age-adjusted DNAm GrimAge.

Descriptive statistics for phenotypes are presented in Supplementary File 1. Data collection protocols have been described fully previously and are described in Supplementary Note 1 [47].

Statistical analyses

DNAm GrimAge was regressed onto chronological age for all LBC1936 participants. These residuals were defined as an accelerated DNAm GrimAge (also known as AgeAccelGrim). Linear regression models were used to investigate relationships between continuous variables and an accelerated DNAm GrimAge, as well as age-adjusted methylation-based surrogates for smoking pack years and the plasma proteins that feed into DNAm GrimAge. Logistic regression was used to test the association between methylation-based predictors and *APOE* ε4 carrier status. An accelerated DNAm GrimAge, age-adjusted DNAm Pack Years or age-adjusted DNAm plasma protein levels were the independent variable of interest in each regression model and all variables were scaled to have a mean of zero and unit variance. Height and smoking status were included as covariates in the models for lung function (forced expiratory volume FEV1; forced vital capacity: FVC; forced expiratory ratio: FER; and peak expiratory flow: PEF). All models were adjusted for chronological age and sex. Mixed-effects models were used to examine the longitudinal association between an accelerated DNAm GrimAge and general cognitive ability. To investigate possible statistical confounding by childhood cognitive ability, all models were repeated with adjustment for age 11 IQ scores. To correct for multiple testing, and given that the methylation-based predictors exhibited a high degree of inter-correlation, we applied the false discovery rate (FDR; [48]) method to phenotypic association analyses ($n = 137$ phenotypes), separately for each predictor. Associations between age-adjusted DNAm GrimAge and regional cortical volume were conducted using the SurfStat toolbox (http://www.math.mcgill.ca/keith/surfstat) for Matrix Laboratory R2018a (The MathWorks Inc, Natick, MA), using the same covariates as above and FDR correction for multiple testing.

Results

Cohort characteristics

Details of LBC1936 participant characteristics at Waves 1 and 2 are presented in Supplementary File 1. Briefly, 47.6% of participants in this study were female. At Wave 1 (relating to the mortality and longitudinal analyses), mean chronological age for both males and females was 69.6 years (SD 0.8), whereas the mean DNAm GrimAge was 67.4 years (SD 5.2). At Wave 2 (relating to cross-sectional analyses), mean chronological age for both males and females was 72.5 years (SD 0.7), whereas the mean DNAm GrimAge was 70.0 years (SD 4.9). The lower mean measure of epigenetic age when compared to chronological age may reflect overall good health of the cohort. However, the variance associated with DNAm GrimAge is much higher than that of chronological age. When calculated across all four available waves of the LBC1936 study, DNAm GrimAge exhibits an intra-class correlation coefficient of 0.85. Mean age 11 IQ scores were 100.69 (SD: 15.37). Notably, lower IQ scores at age 11 ($\beta = -0.11$, $P = 0.02$) were associated with an accelerated DNAm GrimAge. Associations between age 11 IQ and tested phenotypes are presented in Supplementary File 2.

DNAm GrimAge predicts mortality and associates with frailty factors in the LBC1936

Mortality in LBC1936 participants was assessed in relation to an accelerated DNAm GrimAge as well as age-adjusted DNAm-based surrogate markers for plasma protein levels and smoking pack years. DNAm GrimAge was derived for 906 participants with methylation data (at Wave 1: age 70 years). There were 226 deaths (24.9%) over 9 years of follow-up.

A higher DNAm GrimAge was significantly associated with risk of all-cause mortality (Hazard Ratio (HR) = 1.81 per SD increase in DNAm GrimAge, 95% confidence interval (CI) = [1.58, 2.07], $P < 2.0 \times 10^{-16}$). Furthermore, higher levels of age-adjusted DNAm Pack Years were associated with all-cause mortality in the LBC1936 (HR = 1.64 per SD, 95% CI [1.46, 1.86], $P = 2.0 \times 10^{-16}$). In relation to methylation-based surrogates for plasma protein levels, six of the seven DNAm protein surrogates (DNAm ADM, B2M, Cystatin C, GDF15, PAI1 and TIMP1) were significantly associated with all-cause mortality (see Supplementary File 3; Fig. 1a). Following adjustment for age 11 IQ, there was very little change in the HRs and all of the predictors remained significant. Indeed, HRs from all survival models ranged from an attenuation of 2.4% to an increase of 1.8% following adjustment for childhood intelligence.

A Kaplan–Meier survival plot for an accelerated DNAm GrimAge, split into the highest and the lowest quartiles, is presented in Fig. 1b illustrating the higher mortality risk for those with a higher DNAm GrimAge. Kaplan–Meier survival plots for methylation-based surrogates for smoking pack years and plasma protein levels are presented in Supplementary Fig. 2.

For the remainder of the results, only those associations with an FDR-corrected significant P value (< 0.05) are presented herein and in Fig. 2. Full results are presented in Supplementary File 4. In relation to major mortality- and frailty-associated physical traits in the LBC1936, an accelerated DNAm GrimAge was associated with increased levels of interleukin-6 ($\beta = 0.37$, $P = 2.3 \times 10^{-18}$), C-reactive protein ($\beta = 0.25$, $P = 2.8 \times 10^{-8}$), creatinine ($\beta = 0.16$, $P = 1.1 \times 10^{-4}$), an increased body mass index ($\beta = 0.16$, $P = 2.9 \times 10^{-4}$), triglyceride concentration ($\beta = 0.13$, $P = 5.0 \times 10^{-3}$) and body weight ($\beta = 0.09$, $P = 0.04$) (Fig. 2). The relationship between accelerated DNAm GrimAge and triglycerides was no longer significant after controlling for childhood cognitive ability with the effect size decreasing from 0.13 to 0.09 (32.5% attenuation) (Supplementary File 4).

An accelerated DNAm GrimAge was negatively associated with all four measures of lung function ($\beta = [-0.16$ to $-0.27]$, $P = [9.4 \times 10^{-7}$ to $1.7 \times 10^{-16}]$), iron levels ($\beta = -0.24$, $P = 7.2 \times 10^{-7}$), low-density lipoprotein cholesterol levels ($\beta = -0.17$, $P = 1.1 \times 10^{-4}$), total cholesterol levels ($\beta = -0.13$, $P = 1.1 \times 10^{-4}$) and height ($\beta = -0.08$, $P = 0.01$) (Fig. 2). Only the relationship between accelerated DNAm GrimAge and height was non-significant after controlling for childhood intelligence, with the effect size attenuating from -0.08 to -0.06 (% attenuation: 24.5%) (Supplementary File 4). On average, associations were attenuated by 2.5% after controlling for age 11 IQ [ranged from: 19.1% increase (total

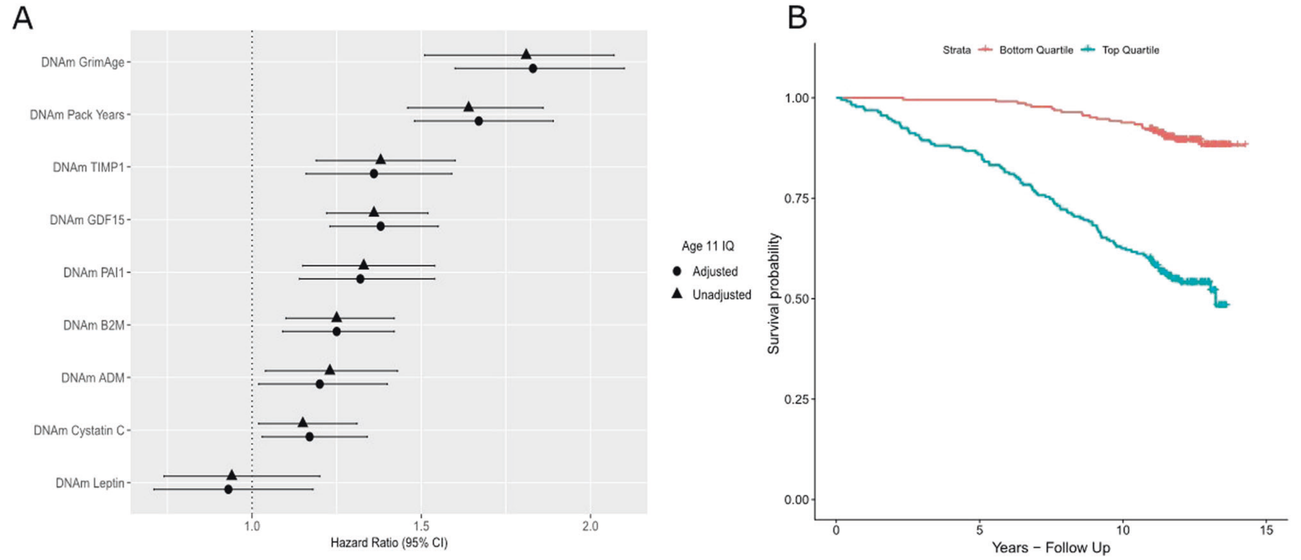

Fig. 1 DNAm GrimAge and its component surrogate markers predict mortality in the LBC1936. **a** Forest plot showing hazard ratios and 95% confidence intervals (horizontal lines) from Cox proportional hazard models for DNAm GrimAge and its constituent DNAm surrogate markers in the LBC1936 ($n = 906$, no. of deaths = 226 following nine years of follow-up). All associations with the exceptions of DNAm Leptin were significant. **b** Kaplan–Meier survival curve exhibiting the survival probabilities for the top (highest DNAm GrimAge) and bottom quartiles (lowest DNAm GrimAge) for DNAm GrimAge in the LBC1936 following 9 years of follow-up

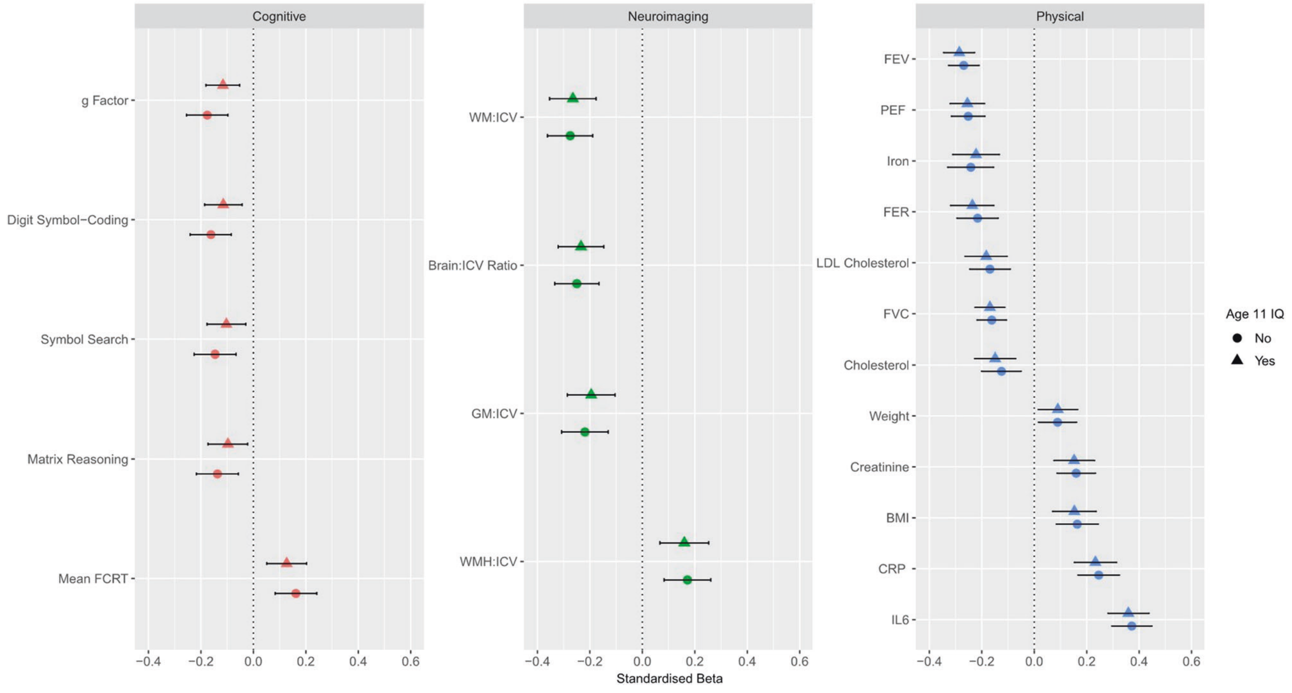

Fig. 2 Cross-sectional association between age-adjusted DNAm GrimAge and cognitive, neuroimaging and physical traits in the LBC1936. *Cognitive*: An accelerated DNAm GrimAge was negatively associated with the general factor of cognitive ability, digit symbol coding, symbol search and matrix reasoning tasks. DNAm GrimAge was also associated with an increased mean four choice reaction time. *Neuroimaging*: Age-adjusted DNAm GrimAge was negatively associated with the ratios of white matter volume, brain volume and grey matter volume to intracranial volume, and positively associated with the ratio of volume of white matter hyperintensities to intracranial volume. *Physical:* An accelerated DNAm GrimAge was negatively associated with four measures of lung function: forced expiratory volume in 1 s, forced vital capacity, forced expiratory ratio and peak expiratory flow, as well as levels of iron, low-density lipoprotein cholesterol and total cholesterol. Age-adjusted DNAm GrimAge was positively associated with weight, levels of creatinine, body mass index as well as levels of C-reactive protein and interleukin-6. Horizontal lines indiciate 95% confidence intervals. BMI body mass index, CRP C-reactive protein, FCRT four choice reaction time, FER forced expiratory ratio, FEV forced expiratory volume, FVC forced vital capacity, GM grey matter, ICV intracranial volume, IL6 interleukin-6, LDL low-density lipoprotein, PEF peak expiratory flow, WM white matter, WHM white matter hyperintensities

cholesterol) to 32.5% attenuation (triglycerides)]. All associations between blood and physical traits and an accelerated DNAm GrimAge in this study are presented in Supplementary Fig. 3. Relationships between all phenotypes tested in this study and age-adjusted DNAm Pack Years as well as age-adjusted plasma protein levels are presented in Supplementary File 5. Significant relationships are further detailed in Supplementary Note 2.

DNAm GrimAge associates with lower cognitive ability in the LBC1936

An accelerated DNAm GrimAge was significantly associated with lower measures of general cognitive ability (g: $\beta = -0.18$, $P = 8.0 \times 10^{-6}$; $n = 709$). Furthermore, an accelerated DNAm GrimAge was negatively associated with all six component tests for fluid intelligence from which *g* was derived (see Section "Phenotypic data"; $\beta = [-0.11$ to $-0.16]$, $P = [0.02$ to $2.4 \times 10^{-4}]$). In addition, an accelerated DNAm GrimAge was associated with an increased four choice reaction time mean ($\beta = 0.16$, $P = 2.9 \times 10^{-4}$). Lower IQ scores at age 70 (which correlated 0.70 with age 11 IQ scores) were associated with age-adjusted DNAm GrimAge ($\beta = -0.11$, $P = 0.02$). An accelerated DNAm GrimAge was also negatively associated with the following measures of crystallised intelligence: the Wechsler Test of Adult Reading ($\beta = -0.13$, $P = 4.0 \times 10^{-3}$) and the National Adult Reading Test ($\beta = -0.10$, $P = 0.03$).

Following adjustment for age 11 IQ, an accelerated DNAm GrimAge remained significantly associated with general cognitive ability (*g*: $\beta = -0.12$, $P = 2.0 \times 10^{-3}$; 33.9% attenuation). Three out of the six tests which constitute the general intelligence factor remained significant after adjustment for age 11 IQ (digit-symbol coding, symbol search, and matrix reasoning). Furthermore, the association between an accelerated DNAm GrimAge and an increased mean four choice reaction time remained significant following adjustment for age 11 IQ (Fig. 2). On average, associations between cognitive tasks and an accelerated DNAm GrimAge were attenuated by 41.1% following controlling for age 11 IQ (ranging from 21.7% attenuation [four choice reaction time] to 77.4% attenuation [National Adult Reading Test]). All associations between cognitive traits and an accelerated DNAm GrimAge in this study are presented in Supplementary Fig. 4. Finally, an accelerated DNAm GrimAge was not associated with *APOE* ε4 carrier status—the strongest genetic risk factor for Alzheimer's disease (odds ratio = 0.96, 95% CI = [0.93, 1.00], $P = 0.06$).

Accelerated DNAm GrimAge showed a borderline significant association with faster cognitive decline (interaction term between an accelerated DNAm GrimAge at Wave 1 and age: $\beta = -0.018$, $P = 0.05$; $n = 906$). This association was attenuated following adjustment for age 11 IQ ($\beta = -0.015$, $P = 0.11$, % attenuation: 16.7%). Secondly, restricting the set of individuals to just those incorporated into our cross-sectional design ($n = 709$), accelerated DNAm GrimAge at Wave 1 was significantly associated with decline in general cognitive ability across the eighth decade ($\beta = -0.020$, $P = 0.03$; $n = 709$). After adjusting for childhood cognitive ability, this association was attenuated to non-significance ($\beta = -0.017$, $P = 0.07$, % attenuation: 15%).

DNAm GrimAge is associated with gross neurostructural differences in the LBC1936

An accelerated DNAm GrimAge was associated with lower white matter volume ($\beta = -0.28$, $P = 1.7 \times 10^{-8}$), total brain volume ($\beta = -0.25$, $P = 1.4 \times 10^{-7}$) and grey matter volume ($\beta = -0.22$, $P = 1.3 \times 10^{-5}$). Furthermore, an accelerated DNAm GrimAge was associated with an increased volume of white matter hyperintensities ($\beta = 0.17$, $P = 1.0 \times 10^{-3}$) (Fig. 2). All associations remained significant following adjustment for age 11 IQ (Supplementary File 4). On average, these associations were attenuated by 6.98% after adjusting for age 11 IQ. All associations between neuroimaging traits and an accelerated DNAm GrimAge in this study are presented in Supplementary Fig. 5.

An accelerated DNAm GrimAge was not significantly associated with general factors of white matter microstructural metrics i.e. fractional anisotropy ($\beta = -0.009$, $P = 0.89$) or mean diffusivity ($\beta = -0.001$, $P = 0.98$), hence additional regional analyses were not performed. However, given that DNAm GrimAge was associated with grey matter volume, we further tested whether there was regional cortical heterogeneity in relation to the DNAm GrimAge-grey matter association. The negative association between accelerated DNAm GrimAge and cortical volume showed a degree of regional heterogeneity across the cortical surface (Fig. 3). The strongest magnitudes were evident in lateral and medial frontal and temporal regions, extending into motor and somatosensory cortex as well as into the posterior cingulate and precuneal areas. In contrast, associations in occipital and inferior lateral and medial frontal regions were non-significant. When the associations were additionally corrected for age 11 IQ, the magnitude of the effect sizes at the FDR-significant loci were weakly attenuated (mean *t*-value attenuation = 3.36%; Supplementary Fig. 6).

Association of DNAm GrimAge with neurological protein biomarkers

Forty of the 92 neurology-related Olink® proteins were significantly associated with an accelerated DNAm GrimAge at FDR-corrected $P < 0.05$ ($n = 709$). These

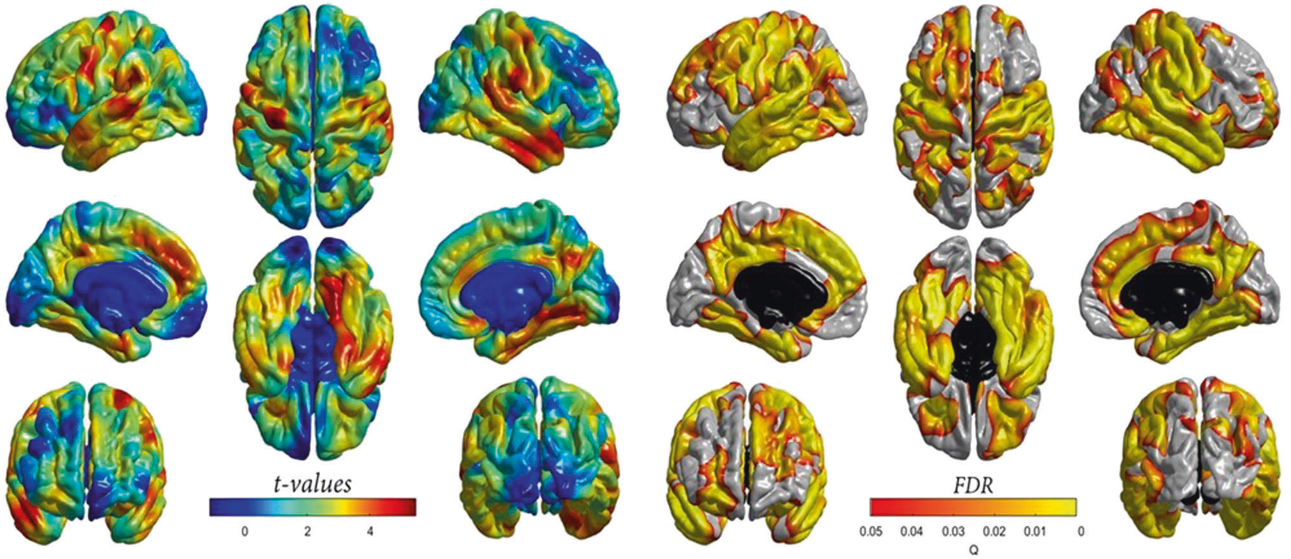

Fig. 3 Cross-sectional association between age-adjusted DNAm GrimAge and regional cortical volume in the LBC1936. Left panel: *t* values indicate the magnitude of the negative association (values have been flipped for visualisation purposes). An accelerated DNAm GrimAge was negatively associated with cortical volume. Right panel: Corresponding FDR-corrected *P* values indicate the spatial distribution of significant associations. FDR false discovery rate

proteins explained between 0.73% ($\beta = -0.09$, NC-Dase) to 7.19% ($\beta = 0.30$, SKR3) of inter-individual variation in an accelerated DNAm GrimAge (in a model which was not adjusted for age and sex; Supplementary File 6). Following adjustment for age 11 IQ, 36/40 associations (90%) remained significant. After adjusting for age 11 IQ, associations were, on average, attenuated by 3.03%.

Correlation between DNAm GrimAge and DNAm Pack Years

We observed that DNAm GrimAge and DNAm Pack Years were highly correlated (correlation coefficient: 0.82) and were cross-sectionally associated with many of the same variables in our phenotypic analyses (Supplementary File 7). Therefore, we carried out a follow-up analysis to determine the difference in magnitude between the effect sizes for DNAm GrimAge or DNAm Pack Years in relation to phenotypes associated with both predictors. Prior to adjusting for age 11 IQ, the effect sizes had a correlation coefficient of 0.88. However, they were, on average, 16.5% greater for DNAm GrimAge when compared to DNAm Pack Years. Following adjustment for age 11 IQ, the correlation coefficient was 0.84, and the effect sizes were, on average, 23.1% greater for DNAm GrimAge upon comparison to DNAm Pack Years. A plot demonstrating the correlation between effect sizes for DNAm GrimAge and DNAm Pack Years from our cross-sectional phenotypic analyses is presented in Supplementary Fig. 7.

Sex-specific differences in associations with DNAm GrimAge

As a sensitivity analysis, we accounted for an interaction between age-adjusted DNAm GrimAge and sex. Prior to adjusting for age 11 IQ, there was evidence for a sex-specific difference only in the relationships between an accelerated DNAm GrimAge and all four measures of lung function (FVC: $\beta_{\mathrm{GrimAge} \times \mathrm{males}} = 0.50$, $P_{\mathrm{GrimAge} \times \mathrm{males}} = 5.6 \times 10^{-39}$; FEV: $\beta_{\mathrm{GrimAge} \times \mathrm{males}} = 0.39$, $\mathrm{P}_{\mathrm{GrimAge} \times \mathrm{males}} = 1.3 \times 10^{-23}$; PEF: $\beta_{\mathrm{GrimAge} \times \mathrm{males}} = 0.17$, $\mathrm{P}_{\mathrm{GrimAge} \times \mathrm{males}} = 1.2 \times 10^{-4}$; FER: $\beta_{\mathrm{GrimAge} \times \mathrm{males}} = -0.14$, $\mathrm{P}_{\mathrm{GrimAge} \times \mathrm{males}} = 0.049$) (Supplementary File 8). The same interaction model was also rerun accounting for age 11 IQ. Three of the lung function tests (all but FER), as well as reading ability and general cognitive ability, exhibited significant interactions between sex and DNAm GrimAge (Supplementary File 9).

Adjustment for educational attainment

In a further sensitivity analysis, we found that an accelerated DNAm GrimAge was significantly associated with years of education ($\beta = -0.12$, $P = 1.7 \times 10^{-3}$). Models adjusted for age 11 IQ were rerun with an additional adjustment for years of education. Of the 57 relationships which remained significant after adjusting for age 11 IQ, eight were attenuated to non-significance when adjusting for education. These included associations with symbol search and weight ($\beta = -0.10$ to -0.08, % attenuation = 22.3%; $\beta = 0.09$ to 0.05, % attenuation = 44.4%, respectively) and with six proteins

(THY 1, RGMA, CDH3, TNFRSF21, NEP and TMPRSS5, mean attenuation: 8.8%) (Supplementary File 10).

Discussion

In this study, we found that a higher-than-expected DNAm GrimAge strongly predicted mortality and was associated with a number of mortality- and frailty-associated traits. This provides the first external replication of the association between DNAm GrimAge and survival. After controlling for childhood cognitive ability, we found that an accelerated DNAm GrimAge was cross-sectionally associated with lower general cognitive ability as well as slower reaction time speed and lower scores on processing speed and perceptual organisation tasks. There was tentative evidence to suggest that an accelerated DNAm GrimAge measured at age 70 may predict decline in general cognitive ability up to age 79. Furthermore, an accelerated DNAm GrimAge was associated with gross neuroanatomical differences and vascular lesions in older age. Finally, a number of neurology-related proteins were associated with an accelerated DNAm GrimAge.

DNAm GrimAge was developed using mortality as a reference and consequently supplants its predecessors in relation to mortality risk prediction. Indeed, in this study, we observed a hazard ratio of 1.81 per standard deviation increase in an accelerated DNAm GrimAge, which outperforms that of previous epigenetic clocks (Hannum Age HR: 1.22, Horvath Age HR: 1.19; DNAm PhenoAge HR: 1.17; all applied to LBC1936) [8, 28]. In relation to mortality- and frailty-associated traits, the strongest association was between DNAm GrimAge and interleukin-6. Furthermore, DNAm GrimAge was strongly associated with C-reactive protein (whose production is stimulated by interleukin-6). Together, this corroborates evidence for the “inflammaging” theory which postulates that chronic, low-grade inflammation significantly influences biological ageing and decline [49]. An accelerated DNAm GrimAge was also associated with lower low-density lipoprotein cholesterol and total cholesterol. In older age, lower levels of these blood-based factors are also associated with higher risk of mortality [50]. In addition, DNAm GrimAge was associated with a higher body mass index which does not agree with previous findings showing that an increased body mass index is protective against mortality risk [39]. However, this may be driven by a strong association between DNAm Leptin and body mass index. Indeed, leptin is an adipose tissue-derived hormone which acts an appetite suppressant, and is strongly correlated with body mass index and obesity [51, 52].

We observed a significant relationship between higher childhood intelligence (as well as age 70 IQ) and a lower DNAm GrimAge in older age. After controlling for childhood cognitive ability, associations between DNAm GrimAge and tests of crystallised intelligence were attenuated to non-significance. This finding is not surprising given that crystallised intelligence remains stable throughout adulthood [53], and that the National Adult Reading Test strongly retrodicts childhood IQ in this sample [54]. However, relationships between DNAm GrimAge and general cognitive ability, as well as fluid intelligence measures, remained significant after adjusting for age 11 IQ. Nevertheless, these associations were attenuated by an average of 41.4% following adjustment for age 11 IQ. Therefore, blood-based methylation changes, as captured by DNAm GrimAge, helps to explain additional variance in late life cognitive ability and fluid intelligence.

An accelerated DNAm GrimAge was significantly associated with gross neurostructural differences, including reductions in total brain, grey matter and white matter volumes and increases in white matter hyperintensity volumes. There was also some heterogeneity in the associations with regional cortical volume, whereby effects were strongest in frontal and temporal regions. These regions also exhibit the largest annual decrease in middle and older age [55], and are most informative for predicting chronological age (albeit using cortical thickness rather than volume; [56]). White matter hyperintensities, which associate with DNAm GrimAge, have also been linked to cortical loss in temporal and lateral frontal regions [57]. This may indicate that altered methylation profiles could help explain mechanistic relationships between neurovascular lesions and cortical atrophy. However, adjustment for vascular risk factors such as hypercholesterolaemia, smoking and diabetes is merited in this context. Furthermore, white matter hyperintensities are also related to physical disability, processing speed and cognitive decline [58, 59]. Additionally, the presence of white matter hyperintensities doubles the risk of dementia, and triples the risk of stroke, and is associated with clinical outcomes in stroke [60, 61]. Therefore, DNAm GrimAge may capture vital aspects of age-related alterations in neurostructural integrity and gross brain pathology.

Here, DNAm GrimAge associated with poorer cognitive ability and neurostructural correlates of dementia. Dementia encompasses strong psychiatric components and overlaps with other psychiatric conditions [62]. In addition, there is a significant genetic or phenotypic overlap between cognitive ability and psychiatric conditions, such as schizophrenia and depression [63, 64]. Furthermore, DNAm GrimAge captured various deleterious aspects of brain health, including altered brain structure and neurological protein biomarkers, which relate to psychiatric disorders. Thus, this composite molecular predictor of mortality should be measured in other large-scale cohorts with incident and prevalent neurological and neuropsychiatric phenotype data to determine its utility in predicting clinically-defined disease.

We observed a very strong correlation between DNAm GrimAge and DNAm Pack Years. Indeed, the associations between smoking and mortality, cognitive decline and brain pathology are well-documented [65–67]. However, the larger effect sizes for DNAm GrimAge suggest that this composite biomarker is supplemented by the inclusion of methylation-based surrogates for plasma protein levels. We identified associations with a number of neurology-related proteins ($n = 40$ before adjustment for age 11 IQ; $n = 36$ after adjustment for age 11 IQ) which may further inform the risk of mortality and age-related morbidities, particularly in relation to neurological disease. Future studies are necessary to define the biological relationships between such proteins and their relevance to age-related pathologies and cognitive decline.

The use of methylation-based proxies for smoking pack years and proteomic data is advantageous as methylation-based predictors are often more accurate than self-reported phenotypes, and the cost of complex proteomic platforms is negated [68]. One strength of this study is that rich data were available across the eighth decade of life, a period in which risk of cognitive decline and compromised brain integrity increases significantly. However, LBC1936 comprises relatively healthy older adults, complicating the generalisability of findings to at-risk clinical populations and broader age ranges.

In conclusion, we demonstrated that an epigenetic predictor of mortality associates with cognitive ability, cognitive decline and neuroimaging phenotypes in a cohort of healthy older ageing adults. These associations were largely independent of another well-known predictor of mortality, childhood intelligence. Indeed, methylation alterations in blood, as captured by DNAm GrimAge, could help provide early indications towards mortality prediction and decline in brain health.

Code availability

Code will be available from the authors on request.

Acknowledgements We thank the LBC1936 participants and study team. We thank the nursing staff at the Wellcome Trust Clinical Research Facility, and radiographers at the Brain Research Imaging Research Centre (www.bric.ed.ac.uk), Western General Edinburgh. The LBC1936 is funded by Age UK as The Disconnected Mind project, and by the Medical Research Council (G0701120, G1001245, MR/M013111/1, MR/R024065/1). SRC, MEB and IJD are also supported by a National Institutes of Health (NIH) research grant R01AG054628. RFH and AJS 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]. AS is supported by a Medical Research Council PhD Studentship in Precision Medicine with funding by the Medical Research Council Doctoral Training Programme and the University of Edinburgh College of Medicine and Veterinary Medicine. DLM and REM are supported by Alzheimer's Research UK major project grant ARUK-PG2017B-10. J.M.W is supported by the Fondation Leducq. SH is supported by 1U01AG060908 – 01.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Ethical permission for LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56), the Lothian Research Ethics Committee (Wave 1: LREC/2003/2/29) and the Scotland A Research Ethics Committee (Waves 2, 3 and 4: 07/MRE00/58).

Informed consent Written informed consent was obtained from all participants.

References

1. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380:2163–96.
2. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging. 2018;10:573–591.
3. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115–R115.
4. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–367.
5. Cole JH, Ritchie SJ, Bastin ME, Valdés Hernández MC, Muñoz Maniega S, Royle N, et al. Brain age predicts mortality. Mol Psychiatry. 2017;23:1385.
6. Vanhooren V, Dewaele S, Libert C, Engelborghs S, De Deyn PP, Toussaint O, et al. Serum N-glycan profile shift during human ageing. Exp Gerontol. 2010;45:738–43.
7. Beck S, Rakyan VK. The methylome: approaches for global DNA methylation profiling. Trends Genet. 2008;24:231–7.
8. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015;16:25–25.

9. McCartney DL, Stevenson AJ, Walker RM, Gibson J, Morris SW, Campbell A, et al. Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. Alzheimers Dement. 2018;10:429–437.
10. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. Clin Epigenetics. 2016;8:64.
11. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging. 2015;7:1130–42.
12. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging. 2016;8:1844–1865.
13. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging. 2019;11:303–327.
14. Tucker-Drob EM. Neurocognitive functions and everyday functions change together in old age. Neuropsychology. 2011;25:368–77.
15. Raz N, Rodrigue KM. Differential aging of the brain: patterns, cognitive correlates and modifiers. Neurosci Biobehav Rev. 2006;30:730–48.
16. Calvin CM, Deary IJ, Fenton C, Roberts BA, Der G, Leckenby N, et al. Intelligence in youth and all-cause-mortality: systematic review with meta-analysis. Int J Epidemiol. 2011;40:626–44.
17. Calvin CM, Batty GD, Der G, Brett CE, Taylor A, Pattie A, et al. Childhood intelligence in relation to major causes of death in 68 year follow-up: prospective population study. Brit Med J. 2017; 357:j2708.
18. Čukić I, Brett CE, Calvin CM, Batty GD, Deary IJ. Childhood IQ and survival to 79: follow-up of 94% of the Scottish Mental Survey 1947. Intelligence. 2017;63:45–50.
19. Wraw C, Deary IJ, Gale CR, Der G. Intelligence in youth and health at age 50. Intelligence. 2015;53:23–32.
20. Gale CR, Booth T, Starr JM, Deary IJ. Intelligence and socio-economic position in childhood in relation to frailty and cumulative allostatic load in later life: the Lothian Birth Cohort 1936. J Epidemiol Community Health. 2016;70:576–82.
21. Wraw C, Der G, Gale CR, Deary IJ. Intelligence in youth and health behaviours in middle age. Intelligence. 2018;69:71–86.
22. Karama S, Bastin ME, Murray C, Royle NA, Penke L, Maniega SMunoz, et al. Childhood cognitive ability accounts for associations between cognitive ability and brain cortical thickness in old age. Mol Psychiatry. 2014;19:555–9.
23. Deary IJ, Bastin ME, Pattie A, Clayden JD, Whalley LJ, Starr JM, et al. White matter integrity and cognition in childhood and old age. Neurology. 2006;66:505–12.
24. Valdés Hernández MDC, Booth T, Murray C, Gow AJ, Penke L, Morris Z, et al. Brain white matter damage in aging and cognitive ability in youth and older age. Neurobiol aging. 2013;34:2740–2747.
25. Deary IJ, Leaper SA, Murray AD, Staff RT, Whalley LJ. Cerebral white matter abnormalities and lifetime cognitive change: a 67-year follow-up of the Scottish Mental Survey of 1932. Psychol Aging. 2003;18:140–8.
26. McHutchison CA, Backhouse EV, Cvoro V, Shenkin SD, Wardlaw JM. Education, socioeconomic status, and intelligence in childhood and stroke risk in later life: a meta-analysis. Epidemiology. 2017;28:608–618.
27. Backhouse EV, McHutchison CA, Cvoro V, Shenkin SD, Wardlaw JM. Early life risk factors for cerebrovascular disease: a systematic review and meta-analysis. Neurology. 2017;88:976–984.
28. Stevenson AJ, McCartney DL, Hillary RF, Redmond P, Taylor AM, Zhang Q, et al., Childhood intelligence attenuates the association between biological ageing and health outcomes in later life. https://www.biorxiv.org/content/10.1101/588293v1. 2019.
29. Deary IJ, Gow AJ, Taylor MD, Corley J, Brett C, Wilson V, et al. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. BMC Geriatr. 2007;7:28–28.
30. Taylor AM, Pattie A, Deary IJ. Cohort profile update: the Lothian Birth Cohorts of 1921 and 1936. Int J Epidemiol. 2018;47:1042–1042r.
31. Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. Genome Res. 2014;24:1725–33.
32. Zhang Q, Marioni RE, Robinson MR, Higham J, Sproul D, Wray NR, et al. Genotype effects contribute to variation in longitudinal methylome patterns in older people. Genome Med. 2018;10:75.
33. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.
34. Velissaris D, Pantzaris N, Koniari I, Koutsogiannis N, Karamouzos V, Kotroni I, et al. C-Reactive protein and frailty in the elderly: a literature review. J Clin Med Res. 2017;9:461–465.
35. Takata Y, Ansai T, Soh I, Awano S, Sonoki K, Akifusa S, et al. Serum albumin levels as an independent predictor of 4-year mortality in a community-dwelling 80-year-old population. Aging Clin Exp Res. 2010;22:31–5.
36. Odden MC, Shlipak MG, Tager IB. Serum creatinine and functional limitation in elderly persons. J Gerontol Ser A, Biol Sci Med Sci. 2009;64:370–376.
37. Cabrera MA, de Andrade SM, Dip RM. Lipids and all-cause mortality among older adults: a 12-year follow-up study. Scientific World J. 2012;2012:930139.
38. Kadoglou NPE, Biddulph JP, Rafnsson SB, Trivella M, Nihoyannopoulos P, Demakakos P. The association of ferritin with cardiovascular and all-cause mortality in community-dwellers: the English longitudinal study of ageing. PLoS ONE. 2017; 12:e0178994.
39. Weiss A, Beloosesky Y, Boaz M, Yalov A, Kornowski R, Grossman E. Body mass index is inversely related to mortality in elderly subjects. J Gen Intern Med. 2008;23:19–24.
40. Celis-Morales CA, Welsh P, Lyall DM, Steell L, Petermann F, Anderson J, et al. Associations of grip strength with cardiovascular, respiratory, and cancer outcomes and all cause mortality: prospective cohort study of half a million UK Biobank participants. Br Med J. 2018;361:k1651.
41. Sin DD, Wu L, Man SF. The relationship between reduced lung function and cardiovascular mortality: a population-based study and a systematic review of the literature. Chest. 2005;127:1952–9.
42. Mannino DM, Davis KJ. Lung function decline and outcomes in an elderly population. Thorax. 2006;61:472–477.
43. Bates, D, Mächler M, Bolker B, Walker S, Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;1:1–48.
44. Wardlaw JM, Bastin ME, Valdes Hernandez MC, Maniega SM, Royle NA, Morris Z, et al. Brain aging, cognition in youth and old age and vascular disease in the Lothian Birth Cohort 1936: rationale, design and methodology of the imaging protocol. Int J Stroke. 2011;6:547–59.
45. Valdes Hernandez Mdel C, Gallacher PJ, Bastin ME, Royle NA, Maniega SM, Deary IJ, et al. Automatic segmentation of brain white matter and white matter lesions in normal aging: comparison of five multispectral techniques. Magn Reson Imaging. 2012;30:222–9.
46. Penke L, Maniega SM, Murray C, Gow AJ, Valdés Hernández MC, Clayden JD, et al. A General factor of brain white matter integrity predicts information processing speed in healthy older people. J Neurosci. 2010;30:7569–7574.
47. Deary IJ, Gow AJ, Taylor MD, Corley J, Brett C, Wilson V, et al. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. BMC Geriatr. 2007;7:28.

48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.
49. Franceschi C, Bonafe M, Valensin S, Olivieri F, De Luca M, Ottaviani E, et al. Inflamm-aging. An evolutionary perspective on immunosenescence. Ann N. Y Acad Sci. 2000;908:244–54.
50. Weverling-Rijnsburger AW, Blauw GJ, Lagaay AM, Knook DL, Meinders AE, Westendorp RG. Total cholesterol and risk of mortality in the oldest old. Lancet. 1997;350:1119–23.
51. Paul RF, Hassan M, Nazar HS, Gillani S, Afzal N, Qayyum I. Effect of body mass index on serum leptin levels. J Ayub Med Coll Abbottabad. 2011;23:40–3.
52. Al Maskari MY, Alnaqdy AA. Correlation between Serum Leptin Levels, Body Mass Index and Obesity in Omanis. Sultan Qaboos Univ Med J. 2006;6:27–31.
53. Horn JL, Cattell RB. Age differences in fluid and crystallized intelligence. Acta Psychol. 1967;26:107–29.
54. Dykiert D, Deary IJ. Retrospective validation of WTAR and NART scores as estimators of prior cognitive ability using the Lothian Birth Cohort 1936. Psychol Assess. 2013;25:1361–6.
55. Zhao, L, Matloff W, Ning K, Kim H, Dinov ID, Toga AW, Age-related differences in brain morphology and the modifiers in middle-aged and older adults. Cereb Cortex. 2018;29:4169–93.
56. Aycheh HM, Seong JK, Shin JH, Na DL, Kang B, Seo SW, et al. Biological brain age prediction using cortical thickness data: a large scale cohort study. Front Aging Neurosci. 2018;10:252.
57. Dickie DA, Karama S, Ritchie SJ, Cox SR, Sakka E, Royle NA, et al. Progression of white matter disease and cortical thinning are not related in older community-dwelling subjects. Stroke. 2016;47:410–416.
58. Sachdev PS, Wen W, Christensen H, Jorm AF. White matter hyperintensities are related to physical disability and poor motor function. J Neurol Neurosurg Psychiatry. 2005;76:362–367.
59. Salarirad S, Staff RT, Fox HC, Deary IJ, Whalley L, Murray AD. Childhood intelligence and brain white matter hyperintensities predict fluid intelligence age 78-81 years: a 1921 Aberdeen birth cohort study. Age Ageing. 2011;40:562–7.
60. Wardlaw JM, Chappell FM, Valdés Hernández MDC, Makin SDJ, Staals J, Shuler K, et al. White matter hyperintensity reduction and outcomes after minor stroke. Neurology. 2017;89:1003–1010.
61. Debette S, Markus HS. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. Br Med J. 2010;341:c3666.
62. Onyike CU. Psychiatric aspects of dementia. Continuum. 2016; 22:600–614.
63. Ohi K, Sumiyoshi C, Fujino H, Yasuda Y, Yamamori H, Fujimoto M, et al. Genetic overlap between general cognitive function and schizophrenia: a review of cognitive GWASs. Int J Mol Sci. 2018;19:3822.
64. Meijsen JJ, Campbell A, Hayward C, Porteous DJ, Deary IJ, Marioni RE, et al. Phenotypic and genetic analysis of cognitive performance in Major Depressive Disorder in the Generation Scotland: Scottish Family Health Study. Transl Psychiatry. 2018;8:63.
65. Sabia S, Elbaz A, Dugravot A, Head J, Shipley M, Hagger-Johnson G, et al. Impact of smoking on cognitive decline in early old age: the Whitehall II cohort study. Arch Gen psychiatry. 2012;69:627–635.
66. Gellert C, Schottker B, Brenner H. Smoking and all-cause mortality in older people: systematic review and meta-analysis. Arch Intern Med. 2012;172:837–44.
67. Power MC, Deal JA, Sharrett AR, Jack Jr CR, Knopman D, Mosley TH, et al. Smoking and white matter hyperintensity progression: the ARIC-MRI study. Neurology. 2015;84:841–848.
68. Zhang Y, Elgizouli M, Schöttker B, et al. DNA methylation markers predict lung cancer incidence. Clin Epigenetics. 2016; 8:127–127.

Affiliations

Robert F. Hillary[1] · Anna J. Stevenson[1] · Simon R. Cox[2,3] · Daniel L. McCartney[1] · Sarah E. Harris[2,3] · Anne Seeboth[1] · Jon Higham[4] · Duncan Sproul[4,5] · Adele M. Taylor[2,3] · Paul Redmond[2,3] · Janie Corley[2,3] · Alison Pattie[2,3] · Maria del. C. Valdés Hernández[2,6] · Susana Muñoz-Maniega[2,6] · Mark E. Bastin[2,6] · Joanna M. Wardlaw[2,6,7] · Steve Horvath[8,9] · Craig W. Ritchie[10] · Tara L. Spires-Jones[7,11] · Andrew M. McIntosh[2,12] · Kathryn L. Evans[1,2] · Ian J. Deary[2,3] · Riccardo E. Marioni[1,2]

[1] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

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

[3] Department of Psychology, University of Edinburgh, Edinburgh, UK

[4] Medical Research Council Human Genetics Unit, Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

[5] Edinburgh Cancer Research Centre, Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

[6] Department of Neuroimaging Sciences, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

[7] UK Dementia Research Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK

[8] Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, USA

[9] Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, California, USA

[10] Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

[11] Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK

[12] Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

8.3 Conclusion

In their initial report of DNAm GrimAge, Lu *et al.* determined that DNAm GrimAge associated with several peripheral and metabolic traits. I supplemented these data by conducting a comprehensive investigation of relationships between DNAm GrimAge and cognitive and neuroimaging phenotypes. Age-adjusted DNAm GrimAge showed strong associations with all-cause mortality, neurological protein biomarkers and many measures of frailty and brain health. I also provided the first external report of an association between DNAm GrimAge and all-cause mortality. Rezwan *et al.* (2020) and Maddock *et al.* (2020) replicated associations between DNAm GrimAge and lung function (460, 461). Furthermore, Maddock *et al.* (2020) reported that age-adjusted DNAm GrimAge was cross-sectionally associated with episodic memory and mental speed (P < 0.001) (460). McCrory *et al*. (2021) showed that DNAm GrimAge outperformed other epigenetic measures of ageing (Horvath Age, Hannum Age and DNAm PhenoAge) in its associations with measures of cognitive ability (462).

Approximately 50 of the phenotypic associations in this study were independent of childhood intelligence and years of education. The possibility of reverse causation cannot be ruled out. However, my analyses suggest that blood methylation profiles reflected in DNAm GrimAge could supplement established health risk factors in predicting late-life frailty and cognitive ability. There was tentative evidence for an association between DNAm GrimAge and cognitive decline from age 70 to 79 years.

Whereas age-adjusted DNAm GrimAge correlated with poorer cognitive ability and brain atrophy, it is unknown whether this ageing biomarker associates with AD. In the next chapter, I assess whether DNAm GrimAge and other epigenetic measures of ageing predict the prevalence and incidence of AD and co-morbidities including type 2 diabetes, depression and CVD.

9 Associations between DNAm GrimAge and common disease states

9.1 Introduction

An accelerated DNAm GrimAge has been associated with the incidence of CVD (313, 379, 463), cancer (313, 376), type 2 diabetes (464) and death due to oropharyngeal cancer (378). Age-adjusted DNAm GrimAge was associated with a reduced risk of all-cause dementia in the Lothian Birth Cohort of 1921 (n = 387, no. of events = 240, HR = 0.89, P < 0.05). The authors concluded that this unexpected association was likely explained by collinearity between DNAm GrimAge and smoking in this sample. The relationship between DNAm GrimAge and causes of dementia such as AD is unclear. Furthermore, no study has examined relationships between DNAm GrimAge and the incidence of multiple common diseases.

Here, I examine associations between DNAm GrimAge and the prevalence and incidence of ten leading causes of death and disease burden, including AD. I utilise blood DNAm and electronic health record linkage data from GS participants (n ≤ 9,537). This is the single largest study on DNAm GrimAge and health outcomes in terms of sample size and the number of conditions considered. I also compare DNAm GrimAge against five other epigenetic ageing biomarkers with respect to their associations with the prevalence and incidence of common disease states. These epigenetic measures of ageing are Horvath Age, Hannum Age, DNAm PhenoAge, DNAm Telomere Length and DunedinPoAm, which are described in Section 2.6.2.

This study was published in *Clinical Epigenetics* (465) in July 2020 and is included in full in Section 9.2.

9.2 Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden

Clinical Epigenetics

RESEARCH Open Access

Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden

Robert F. Hillary[1], Anna J. Stevenson[1], Daniel L. McCartney[1], Archie Campbell[1], Rosie M. Walker[1], David M. Howard[2,3], Craig W. Ritchie[4], Steve Horvath[5,6], Caroline Hayward[7], Andrew M. McIntosh[1,3], David J. Porteous[1], Ian J. Deary[8], Kathryn L. Evans[1] and Riccardo E. Marioni[1*]

Abstract

Background: Individuals of the same chronological age display different rates of biological ageing. A number of measures of biological age have been proposed which harness age-related changes in DNA methylation profiles. These measures include five 'epigenetic clocks' which provide an index of how much an individual's biological age differs from their chronological age at the time of measurement. The five clocks encompass methylation-based predictors of chronological age (HorvathAge, HannumAge), all-cause mortality (DNAm PhenoAge, DNAm GrimAge) and telomere length (DNAm Telomere Length). A sixth epigenetic measure of ageing differs from these clocks in that it acts as a speedometer providing a single time-point measurement of the pace of an individual's biological ageing. This measure of ageing is termed DunedinPoAm. In this study, we test the association between these six epigenetic measures of ageing and the prevalence and incidence of the leading causes of disease burden and mortality in high-income countries ($n \leq 9537$, Generation Scotland: Scottish Family Health Study).

Results: DNAm GrimAge predicted incidence of clinically diagnosed chronic obstructive pulmonary disease (COPD), type 2 diabetes and ischemic heart disease after 13 years of follow-up (hazard ratios = 2.22, 1.52 and 1.41, respectively). DunedinPoAm predicted the incidence of COPD and lung cancer (hazard ratios = 2.02 and 1.45, respectively). DNAm PhenoAge predicted incidence of type 2 diabetes (hazard ratio = 1.54). DNAm Telomere Length associated with the incidence of ischemic heart disease (hazard ratio = 0.80). DNAm GrimAge associated with all-cause mortality, the prevalence of COPD and spirometry measures at the study baseline. These associations were present after adjusting for possible confounding risk factors including alcohol consumption, body mass index, deprivation, education and tobacco smoking and surpassed stringent Bonferroni-corrected significance thresholds.

Conclusions: Our data suggest that epigenetic measures of ageing may have utility in clinical settings to complement gold-standard methods for disease assessment and management.

Keywords: DNA methylation, Biological ageing, Epigenetic age acceleration, Epidemiology

* Correspondence: riccardo.marioni@ed.ac.uk
[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK
Full list of author information is available at the end of the article

Background

The sustained increase in global life expectancy and population size has prompted a concomitant elevation in the prevalence of chronic disease states [1]. The World Health Organisation specifies ten leading causes of mortality and ten leading causes of disease burden. In high-income countries, six diseases are present in both sets: ischemic heart disease, stroke, lung cancer, Alzheimer's disease (AD) and other dementias, diabetes and chronic obstructive pulmonary disease (COPD). The remaining four leading causes of mortality are lower respiratory tract diseases, bowel cancer, kidney disease and breast cancer [2]. The additional four causes of disease burden are back or neck pain, skin disease, sense organ disease and depression [3]. Many of these disease states encompass heterogeneous, complex aetiologies resulting in a paucity of effective treatment paradigms. Given the number of individuals affected by such disorders and the associated burden, there is an urgent need for effective molecular predictors in clinical settings that can identify individuals on trajectories towards disease.

Ageing is a major risk factor for many common disease states. However, individuals of the same chronological age exhibit disparate rates of biological ageing and susceptibilities to common morbidities and mortality. Differential patterns of biological ageing among individuals may be exploited to identify novel predictors of disease [4]. Recently, a number of strategies have been proposed to estimate biological age by leveraging inter-individual variation in DNA methylation (DNAm) profiles. These epigenetic measures of ageing, many of which are called 'epigenetic clocks', correlate strongly with chronological age [5]. Moreover, for a given chronological age, an accelerated epigenetic age or faster rate of ageing is associated with an increased risk of mortality and shows cross-sectional relationships with age-related morbidities [6–10].

In this paper, we focus on six epigenetic predictors of ageing. In 2013, Horvath developed a pan-tissue epigenetic clock, termed 'Horvath Age', derived from the linear combination of 353 CpG sites in multiple tissues [11]. Hannum created a DNAm-based clock termed 'Hannum Age' based on 71 CpG sites in blood tissue [12]. Levine et al. proposed a predictor of lifespan and health by developing a methylation-based predictor of an individual's 'phenotypic age' ('DNAm PhenoAge'). Phenotypic age is informed by chronological age as well as haematological and biochemical measures, including creatinine levels and lymphocyte percent [13]. Lu et al. proposed 'DNAm GrimAge' as a predictor of mortality and demonstrated that it outperforms existing clocks in predicting death and age-related conditions, including cardiovascular disease [14]. DNAm GrimAge was developed in two stages. In the first stage, DNAm-based surrogates for 88 plasma protein levels and smoking pack years were developed using elastic net regression models. Only DNAm-based surrogates which exhibited a correlation coefficient of at least 0.35 with their respective phenotype were considered for stage two. In addition to smoking pack years, 12/88 protein proxies satisfied this condition. In the second stage, time-to-death due to all-cause mortality was regressed on chronological age, sex, DNAm-based surrogates for smoking pack years and 12 plasma protein levels. The model selected chronological age, sex, DNAm-based proxies for smoking pack years and the levels of 7/12 plasma proteins; the linear combination of these variables provides a measure of DNAm GrimAge. Furthermore, telomere length is associated with cardiovascular disease, cancer risk and all-cause mortality [15–17]. Lu et al. proposed a DNAm-based estimator of telomere length termed 'DNAm Telomere Length' (DNAm TL) which exhibits stronger associations with lifespan, smoking history and body mass index when compared to phenotypic telomere length as measured by quantitative polymerase chain reaction or Southern blotting [18]. These five measures of biological ageing provide an index of the difference between an individual's biological age and chronological age at the time of measurement and are derived from cross-sectional measurements across individuals of varying ages. This can approximate a longitudinal ageing trajectory but can also be confounded by the possibility that individuals who were born in different years may have been exposed to different early-life exposures [4]. As a result, individuals of different chronological ages may display differential DNAm patterns that do not reflect age-related DNAm changes but rather reflect differential early-life environmental influences. To address this, Belsky et al. (2015) developed a longitudinal measure of biological age by examining the rate of change in 18 blood-chemistry and organ-system-function biomarkers at three successive time points from ages 26 to 38 in participants of the Dunedin study (n = 954) [19]. This measure was termed 'Pace of Aging' (PoA), and all of the individuals in the sample were born in 1972–1973. Recently, Belsky et al. (2020) derived a DNAm-based proxy of PoA termed 'DunedinPoAm' [20]. The authors state that this measure reflects a speedometer which tracks how fast the subject is ageing whereas the previous measures of epigenetic ageing represent clocks recording how much time has passed.

For epigenetic clocks, the difference between an individual's methylation-based age and their chronological age provides a measure of accelerated or decelerated ageing. DunedinPoAm returns a measure in years of biological ageing per each calendar year with higher values reflecting a faster rate of ageing. Higher values of DunedinPoAm as well as age-adjusted Horvath Age, Hannum

Age, DNAm PhenoAge and DNAm GrimAge are hypothesised to associate with poorer health outcomes as these measures capture accelerated biological ageing. Lower values of age-adjusted DNAm TL are hypothesised to correlate with poorer health as this reflects shorter telomere length. To date, a number of studies have demonstrated associations between epigenetic measures of ageing and risk of mortality and disease states [21–24] or have provided comparisons of such epigenetic measures [25–30]. However, no study has compared all six epigenetic measures of ageing with respect to their association with a broad range of common health conditions.

In this study, we test the association between the six epigenetic measures of ageing and the prevalence, and incidence, of the ten leading causes of mortality and disease burden (as indexed by disability-adjusted life years; DALYs) [2, 3]. In addition, we examine their association with continuous traits underlying these conditions, such as lung function tests for chronic obstructive pulmonary disease (COPD). We utilise DNA methylation array data and electronic health record data from a Scottish cohort, Generation Scotland: Scottish Family Health Study (GS: SFHS or GS). GS is a family-based cohort consisting of over 20,000 individuals with rich health and lifestyle information. Genome-wide methylation data were generated on approximately 10,000 participants making it one of the largest DNAm resources in the world. We examine associations between epigenetic measures of ageing and prevalent disease as well as an assessment of their ability to predict time-to-disease onset. These findings may expedite the future use and refinement of large-scale molecular data-based approaches for predicting clinically defined outcomes and subsequent individual disease risk prediction.

Results

Demographics and epigenetic measures of ageing

In the discovery cohort, 56.3% of the participants were female with a mean age of 51.4 years (standard deviation (SD) = 13.2) (*n* = 4450). The mean values for epigenetic measures of ageing were as follows: Horvath age (60.1 years, SD = 9.8), Hannum age (47.4 years, SD = 9.6), DNAm PhenoAge (43.7 years, SD = 11.5), DNAm GrimAge (48.8 years, SD = 10.9), DNAm Telomere Length (7.4 kilobase pairs, SD = 0.3), and DunedinPoAm (1.1 years of biological ageing per each calendar year, SD = 0.1). Summary data for all variables in this study are presented in Additional file 1.

In the replication cohort, 61.4% were female with a mean age of 50.0 years (SD = 12.5) (*n* = 2578). Values for all phenotypes were comparable between discovery and replication cohorts with the exception of DNAm GrimAge (discovery: 48.8 years, SD = 10.9, replication: 60.5 years, SD = 10.6), and the incidence of self-reported depression (discovery: 8.4%, replication: 16.4%), and SCID (Structured Clinical Interview for DSM)-identified Depression (discovery: 18.5%, replication: 38.2%). This disparity in depression prevalence reflects an over-sampling of depression cases in the replication cohort. It is unclear as to why the replication cohort shows a higher mean DNAm GrimAge. However, it is possible that this difference may be driven by a latent aspect of poorer overall health that may be associated or correlated with depression.

Epigenetic measures of ageing and disease prevalence

In a basic model adjusting for age and sex, 51 phenotypes were significant at Bonferroni-corrected levels of significance in both the discovery and replication cohorts (Additional file 2: Note 1 and Additional file 3: Tables S1-S4). In the discovery cohort, a Bonferroni-corrected threshold of $P < 2.54 \times 10^{-4}$ was applied as this corrected for all tests performed (0.05/197 tests). Of these 197 models, 78 were significant at $P < 2.54 \times 10^{-4}$ in the discovery set and were therefore carried forward to the replication stage. In the replication set, associations which surpassed a Bonferroni-corrected threshold of $P < 6.41 \times 10^{-4}$ were deemed significant (0.05/78 tests). Additional file 4: Fig. S1-S3 highlight significant associations present in both sets for categorical traits, continuous traits and all-cause mortality, respectively. A measure-by-measure comparison of associations with categorical and continuous phenotypes from fully adjusted models in the replication cohort, stratified by disease type, is shown in Additional file 5. For all models, beta coefficients for continuous traits were correlated 0.96 between discovery and replication sets. For categorical phenotypes, the correlation coefficient for log odds was 0.79 between sets (Additional file 4: Fig. S4).

Fifteen relationships remained significant in both discovery and replication sets in a fully adjusted model accounting for age, sex and five common risk factors (Additional file 3: Tables S5 and S6, respectively). Those relationships which were significant in both cohorts at a Bonferroni-corrected significance threshold of $P < 6.41 \times 10^{-4}$ (reflecting the same stringent threshold as above) are reported herein and presented in Table 1 and Fig. 1.

Associations with disease

In relation to prevalent disease data, only the association between an accelerated DNAm GrimAge and COPD remained significant in both cohorts in the fully adjusted model (replication cohort: odds ratio (OR) per SD = 3.29, 95% confidence interval (CI) = [1.73, 6.30], $P = 3.4 \times 10^{-4}$; Fig. 1)

Table 1 Significant and replicated relationships between epigenetic age measures and prevalent disease data, and continuous traits

		Discovery cohort			Replication cohort		
Categorical phenotypes							
Measure	Variable	*n* event	OR	*P*	*n* event	OR	*P*
DNAm GrimAge	COPD	48	2.00	1.0×10^{-4}	32	3.29	3.4×10^{-4}
Continuous phenotypes							
Measure	Variable	*n*	β	*P*	*n*	β	*P*
DunedinPoAm	Pack Years	2419	0.45	1.2×10^{-112}	1340	0.33	7.6×10^{-36}
DNAm TL	Pack Years	2419	-0.14	1.1×10^{-11}	1340	-0.18	2.7×10^{-11}
DNAm PhenoAge	Pack Years	2419	0.11	3.0×10^{-08}	1340	0.16	9.5×10^{-10}
DNAm GrimAge	SIMD	2419	-0.13	5.9×10^{-08}	1340	-0.19	4.6×10^{-09}
DNAm GrimAge	Average Heart Rate	2416	0.19	1.4×10^{-12}	1339	0.20	1.6×10^{-08}
DunedinPoAm	SIMD	2419	-0.13	1.8×10^{-09}	1340	-0.16	2.2×10^{-08}
Hannum Age	Creatinine	2406	0.21	1.4×10^{-26}	1334	0.13	4.2×10^{-07}
DNAm GrimAge	FEF	2055	-0.12	1.2×10^{-06}	1149	-0.15	1.4×10^{-06}
DNAm PhenoAge	Body Mass Index	2419	0.12	2.5×10^{-10}	1340	0.12	7.4×10^{-06}
DNAm PhenoAge	Average Heart Rate	2416	0.11	2.1×10^{-07}	1339	0.12	7.7×10^{-06}
DNAm GrimAge	FEV	2074	-0.08	2.0×10^{-05}	1151	-0.10	1.4×10^{-04}
DNAm GrimAge	Creatinine	2406	0.19	3.0×10^{-15}	1334	0.13	2.0×10^{-04}
DunedinPoAm	Average Heart Rate	2416	0.19	1.6×10^{-15}	1339	0.11	2.2×10^{-04}
Mortality analysis							
Measure	Variable	*n* event	HR	*P*	*n* events	HR	*P*
DNAm GrimAge	All-cause mortality	89	1.62	1.4×10^{-4}	30	2.10	5.6×10^{-4}

Analyses were performed using a fully adjusted model accounting for age, sex, alcohol consumption, body mass index, deprivation, education and smoking pack years

COPD chronic obstructive pulmonary disease, *FEF* forced expiratory flow, *FEV* forced expiratory volume, *HR* hazard ratio, *OR* odds ratio, *SIMD* Scottish Index of Multiple Deprivation, *TL* telomere length

Associations with all-cause mortality

An accelerated DNAm GrimAge alone was associated with all-cause mortality following adjustment for the lifestyle risk factors (replication cohort: hazard ratio (HR) per SD = 2.10, 95% CI = [1.36, 3.25], $P = 5.6 \times 10^{-4}$; Fig. 1).

Associations with continuous clinically associated traits

An accelerated DNAm GrimAge was associated with greater deprivation (a lower Scottish Index of Multiple Deprivation (SIMD) rank; $\beta_{\text{replication}} = -0.19$, 95% CI = [-0.25, -0.13], $P = 4.6 \times 10^{-9}$), an increased average heart rate ($\beta_{\text{replication}} = 0.20$, 95% CI = [0.13, 0.27], $P = 1.6 \times 10^{-8}$), a reduced forced expiratory flow ($\beta_{\text{replication}} = -0.15$, 95% CI = [-0.21, -0.09], $P = 1.4 \times 10^{-6}$), a reduced forced expiratory volume ($\beta_{\text{replication}} = -0.10$, 95% CI = [-0.15, -0.05], $P = 1.4 \times 10^{-4}$) and increased serum creatinine levels ($\beta_{\text{replication}} = 0.13$, 95% CI = [0.06, 0.20], $P = 2.0 \times 10^{-4}$).

Higher values of DunedinPoAm, indicating a faster rate of ageing, were positively associated with smoking pack years ($\beta_{\text{replication}} = 0.33$, 95% CI = [0.28, 0.38], $P = 7.6 \times 10^{-36}$), greater deprivation (lower SIMD rank; $\beta_{\text{replication}} = -0.16$, 95% CI = [-0.22, -0.10], $P = 2.2 \times 10^{-8}$) and average heart rate ($\beta_{\text{replication}} = 0.11$, 95% CI = [0.05, 0.17], $P = 2.2 \times 10^{-4}$).

An accelerated DNAm PhenoAge was associated with smoking pack years ($\beta_{\text{replication}} = 0.16$, 95% CI = [0.11, 0.21], $P = 9.5 \times 10^{-10}$), an increased body mass index ($\beta_{\text{replication}} = 0.12$, 95% CI = [0.07, 0.17], $P = 7.4 \times 10^{-6}$) and an increased average heart rate ($\beta_{\text{replication}} = 0.12$, 95% CI = [0.07, 0.17], $P = 7.7 \times 10^{-6}$).

Age-adjusted DNAm Telomere Length was negatively associated with smoking pack years ($\beta_{\text{replication}} = -0.18$, 95% CI = [-0.23, -0.13], $P = 2.7 \times 10^{-11}$). An accelerated DNAm Hannum Age (EEAA) was associated with increased serum creatinine levels ($\beta_{\text{replication}} = 0.13$, 95% CI = [0.08, 0.18], $P = 4.2 \times 10^{-7}$).

Covariate-specific attenuation

To examine the contribution of each of the five common disease risk factors in attenuating the 51 significant associations brought forward to the fully adjusted model, we repeated each model including only one of these five covariates at a time. These risk factors were alcohol consumption, body mass index, deprivation, education and

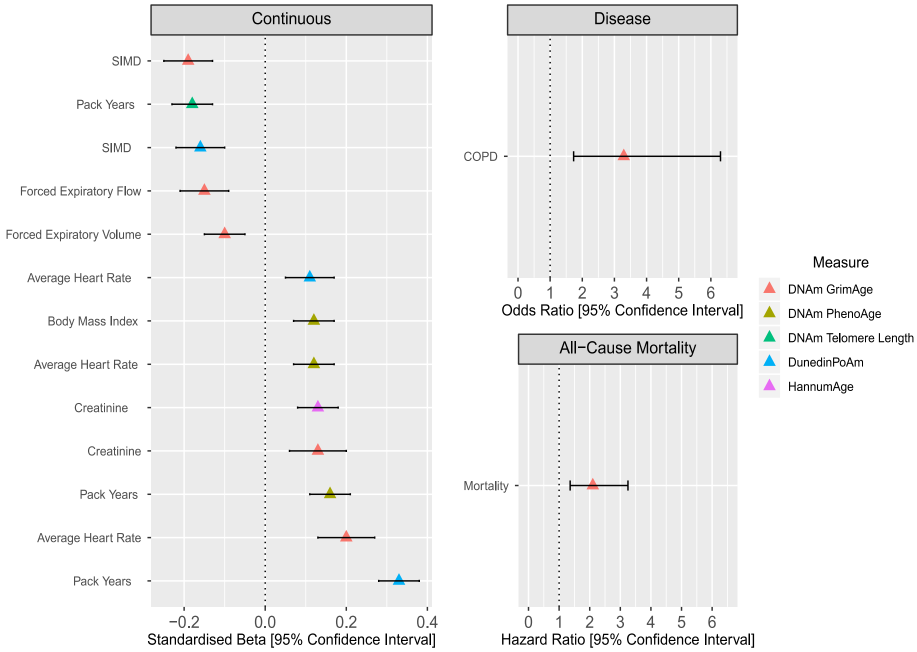

Fig. 1 The associations between epigenetic measures of ageing and disease prevalence, continuous traits and all-cause mortality in Generation Scotland. Only associations present in discovery and replication sets are shown, and replication test statistics are presented. *Continuous:* Age-adjusted DNAm GrimAge was associated with greater deprivation (lower SIMD rank), reduced forced expiratory flow and forced expiratory volume. Age-adjusted DNAm GrimAge was positively associated with serum creatinine levels and average heart rate. Age-adjusted DNAm PhenoAge was positively associated with body mass index, average heart rate and smoking pack years. Age-adjusted DNAm Telomere Length was negatively associated with smoking pack years. Higher values for DunedinPoAm were associated with greater deprivation (lower SIMD rank), a higher average heart rate and smoking pack years. Age-adjusted Hannum Age was positively associated with serum creatinine levels. *Disease:* Age-adjusted DNAm GrimAge alone was associated with the prevalence of COPD in both discovery and replication sub-cohorts after correction for multiple testing. *All-Cause Mortality:* Age-adjusted DNAm GrimAge alone was associated with all-cause mortality in both sets after multiple testing correction. Associations represent a one standard deviation increase in the respective measure of biological ageing. Models were adjusted for age, sex, alcohol consumption, body mass index, deprivation, education and smoking. Models involving lung function tests were also corrected for height. COPD (chronic obstructive pulmonary disease), SIMD (Scottish Index of Multiple Deprivation)

smoking pack years. The ranges of mean attenuation in traits by these covariates were 9.5 to 15.1% in the discovery set and 4.7 to 20.3% in the replication set (Additional file 3: Tables S7 and S8, respectively). Smoking pack years exhibited the greatest mean attenuation in both cohorts (discovery = 15.1%, replication = 20.3%).

Epigenetic measures of ageing and disease incidence

For incident disease outcomes, there were 17 Bonferroni-corrected significant associations at $P < 8.33 \times 10^{-4}$ ($P < 0.05/60$ tests; full output in Additional file 3: Table S9, see also Additional file 4: Fig. S5 and Additional file 6: Note 2). Of these, 7 remained significant in a fully adjusted model at a Bonferroni-corrected significance threshold of 8.33×10^{-4} (Additional file 3: Table S10). These relationships are presented herein and in Fig. 2.

A one standard deviation increase in DNAm GrimAge at baseline was associated with the incidence of COPD (HR = 2.22, 95% CI = [1.81, 2.72], $P = 2.4 \times 10^{-14}$), type 2 diabetes (HR = 1.52, 95% CI = [1.20, 1.90], $P = 3.1 \times 10^{-4}$) and heart disease (HR = 1.41, 95% CI = [1.18, 1.68], $P = 1.1 \times 10^{-4}$). Higher values of DunedinPoAm (per SD) associated with the incidence of COPD (HR = 2.02, 95% CI = [1.59, 2.57], $P = 8.4 \times 10^{-9}$) and lung cancer (HR = 1.45, 95% CI = [1.18, 1.79], $P = 5.3 \times 10^{-4}$). An accelerated DNAm PhenoAge (per SD) associated with a higher incidence of type 2 diabetes (HR = 1.54, 95% CI = [1.21, 1.97], $P = 4.5 \times 10^{-4}$). Age-adjusted DNAm Telomere Length (per SD) associated with a lower incidence of heart disease (HR = 0.80, 95% CI = [0.69, 0.92], $P = 2.5 \times 10^{-4}$).

Sex-specific analyses of epigenetic measures of ageing and phenotypes in Generation Scotland

As the occurrence of common diseases differs between the sexes, we ran sensitivity analyses using cross-sectional data to determine the correlation between effect sizes for males versus females. In the discovery cohort, continuous phenotypes had a correlation

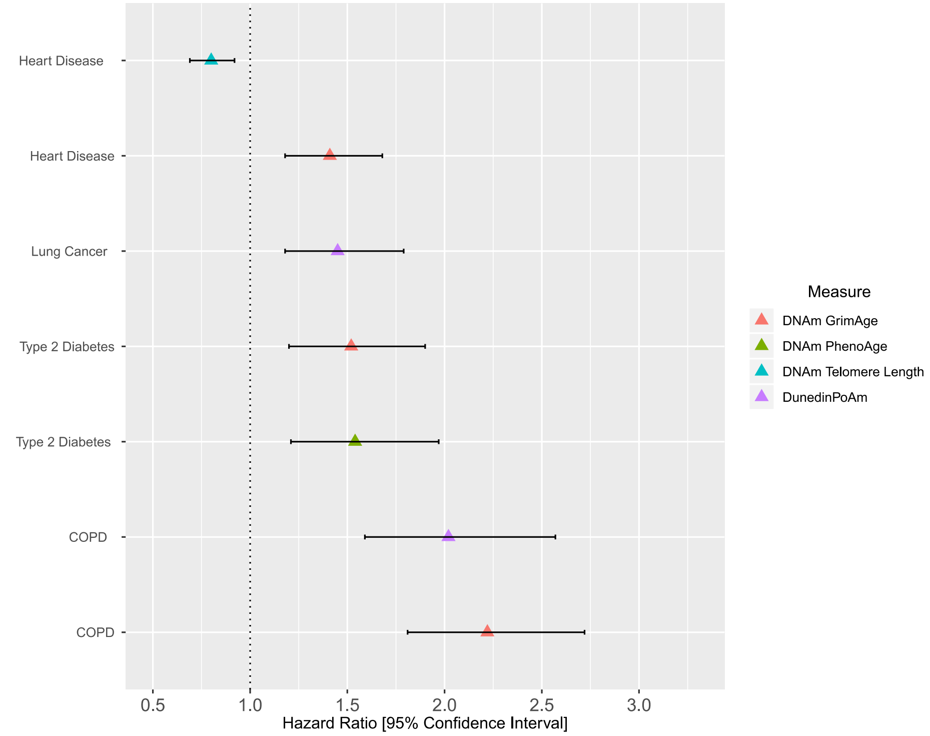

Fig. 2 The associations between epigenetic measures of ageing and incidence of common disease states in Generation Scotland. Age-adjusted DNAm GrimAge was associated with the incidence of COPD, type 2 diabetes and ischemic heart disease after 13 years of follow-up. Age-adjusted DNAm PhenoAge associated with the incidence of type 2 diabetes. Age-adjusted measures of DNAm Telomere Length associated with the incidence of ischemic heart disease. Higher DunedinPoAm values, indicating a faster pace of ageing, were associated with the incidence of COPD and lung cancer. Associations represent a one standard deviation increase in the respective epigenetic measure of ageing. Models were adjusted for age, sex, alcohol consumption, body mass index, deprivation, education and smoking. COPD (chronic obstructive pulmonary disease)

coefficient of 0.93 between sexes whereas categorical disease phenotypes exhibited a correlation coefficient of 0.81 (Additional file 4: Fig. S6). In the replication cohort, there was a correlation of 0.86 and 0.70 between effect sizes for continuous and categorical phenotypes, respectively (Additional file 4: Fig. S7). Excluding diseases ≤ 10 cases (lung and bowel cancer), the largest difference between males and females was for the DunedinPoAm-chronic kidney disease relationship (males: no. of events = 40, OR = 1.23, females: no. of events = 45, OR = 1.72, absolute difference = 0.49). On average, the largest difference between males and females across measures was observed for COPD with males having a higher odds ratio for each measure of ageing (mean difference in effect sizes across measures = 0.25, range = [0.12, 0.37], discovery cohort; Additional file 3: Table S11).

Discussion

In this study, we examined associations between six major epigenetic measures of ageing and the prevalence and incidence of the leading causes of mortality and disease burden in high-income countries. DNAm GrimAge, a predictor of mortality, associated with the prevalence of COPD and incidence of various disease states, including COPD, type 2 diabetes and cardiovascular disease. It was associated with death due to all-cause mortality and outperformed competitor epigenetic measures of ageing in capturing variability across clinically associated continuous traits. Higher values for DunedinPoAm, which captures faster rates of biological ageing, associated with the incidence of COPD and lung cancer. Higher-than-expected DNAm PhenoAge predicted the incidence of type 2 diabetes in the present study. Age-adjusted measures of DNAm Telomere Length associated with the incidence of ischemic heart disease. Our results replicate previous cross-sectional findings between DNAm PhenoAge and body mass index, diabetes [21] and socioeconomic position (in a basic model) [28]. We also replicated associations between DNAm GrimAge and heart disease [14]. Lastly, we replicated the relationship between Hannum Age and creatinine [31] and between DNAmTLadjAge and smoking pack years [18]. This is also the first external study examining the association between DunedinPoAm and a wide range of health outcomes.

DNAm GrimAge served as a powerful correlate of various phenotypes in our study and has been previously shown to associate with incident heart disease, time-to-cancer and neurological health [14, 22]. DNAm GrimAge is derived from chronological age, sex and methylation-based surrogates of smoking pack years and seven plasma proteins (including DNAm-based estimators of plasminogen activator inhibitor 1, growth differentiation factor 15 and cystatin C). Here, we show that this blood-based epigenetic predictor of mortality risk is associated with poorer performance in lung function tests and predicted incidence of COPD. Compromised lung function has previously been linked to mortality [32, 33]. While it is possible that the associations are mainly driven by the inclusion of smoking pack years, DNAm GrimAge remained associated with COPD and spirometry tests when controlling for self-reported smoking pack years. Similarly, DunedinPoAm associated with time-to-onset of COPD and lung cancer. DunedinPoAm also demonstrated a strong correlation with smoking pack years in our study; however, the associations between DunedinPoAm and incident disease outcomes remained after adjusting for common disease risk factors, including smoking behaviour. In their original study, Belsky et al. identified that the *AHRR* probe cg05575921 was among the 46 CpG sites used to calculate DunedinPoAm. This probe has been strongly associated with smoking behaviour [34–41]. The authors also demonstrated that a version of DunedinPoAm calculated without this probe correlated 0.94 with the DunedinPoAm measure including all probes [20]. DNAm PhenoAge predicted the incidence of type 2 diabetes; however, this may reflect the inclusion of HbA1c in the phenotypic age measure which is used to diagnose diabetes. In our study, an epigenetic predictor of telomere length predicted time-to-onset of ischemic heart disease. A shorter leukocyte telomere length has been shown to associate with heart disease in diverse populations, suggesting that the DNAm Telomere Length predictor may capture key facets of this clinical association [42–44]. Our rich resource of genome-wide DNA methylation and longitudinal health data is the first to show the association of epigenetic measures of ageing with a wide range of common disease states, even after accounting for major confounding influences. These findings have implications for the potential utility of epigenetic measures of ageing in clinical settings.

The majority of our prevalent disease data relied on self-report. Self-report prevalence data have been shown to have a high degree of sensitivity and specificity [45]. Our incident data was obtained using ICD-10 codes from health record linkage. Strikingly, measures of biological ageing showed strong associations with the incidence of common diseases following 13 years of follow-up from the study baseline. These measures performed better at predicting incident rather than prevalent data. However, this may reflect the inclusion of health record-linked versus self-report data and the larger sample size in incidence analyses. Notably, the six epigenetic measures of ageing in our study are correlated with one another among study participants. As well as this, the incidence or prevalence of different disease states, as well as associated continuous traits, may be correlated with one another as they may reflect patterns of poor overall health and disease risk behaviours. Therefore, our application of Bonferroni-corrected significance thresholds is stringent and only captured the most high-confidence associations in our study. These associations were also independent of common disease risk factors and therefore may reflect important associations between age-related physiological changes and risk of disease.

An important limitation is the lack of adjustments for medication use, which may confound associations between epigenetic measures and chronic conditions. Furthermore, studies examining causality between the relationships shown are merited. It is also unclear whether the risk factors examined in this study play a causal role in driving associations between epigenetic measures of ageing and phenotypes, or whether these pleiotropically affect altered DNA methylation and adverse health outcomes. Genetic influences may contribute to differences in DNA methylation and the subsequent estimation of epigenetic age or pace of ageing; therefore, it is possible that our findings may not be generalisable to individuals of non-European ancestry [46, 47].

Conclusions

In conclusion, using a large cohort with rich health and DNA methylation data, we provide the first comparison of six major epigenetic measures of biological ageing with respect to their associations with leading causes of mortality and disease burden. DNAm GrimAge outperformed the other measures in its associations with disease data and associated clinical traits. This may suggest that predicting mortality, rather than age or homeostatic characteristics, may be more informative for common disease prediction. Thus, proteomic-based methods (as utilised by DNAm GrimAge) using large, physiologically diverse protein sets for predicting ageing and health may be of particular interest in future studies. Our results may help to refine the future use and development of biological age estimators, particularly in studies which aim to comprehensively examine their ability to predict stringent clinically defined outcomes. Our analyses suggest that epigenetic measures of ageing can predict the incidence of common disease states, even after accounting for major confounding risk factors. This may have significant implications for their potential utility in clinical settings to complement gold-standard methods of clinical disease assessment and management.

Methods

Generation Scotland

Details of the Generation Scotland (GS) study have been described previously [48, 49]. Briefly, the cohort includes 23,960 individuals, where most individuals (94.2%) have at least one other first-degree family member participating in the study. This encompasses 5573 families with a median family size of 3 (interquartile range = 2–5 members; excluding 1400 singletons without any relatives in the study). For prevalence analyses, the discovery cohort comprised unrelated GS participants with genome-wide methylation data ($n_{\text{discovery}}$ = 4450). The replication cohort was also derived from GS participants, unrelated to those in the discovery cohort, who had genome-wide DNA methylation measured in a separate batch (n = 5087). Within the replication cohort, 2578 participants were also unrelated to one another and these unrelated individuals were considered for cross-sectional analyses ($n_{\text{replication}}$ = 2578). For incidence analyses, all individuals with available methylation and phenotypic data in GS were considered (n = 4450 + 5087 = 9537).

DNA methylation and calculation of biological ageing measures

DNA methylation levels were measured using the Illumina HumanMethylationEPIC BeadChip Array on blood samples from GS participants. Further details on the processing of DNAm data and the calculation of the six measures of ageing, or pace of ageing, are outlined in Additional file 7; the five clocks (other than DunedinPoAm) were calculated using Horvath's online age calculator (https://dnamage.genetics.ucla.edu/). Normalised GS methylation data were uploaded as input for the algorithm. Data underwent a further round of normalisation by the age calculator. Briefly, Horvath Age provides an estimate of biological ageing termed "intrinsic epigenetic age acceleration (IEAA)" as it is independent of age-related changes in blood composition. IEAA is derived from regressing Horvath Age onto chronological age. In contrast, Hannum Age provides a measure of ageing referred to as "extrinsic epigenetic age acceleration (EEAA)" as it encompasses age-related changes in blood cell composition. EEAA is derived from regressing a weighted average of Hannum Age and three blood cell types (naive and exhausted cytotoxic T cells, and plasmablasts) onto chronological age. DNAm PhenoAge reflects an individual's 'Phenotypic Age' and, when regressed onto chronological age, provides an index of age acceleration termed 'AgeAccelPheno'. Similarly, when age-adjusted, DNAm GrimAge is termed 'AgeAccelGrim'. Lastly, age-adjusted 'DNAm Telomere Length' is referred to as 'DNAmTLadjAge'. DunedinPoAm was calculated using DNAm beta values as input and the *DunedinPoAm38* package in *R* developed by the original study's authors (https://github.com/danbelsky/DunedinPoAm38 [20]). The five aforementioned epigenetic clocks capture a state of accelerated or decelerated biological ageing reflecting how much ageing has occurred in the individual. However, DunedinPoAm was trained to provide a single time-point, blood-based measurement of the pace of biological ageing in individuals. DunedinPoAm is a DNAm-based proxy of the 'Pace of Aging' (PoA) measure. PoA was derived by examining the rate of change in 18 blood-chemistry and organ-system-function biomarkers at three successive time points in participants of Dunedin Study (n = 954). The participants were all born in 1972–1973 and were aged 26, 32 and 38 at the time of biomarker measurements. Mixed-effects growth modelling of longitudinal changes in biomarker levels among participants allowed for estimations of the rate of change in biomarker levels for each participant. The sum of random slopes for the biomarker levels (rate of change for each participant) provided a measure of PoA [19]. An elastic net regression model using DNAm data and PoA calculated at age 38 in participants of the Dunedin Study identified 46 CpG sites as informative for predicting PoA, thereby creating a single time-point measure of PoA called DunedinPoAm. DunedinPoAm reflects years of biological ageing per each calendar year. These six measures of biological ageing were input as independent variables in statistical models. Correlations between these predictors are shown in Additional file 4: Fig. S8 for the discovery and replication sets. The correlation structure between these predictors was similar in both sets. DNAmTLadjAge was negatively correlated with the other five indices of ageing (discovery: mean coefficient = − 0.34, range = − 0.12 to − 0.47). This negative correlation was present as shorter telomere lengths typically correspond to an advanced age. The mean correlation coefficient between the remaining five predictors was 0.35 (discovery: range = 0.07 to 0.73).

Phenotype preparation

For continuous phenotypes, outliers were defined as those values which were beyond 3.5 standard deviations from the mean for a given trait. These outliers were removed prior to analyses. Body mass index was log-transformed. To reduce skewness in the distribution of alcohol consumption and smoking pack years, a log(units +1) or log(pack years +1) transformation was performed. The interval from the start of the Q wave to the end of the T wave on electrocardiogram tests (QT interval) was corrected for heart rate. A general fluid ('*gf*') cognitive ability score was derived from principal components analysis of three tests examining different cognitive domains. These domains were processing speed (Wechsler Digit Symbol Substitution Test), verbal declarative memory (Wechsler Logical Memory Test) and verbal fluency (the phonemic verbal fluency test). To derive a general ('*g*') cognitive ability score, the principal component analysis was performed on the above three tests and a measure of crystallised intelligence: The Mill Hill Vocabulary test. The first unrotated principal components from these analyses were extracted and labelled as '*gf*' and '*g*', respectively.

For categorical phenotypes, we aimed to examine the ten leading causes of mortality in high-income countries [2]. We also aimed to examine the ten leading causes of disease burden, six of which overlap with the top causes of mortality. This represents fourteen diseases. We had self-report phenotypic information for the prevalence of nine of these diseases (Additional file 1); specifically, we lacked self-report information on lower respiratory diseases and kidney disease (mortality), skin and sense organ diseases (disease burden), and Alzheimer's disease (AD; present in both the leading causes of mortality and disease burden). We were able to use proxy phenotypes for two of these conditions. We used self-reported maternal history and paternal history as proxies for AD. For kidney disease, we estimated glomerular filtration rate (eGFR) from serum creatinine levels using the chronic kidney disease epidemiology collaboration CKD-EPI equation [50] from which we inferred the prevalence of chronic kidney disease (CKD). Individuals with an eGFR < 60 ml/min/1.73 m^2 were considered to have CKD. In addition to self-report depression, we also had available information on SCID (Structured Clinical Interview for DSM)-identified depression [51]. Lastly, we separated self-reported back and neck pain into distinct phenotypes for analyses. Together, this resulted in a total of fourteen disease phenotypes for prevalence analyses.

In relation to disease incidence, health record linkage was available for up to 13 years of follow-up since the study baseline (median time-of-onset from baseline = 5.75 years, range = [< 1 month, 13 years]). For each disease state, those individuals who self-reported disease at study baseline were excluded. For cancer, individuals present on the Scottish Cancer Registry (SMR06) were included as cases for incidence analyses. Additionally, for incident cancer analyses, individuals who were recorded on the General Acute Inpatient and Day Case - Scottish Morbidity Records (SMR01) were removed from the control set. For a given condition, individuals who self-reported no disease at study baseline but had prior evidence of diagnosis through health record linkage were removed from analyses. Discovery and replication cohorts were combined to consider all participants for follow-up and to provide a sufficient number of cases for analyses. For incident disease analyses, ICD-10-coded data were retrieved for the following ten conditions: AD, bowel cancer, breast cancer, COPD, depression, type 2 diabetes, dorsalgia (neck and back pain combined), ischemic heart disease, lung cancer and stroke. These reflect the disease states examined in the prevalence analyses with the exception of chronic kidney disease. Furthermore, the two proxies of AD, two measures of depression and separate measures of neck and back pain were replaced by single, clinically defined counterparts in the incidence analyses. Additional file 4: Fig. S9 shows a heatmap for effect sizes from Cox regression models between epigenetic measures of ageing and incident disease outcomes in a fully adjusted model.

Statistical analyses

Linear regression models were used to examine the association between continuous traits and age-adjusted epigenetic clock measures (reflecting the difference between an individual's estimated biological age and chronological age) or DunedinPoAm (reflecting the rate of biological ageing). In cross-sectional analyses, logistic regression was used to test the association between categorical disease phenotypes and these epigenetic measures of ageing. In longitudinal analyses, Cox proportional hazards regression models were used to examine whether measures of biological ageing were associated with the incidence of disease. Cox models were also used to examine whether these measures were associated with all-cause mortality in discovery and replication cohorts. There were 182 (4.09%) and 57 (2.25%) deaths in the discovery and replication sets, respectively. The proportional hazards assumption was tested using the *cox.zph()* function in the *survival* package in *R* [52, 53]. There was no strong evidence (P > 0.05) of assumption violation for the reported significant associations. Phenotypes were scaled to mean zero and unit variance. Continuous or categorical phenotypes were input as dependent variables with measures of biological ageing incorporated as independent variables.

In a basic model, all analyses were adjusted for chronological age and sex. Additional adjustments for height were carried out for measures of lung function. All significant tests from the basic model were then repeated adjusting for additional five covariates, which represent important risk factors for common diseases. These covariates were alcohol intake (units consumed/ week), body mass index, educational attainment, deprivation (Scottish Index of Multiple Deprivation) and tobacco smoking pack years.

a. *Basic model*: Phenotype ~ Epigenetic Measure + age + sex

b. *Fully adjusted model*: Phenotype ~ Epigenetic Measure + age + sex + alcohol units consumed per week + body mass index + educational attainment + Scottish Index of Multiple Deprivation + smoking pack years

In relation to cross-sectional prevalence data, the discovery analyses consisted of 33 phenotypes which were tested against every epigenetic measure of ageing (all-cause mortality, fourteen disease and eighteen continuous phenotypes; Additional file 3: Table S1). This led to a total of 198 (33 × 6 measures) tests; however, the DNAm GrimAge versus smoking pack years comparison was excluded given the inclusion of a DNAm-based surrogate of pack years in the development of DNAm GrimAge. This led to a Bonferroni-corrected significance threshold of $P < 0.05/197$ tests $= 2.54 \times 10^{-4}$. Of these 197 tests, 78 were significant; thus, in the replication cohort, a Bonferroni-corrected significance threshold of $P < 0.05/78$ tests $= 6.41 \times 10^{-4}$ was set. In total, 51 associations were significant in both cohorts. The fully adjusted model was then applied to these 51 associations in both the discovery and replication cohorts, holding the same stringent Bonferroni-corrected threshold of $P < 0.05/78$ tests $= 6.41 \times 10^{-4}$.

In relation to incidence data, all ten phenotypes were tested against each of the six measures of ageing. In the basic model, this resulted in a Bonferroni-corrected significance threshold of $P < 0.05/60$ tests $= 8.33 \times 10^{-4}$. In total, seventeen associations were significant and brought forward to the fully adjusted analysis stage. In the fully adjusted model, the same Bonferroni-corrected significance threshold of $P < 0.05/60$ tests $= 8.33 \times 10^{-4}$ was applied.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13148-020-00905-6.

Additional file 1. Demographics and Descriptive Statistics for Discovery and Replication Cohorts.

Additional file 2. Supplementary Note 1. Significant cross-sectional associations between phenotypes and epigenetic measures of ageing in both discovery and replication cohorts in a basic model adjusting for age and sex.

Additional file 3. Supplementary Tables. The associations between epigenetic measures of ageing and disease phenotypes in the discovery cohort (Bonferroni-corrected threshold: $P < 2.54 \times 10^{-4}$; significant results are emboldened). (Table S1). The associations between epigenetic measures of ageing and continuous phenotypes in the discovery cohort (Bonferroni-corrected threshold: $P < 2.54 \times 10^{-4}$; significant results are emboldened). (Table S2). The associations between epigenetic measures of ageing and all-cause mortality in the discovery cohort (Bonferroni-corrected threshold: $P < 2.54 \times 10^{-4}$; significant results are emboldened). (Table S3). Associations between significant phenotypes (identified in the discovery set) and epigenetic measures of ageing in the replication cohort at $P < 6.41 \times 10^{-4}$. (Table S4). Discovery Cohort: Associations between phenotypes (significant in basic model) and epigenetic measures of ageing in a fully-adjusted model (Bonferroni threshold: $P < 6.41 \times 10^{-4}$). (Table S5). Replication Cohort: Associations between phenotypes (significant in basic model) and epigenetic measures of ageing in a fully-adjusted model (Bonferroni threshold: $P < 6.41 \times 10^{-4}$). (Table S6). Covariate-specific analyses of trait-epigenetic age relationship attenuation in the discovery cohort. (Table S7). Covariate-specific analyses of trait-epigenetic age relationship attenuation in the replication cohort. (Table S8). The associations between epigenetic measures of ageing calculated at study baseline and ICD-10-coded incident disease data in Generation Scotland in a basic model adjusted for age and sex. Significant associations that survived a multiple testing correction threshold of 8.33×10^{-4} (0.05/60 tests) are emboldened. Nominally significant associations are italicised. (Table S9). The associations between epigenetic measures of ageing measured at study baseline and ICD-10-coded incident disease data in Generation Scotland in a fully-adjusted model adjusted for age, sex and common disease risk factors. Significant associations that survived a multiple testing correction threshold of 8.33×10^{-4} (0.05/60 tests) are emboldened. Nominally significant associations are italicised. (Table S10). Sex-specific differences in categorical phenotype-epigenetic age relationships within the discovery cohort. (Table S11).

Additional file 4. Significant associations between epigenetic measures of ageing and prevalent disease phenotypes present in both discovery and replication sets in a basic model adjusted for age and sex. (Fig. S1). Significant associations between epigenetic measures of ageing and continuous phenotypes present in both discovery and replication sets in a basic model adjusted for age and sex. (Fig. S2). Associations between epigenetic measures of ageing and all-cause mortality in both discovery (A) and replication (B) sets in a basic model adjusted for age and sex. (Fig. S3). Degree of correlation for continuous variables (A) or categorical variables (B) between discovery and replication cohorts. (Fig. S4). Significant associations between epigenetic measures of ageing and incidence of common disease states in Generation Scotland in a basic model adjusting for age and sex. (Fig. S5). Degree of correlation between males and females in relation to continuous variables (A) or categorical variables (B) in the discovery cohort. (Fig. S6). Degree of correlation between males and females in relation to continuous variables (A) or categorical variables (B) in the replication cohort. (Fig. S7). Correlation structure between different epigenetic measures of biological ageing in discovery (A) and replication (B) sets. (Fig. S8). Heatmap demonstrating the relationship between epigenetic measures of ageing and incident disease outcomes in a fully-adjusted Cox regression model in Generation Scotland. (Fig. S9).

Additional file 5. Comparison of epigenetic age measures in terms of their associations with categorical and continuous phenotypes from fully-adjusted models in the replication cohort, stratified by disease type.

Additional file 6. Supplementary Note 2. Associations between epigenetic measures of ageing and incidence of ICD-10-coded common diseases in a basic model adjusting for age and sex.

Additional file 7. Details of Supplementary Methods.

Abbreviations

AD: Alzheimer's disease; CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; CKD: Chronic kidney disease; DNAm: DNA methylation; DNAm TL: DNAm methylation-based estimator of telomere length; DSM: Diagnostic and Statistical Manual of Mental Disorders; EEAA: Extrinsic epigenetic age acceleration; eGFR: Estimated glomerular filtration rate; FEF: Forced expiratory flow; FEV: Forced expiratory volume; GS: Generation Scotland; GS:SFHS: Generation Scotland: Scottish Family Health Study; HR: Hazard ratio; ICD-10: International Classification of Diseases, 10th

Revision; IEAA: Intrinsic epigenetic age acceleration; OR: Odds ratio; SCID: Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders; SD: Standard deviation; SIMD: Scottish index of multiple deprivation; SMR: Scottish morbidity records; TL: Telomere length

Acknowledgements

The authors thank all of the participants of the Generation Scotland: Scottish Family Health Study as well as study team members for their previous and ongoing contribution to this study.

Authors' contributions

RFH, AJS and REM designed the research. RFH, AJS, DLMcC and REM contributed to the analyses. RFH and REM wrote the manuscript. Data were prepared and collected by AC, RMW, DMH, CH, AMM, DJP, IJD and KLE. All authors read, revised and approved the final manuscript.

Funding

GS received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). DNA methylation profiling of the GS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland, and was funded by the Medical Research Council UK, the Brain & Behavior Research Foundation (Ref: 27404) and the Wellcome Trust (Wellcome Trust Strategic Award 'STratifying Resilience and Depression Longitudinally' ((STRADL) Reference 104036/Z/14/Z)). AMM is supported by the Wellcome Trust (104036/Z/14/Z, 216767/Z/19/Z), UKRI MRC (MC_PC_17209, MR/S035818/1) and the European Union H2020 (SEP-210574971). RFH and AJS 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 [RFH: 108890/Z/15/Z; AJS: 203771/Z/16/Z]. DLMcC and REM are supported by Alzheimer's Research UK major project grant ARUK-PG2017B-10. SH is supported by 1U01AG060908 – 01. DMH is supported by a Sir Henry Wellcome Postdoctoral Fellowship (Reference 213674/Z/18/Z) and a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (Ref: 27404)

Availability of data and materials

According to the terms of consent for GS participants, access to data must be reviewed by the GS Access Committee. Applications should be made to access@generationscotland.org.

Ethics approval and consent to participate

All components of the Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Numbers: 05/S1401/89 and 10/S1402/20). All participants provided broad and enduring written informed consent for biomedical research. The Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 15/0040/ES). This study was performed in accordance with the Helsinki declaration.

Consent for publication

Not applicable.

Competing interests

AMM has received research support from Eli Lilly, Janssen and the Sackler Foundation. AMM has also received speaker fees from Illumina and Janssen. The other authors declare that they have no competing interests.

Author details

[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. [2]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE5 8AF, UK. [3]Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh EH16 4UX, UK. [4]Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh EH16 4UX, UK. [5]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles 90095-7088, USA. [6]Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles 90095-1772, USA. [7]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. [8]Lothian Birth Cohorts, University of Edinburgh, Edinburgh EH8 9JZ, UK.

Received: 12 May 2020 Accepted: 14 July 2020
Published online: 31 July 2020

References

1. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380(9859):2163–96.
2. 2016 GHE. Deaths by cause, age, sex, by country and by region, 2000–2016. Geneva: World Health Organization; 2018.
3. Hay SI, Abajobir AA, Abate KH, Abbafati C, Abbas KM, Abd-Allah F, et al. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1260–344.
4. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: challenges and recommendations. Genome Biol. 2019;20(1):249.
5. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018;19(6):371–84.
6. McCartney DL, Stevenson AJ, Walker RM, Gibson J, Morris SW, Campbell A, et al. Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. Alzheimers Dement (Amsterdam, Netherlands). 2018;10:429-437.
7. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. Clin Epigenetics. 2016;8(1):64.
8. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging (Albany NY). 2015;7(12):1130–42.
9. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). 2016;8(9):1844–65.
10. Han LKM, Aghajani M, Clark SL, Chan RF, Hattab MW, Shabalin AA, et al. Epigenetic aging in major depressive disorder. Am J Psychiatry. 2018;175(8):774–82.
11. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115.
12. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.
13. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging. 2018;10(4):573–91.
14. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019;11(2):303–27.
15. Ma H, Zhou Z, Wei S, Liu Z, Pooley KA, Dunning AM, et al. Shortened telomere length is associated with increased risk of cancer: a meta-analysis. PLoS One. 2011;6(6):e20466.
16. Honig LS, Kang MS, Schupf N, Lee JH, Mayeux R. Association of shorter leukocyte telomere repeat length with dementia and mortality. Arch Neurol. 2012;69(10):1332–9.
17. Wang Q, Zhan Y, Pedersen NL, Fang F, Hagg S. Telomere length and all-cause mortality: a meta-analysis. Ageing Res Rev. 2018;48:11–20.
18. Lu AT, Seeboth A, Tsai PC, Sun D, Quach A, Reiner AP, et al. DNA methylation-based estimator of telomere length. Aging (Albany NY). 2019.
19. Belsky DW, Caspi A, Houts R, Cohen HJ, Corcoran DL, Danese A, et al. Quantification of biological aging in young adults. Proc Natl Acad Sci U S A. 2015;112(30):E4104–E10.
20. Belsky DW, Caspi A, Arseneault L, Baccarelli A, Corcoran DL, Gao X, et al. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. eLife. 2020;9.
21. Stevenson AJ, McCartney DL, Hillary RF, Redmond P, Taylor AM, Zhang Q, et al. Childhood intelligence attenuates the association between biological ageing and health outcomes in later life. Translational psychiatry. 2019;9(1):323.

22. Hillary RF, Stevenson AJ, Cox SR, McCartney DL, Harris SE, Seeboth A, et al. An epigenetic predictor of death captures multi-modal measures of brain health. Molecular psychiatry. 2019.
23. Rosen AD, Robertson KD, Hlady RA, Muench C, Lee J, Philibert R, et al. DNA methylation age is accelerated in alcohol dependence. Transl Psychiatry. 2018;8(1):182.
24. Horvath S, Garagnani P, Bacalini MG, Pirazzini C, Salvioli S, Gentilini D, et al. Accelerated epigenetic aging in Down syndrome. Aging Cell. 2015;14(3): 491–5.
25. Fransquet PD, Wrigglesworth J, Woods RL, Ernst ME, Ryan J. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. Clin Epigenetics. 2019;11(1):62.
26. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015;16(1):25.
27. Ryan J, Wrigglesworth J, Loong J, Fransquet PD, Woods RL. A systematic review and meta-analysis of environmental, lifestyle and health factors associated with DNA methylation age. J Gerontol A Biol Sci Med Sci. 2019.
28. Fiorito G, McCrory C, Robinson O, Carmeli C, Rosales CO, Zhang Y, et al. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. Aging. 2019;11(7):2045–70.
29. Zhao W, Ammous F, Ratliff S, Liu J, Yu M, Mosley TH, et al. Education and lifestyle factors are associated with DNA methylation clocks in older African Americans. Int J Environ Res Public Health. 2019;16(17):3141.
30. McCrory C, Fiorito G, Hernandez B, Polidoro S, O'Halloran AM, Hever A, et al. Association of 4 epigenetic clocks with measures of functional health, cognition, and all-cause mortality in The Irish Longitudinal Study on Ageing (TILDA). bioRxiv. 2020:2020.04.27.063164.
31. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016;17(1):171.
32. Mannino DM, Davis KJ. Lung function decline and outcomes in an elderly population. Thorax. 2006;61(6):472–7.
33. Mannino DM, Buist AS, Petty TL, Enright PL, Redd SC. Lung function and mortality in the United States: data from the First National Health and Nutrition Examination Survey follow up study. Thorax. 2003;58(5):388–93.
34. Tsaprouni LG, Yang T-P, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382–96.
35. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9(5):436–47.
36. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS ONE. 2013;8(5):e63812-e.
37. Zhang Y, Breitling LP, Balavarca Y, Holleczek B, Schottker B, Brenner H. Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality. Int J Cancer. 2016;139(11):2482–92.
38. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6(1):4.
39. Philibert RA, Beach SRH, Brody GH. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. Epigenetics. 2012; 7(11):1331–8.
40. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.
41. Kodal JB, Kobylecki CJ, Vedel-Krogh S, Nordestgaard BG, Bojesen SE. AHRR hypomethylation, lung function, lung function decline and respiratory symptoms. Eur Respir J. 2018;51(3):1701512.
42. Bhattacharyya J, Mihara K, Bhattacharjee D, Mukherjee M. Telomere length as a potential biomarker of coronary artery disease. Indian J Med Res. 2017; 145(6):730–7.
43. Stefler D, Malyutina S, Maximov V, Orlov P, Ivanoschuk D, Nikitin Y, et al. Leukocyte telomere length and risk of coronary heart disease and stroke mortality: prospective evidence from a Russian cohort. Sci Rep. 2018;8(1): 16627.
44. Haycock PC, Heydon EE, Kaptoge S, Butterworth AS, Thompson A, Willeit P. Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis. BMJ : British Medical Journal. 2014; 349:g4227.
45. Oksanen T, Kivimaki M, Pentti J, Virtanen M, Klaukka T, Vahtera J. Self-report as an indicator of incident disease. Ann Epidemiol. 2010;20(7):547–54.
46. Lu AT, Xue L, Salfati EL, Chen BH, Ferrucci L, Levy D, et al. GWAS of epigenetic aging rates in blood reveals a critical role for TERT. Nat Commun. 2018;9(1):387.
47. Jylhava J, Hjelmborg J, Soerensen M, Munoz E, Tan Q, Kuja-Halkola R, et al. Longitudinal changes in the genetic and environmental influences on the epigenetic clocks across old age: evidence from two twin cohorts. EBioMedicine. 2019;40:710–6.
48. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. Int J Epidemiol. 2013;42(3):689–700.
49. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC Med Genet. 2006;7(1):74.
50. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604–12.
51. Fernandez-Pujals AM, Adams MJ, Thomson P, McKechanie AG, Blackwood DHR, Smith BH, et al. Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in Generation Scotland: Scottish Family Health Study (GS:SFHS). PLoS ONE. 2015;10(11): e0142197-e.
52. Grambsch PM, Therneau TM. Modeling survival data: extending the Cox model. Statistics for Biology and Health. 2000.
53. Therneau T. A package for survival analysis in R. version 2.42-6. 2018. Reference Source. 2019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

9.3 Conclusion

None of the six epigenetic ageing measures in this study were significantly associated with maternal or paternal history of AD. DNAm GrimAge and DunedinPoAm were nominally associated with the incidence of AD in a basic model adjusted for chronological age and sex (HR = 1.70 and 1.67, respectively, and P = 0.03). These associations did not surpass multiple testing correction and therefore were not brought forward to the fully-adjusted model. The number of incident cases for AD was low during 14 years of follow-up (n = 20 cases and 1,936 controls). Therefore, my study may have been underpowered. These relationships should be tested in samples with longer periods of follow-up or more incident cases of AD and other forms of dementia.

DNAm GrimAge outperformed other epigenetic measures of ageing in its associations with prevalent and incident disease in the GS cohort. This is in agreement with findings from Lu *et al.* in their initial description of DNAm GrimAge (313). Further, McCrory *et al.* (2021) showed that DNAm GrimAge associated with a greater number of physical and cognitive performance measures than DNAm PhenoAge, Horvath Age and Hannum Age (462).

It is unclear why DNAm GrimAge explains more of the variance in health-related characteristics than other epigenetic ageing biomarkers. Data from my study and others suggest that predicting the risk of mortality might be of greater clinical utility than estimating chronological age or telomere shortening. Further, the 'state' measure of DNAm GrimAge outperformed DunedinPoAm, which estimates the rate of biological ageing. DNAm GrimAge and DNAm PhenoAge are both based on the prediction of all-cause mortality. The inclusion of smoking-related damage in the DNAm GrimAge estimate may explain its stronger associations with health-related phenotypes when compared to DNAm PhenoAge (466). However, DNAm GrimAge associated with the incidence of chronic obstructive pulmonary disease (COPD), type 2 diabetes and ischemic heart disease independently of common risk factors including smoking. DNAm GrimAge might capture risk mechanisms in age-related diseases that are not explained by common lifestyle-related factors. Risk factors were based on self-report data. Therefore, there may have been residual confounding due to measurement error in covariates such as smoking pack years. Further work is needed to assess causality between epigenetic biomarkers, lifestyle risk factors and health outcomes.

In this chapter, I comprehensively investigated cross-sectional and longitudinal relationships between six major epigenetic measures of ageing and common diseases. My findings in Chapters 8 and 9 suggest that DNAm GrimAge, which integrates methylomic and blood proteomic data, serves as a powerful correlate of cognitive fitness as well as incident COPD, heart disease and type 2 diabetes. In the following chapter, I discuss my findings from Chapters 5-9 in the context of the wider literature, the limitations of these studies and recommendations for future work.

10 Discussion

In this thesis, I explored the integration of genomics, epigenomics and proteomics to identify molecular correlates of cognitive function and dementia. First, I carried out GWAS and EWAS on 422 plasma protein levels to probe molecular pathways that link genotype, lifestyle factors and blood proteins to AD risk (Chapters 5-7). Second, I considered a blood-based predictor of mortality risk termed DNAm GrimAge that is derived from methylation-based estimates of seven plasma protein levels and smoking behaviour. I examined associations between DNAm GrimAge and brain health and the incidence of common diseases, including AD (Chapters 8 and 9). In this chapter, I first summarise the findings from Chapters 5-9. I then critically evaluate the strengths and limitations of the samples and methodologies used in this thesis. I finish by recommending future research directions.

10.1 GWAS and EWAS on plasma protein levels

I outlined in Chapter 1 that identifying blood-based markers of neurological diseases such as dementia could allow for non-invasive and repeated monitoring of disease risk and progression. GWAS and EWAS on complex disease states reveal genomic regions that might influence disease risk mechanisms. Combining these approaches with large-scale proteomics further informs underlying disease-associated pathways (Chapter 2) (410, 467). There are over 30 GWAS with multiplexed blood protein measurements; however, only two such EWAS are reported in the literature (Chapter 3). Further, only Ahsan *et al.* have combined genomics, epigenomics and proteomics to examine the molecular architecture of cancer and CVD protein biomarkers (326). In this thesis, I conducted both GWAS and EWAS on plasma levels of neurology-related proteins (Chapter 5), peripheral inflammatory biomarkers (Chapter 6) and AD-associated proteins (Chapter 7). The studies advance our understanding of the genetic and epigenetic regulation of the plasma proteome. This is the first time that GWAS and EWAS were conducted on panels of blood proteins enriched for their relevance to neurological conditions. Using a range of molecular genetic techniques, I highlighted possible molecular pathways and causative relationships between plasma proteins and AD risk. Further, I estimated the contribution of common genetic and epigenetic variation towards inter-individual differences in plasma levels of 342 proteins (BayesR+, Chapters 6 and 7).

In Chapter 5, I performed separate GWAS and EWAS on plasma levels of 92 neurological protein biomarkers using linear regression models (Olink neurology panel). In the GWAS stage, 41 pQTLs were identified across 33 plasma proteins. Twenty-seven of these associations were not reported in the existing literature. I found 26 novel CpG associations across nine plasma proteins. Linear mixed-effects models (implemented in OSCA software) outperform linear regression methods in controlling for inflation and intercorrelations among CpG probes (132). Therefore, I also carried out EWAS using OSCA. Twenty-three of the 26 CpG associations identified by linear regression were present in OSCA. Integrating GWAS, EWAS and mQTL data suggested that a *trans* pQTL in *ITIH4* affects neprilysin (NEP) levels through altered methylation within *ITIH1* and *ITIH4*. NEP is a zinc-dependent metalloprotease that cleaves beta-amyloid in neural tissue. The loss of NEP causes AD-like behaviour in mice (468, 469). Inter-alpha-trypsin inhibitor heavy chain H1 and H4 (encoded by *ITIH1* and *ITIH4*, respectively) are up-regulated in serum samples from AD patients and murine AD models (470, 471). Whether the co-regulation of these proteins occurs in human central nervous tissue is unclear. Plasma PVR levels were causally associated with the risk of AD in MR analyses. *PVR* is located within the *TOMM40-APOE-APOC2* cluster on chromosome 19, which harbours the most significant genomic risk loci for AD (472, 473). Decreased expression of *PVR* in monocytes has been associated with AD; however, the authors showed that this association was dependent on *APOE* genotype (474). Further, PVR levels in the dorsolateral prefrontal cortex were associated with AD risk. Colocalisation analyses suggested that distinct causal variants underpinned brain PVR levels and AD risk (475). This mirrors the findings from my study that two independent causal SNPs in *PVR* associated with blood PVR levels and AD risk. The multi-omics strategy highlighted plausible networks of proteins and candidate disease markers for follow-up mechanistic studies in the context of AD pathology.

In Chapter 6, I used a novel Bayesian penalised regression framework (BayesR+) to conduct the first integrated GWAS and EWAS on blood protein levels. I performed sensitivity analyses using linear regression (GWAS and EWAS) and OSCA (EWAS alone). The analyses were applied to a panel of inflammation-associated proteins (Olink inflammation panel). Thirteen pQTLs were detected in both BayesR+ and linear regression GWAS. Three CpG associations were concordant across linear regression EWAS, OSCA and BayesR+. Two pQTL and two CpG associations were not previously reported in the literature. I used two-sample MR to test whether the 13 proteins with robust pQTL associations were causally associated with AD risk. Whereas IL18R1 showed a nominally significant relationship with AD risk (P = 0.04), no association withstood multiple testing correction. There was a limited number of instruments available to proxy for plasma protein levels in MR analyses. After LD pruning, 11/13 proteins had only one available instrument. The power of MR is affected by how much of the variance in the exposure variable instruments explain. Including additional pQTLs in allele scores that explain more variability in inflammatory protein levels can improve power (476). However, the pQTLs should directly regulate inflammatory proteins. External meta-analyses of GWAS on inflammation (and neurology-related) proteins are ongoing. The increased sample sizes afforded by meta-analyses and consortia are poised to uncover more pQTL associations. This will help to clarify causal mechanisms between peripheral proteins and neurological disease states.

I estimated the proportion of phenotypic variance in 70 inflammatory protein levels explained by genotyped SNPs (MAF > 1%) and methylation probes on the Illumina 450k array. Up to 66.4% of the variance in protein levels (for vascular endothelial growth factor A, VEGFA) was accounted for by genotype and methylation data when conditioned on one another. Methylation-based variance components estimates were strongly correlated between BayesR+

and OSCA in Chapter 6 (*r* = 0.73). BayesR+ shows a higher correlation between estimated and simulated phenotype-epigenetic probe associations with reduced mean squared errors when compared to linear regression and OSCA. The simulated data pertained to complex traits (133). The relative performances of these methods in the context of simulated proteomic data are unknown. BayesR+ is more flexible than OSCA as different prior distributions can be assigned to individual covariates (i.e. CpG probes of small, medium and large effects) or groups of covariates (i.e. genotype and methylation data). However, pre-specifying series of prior Gaussian distributions can increase computational cost over other methods. In a given study, the most appropriate method for understanding the molecular architecture of the human proteome may depend on sample size, proteome coverage, computational cost and controlling for false positives. Nevertheless, BayesR+ is a powerful and accurate method for the joint inference of genetic and epigenetic effects over phenotypes.

In Chapter 7, I performed a structured literature review to identify plasma proteins that have been associated with AD-related phenotypes (SOMAscan platform). In total, 359 proteins were identified, of which 282 were available to analyse in the Generation Scotland STRADL cohort. I showed that, in combination, genotyped SNPs and methylation data explained between 21.8% and 93.3% of the variance in SOMAmer levels. I identified 64 pQTLs across 39 proteins and 26 CpG sites that associated with 20 proteins. Seven pQTLs and 19 CpG loci were novel. There was evidence for causal associations between AD risk and plasma levels of TREM2 and TBCA (Tubulin-specific chaperone A). The role of TREM2 in AD is well described. A rare loss of function mutation in TREM2 (frequency ~ 0.2%) increases the risk of AD by three-fold (477, 478). TREM2 regulates the recruitment of microglia to amyloid plaques and limits the spread of AD pathology (479-482). Further, MS4A4A is a key regulator of TREM2 production in the CNS (448, 483). TBCA regulates the proper folding of beta-tubulin, which interacts with tau (484, 485). However, the relationship between APOE, TBCA and AD risk is unknown. My findings also suggested that the sigma-2 receptor encoded by *TMEM97* regulates MAP kinase-activated protein kinase 5 (MAPKAPK5) levels. The sigma-2 receptor promotes neuroinflammation and cognitive deficits in animal models and the uptake of toxic amyloid species in neurons *in vitro* (450, 486, 487). Future studies should investigate whether MAPKAPK5 lies on the causal path between the sigma-2 receptor and synaptotoxicity.

The novel findings in Chapters 5-7 included 36 pQTLs and 47 CpG site associations. The *NLRC5* probe cg07839457 was linked to differential plasma levels of five inflammatory proteins (CXCL9, CXCL10, CXCL11, CSF1R and IL18BP). This CpG site was associated with six other inflammatory proteins (CD163, CD48, FCGR3B, IL12, IL18 and LAG3) in the studies by Ahsan *et al.* (326) and Zaghlool *et al.* (366). *NLRC5* methylation may therefore serve as a read-out for low-grade chronic inflammation. Six proteins (CCL11, GHR, PIGR, TGF-alpha, TN-R, and WFDC2) were associated with CpG sites in *AHRR* and *F2RL3*, which are among the strongest methylation-based correlates of cigarette smoking (368, 369, 444). Further work is required to test whether these proteins mediate associations between cigarette smoking and health outcomes. I showed that pQTLs might influence plasma levels of eleven proteins via their effects on DNA methylation. There was strong evidence that pQTLs for nine proteins exert their effects at the level of transcription. Integrating genomics, epigenomics and proteomics shed light on the molecular regulation of individual proteins and their relationships with lifestyle factors, inflammation and AD risk. In the next section, I discuss the main findings from Chapters 8 and 9 in which I consider the composite biomarker DNAm GrimAge.

10.2 Epigenetic ageing and neurological outcomes

In Chapter 8, I assessed cross-sectional relationships between DNAm GrimAge and 137 phenotypes of interest in older age (mean age = 73 years). Age-adjusted DNAm GrimAge (AgeAccelGrim) associated with lower age 11 IQ (β = -0.11), age 70 IQ (β = -0.11), general cognitive ability (*'g'*, β = -0.18) and nine additional measures of crystallised and fluid intelligence.

AgeAccelGrim associated with lower white matter, grey matter and total brain volumes on MRI scans (range of β = [-0.22, -0.28]). AgeAccelGrim was positively associated with white matter hyperintensities, which predict cognitive impairment and incident dementia (488, 489). I also observed significant relationships between an accelerated DNAm GrimAge and 14 spirometric, physical and blood chemistry measures and 40 Olink neurology proteins. I tested whether early-life cognitive ability confounded associations between health- and brain-related phenotypes in older age. Twelve of the 69 associations outlined above (excluding age 11 IQ) were non-significant after adjusting for age 11 IQ. The twelve traits included six cognitive measures, triglycerides, height and four neurological protein biomarkers (CTSS, GDNF, NC-Dase and VWC2). Effect sizes for cognitive traits were attenuated, on average, by 41%. Associations between AgeAccelGrim and *'g'* and neuroimaging traits remained significant after controlling for age 11 IQ. These findings suggest that DNAm GrimAge explains additional variance in frailty and cognitive traits in older age beyond an established mortality predictor, childhood intelligence.

There was a weak association between an accelerated DNAm GrimAge at age 70 and faster cognitive decline over the eighth decade of life (P = 0.05). This association was attenuated after adjusting for age 11 IQ (P = 0.11). Maddock *et al.* (2020) showed that AgeAccelGrim cross-sectionally associated with episodic memory and mental speed at age 53 years in up to 1,368 participants from the National Survey of Health and Development Cohort. Models were adjusted for age and sex. AgeAccelGrim did not associate with cognitive decline from age 53 to age 69 years in this cohort (460). Longitudinal relationships between AgeAccelGrim and normal age-related cognitive decline should be tested with longer periods of follow-up, for instance from middle age to age 79 years or above. McCrory *et al*. (2021) reported that AgeAccelGrim associated with errors on three cognitive tasks (MMSE, Montreal Cognitive Assessment and Sustained Attention to Response Task) and slower reaction time in models adjusted for age, sex and white blood cell counts. Associations were attenuated to non-significance after controlling for body mass index, physical activity, socioeconomic status and smoking (462). I demonstrated in Chapter 8 that DNAm GrimAge and DNAm Pack Years were highly correlated (*r* = 0.84). In Chapter 9, the association between AgeAccelGrim and *'g'* was attenuated by 64% when adjusting for common lifestyle factors including smoking. Smoking behaviour likely accounts for most of the variance in cognitive traits explained by AgeAccelGrim in these studies. However, my analyses showed that effect sizes were, on average, 23% greater for the full DNAm GrimAge predictor than the smoking proxy alone. Methylation-based predictors of circulating protein levels explained additional variance in cognitive ability and frailty over the smoking predictor. Together, the findings from this study (and others) show that DNAm GrimAge, a composite biomarker of mortality risk, correlates strongly with poorer cognitive function and brain lesions.

In Chapter 9, I investigated whether DNAm GrimAge associated with the prevalence and incidence of AD and nine other common diseases. I considered five additional epigenetic measures of ageing that predict chronological age (Horvath Age and Hannum Age), all-cause mortality (DNAm PhenoAge), telomere length (DNAm Telomere Length) or the rate of biological ageing (DunedinPoAm) (314, 315, 317, 318, 323). DNAm GrimAge outperformed other epigenetic ageing measures in its associations with continuous health measures and prevalent and incident disease. This agrees with findings from Lu *et al.* (2019) and McCrory *et al.* (2021) (313, 462). There were no significant associations between epigenetic ageing biomarkers and prevalent or incident AD. Parental history of AD was used as a proxy-phenotype for prevalent cases at GS baseline. Prevalence estimates for AD are twice as high for women than men (490). Analyses were therefore conducted separately for self-reported measures of maternal and paternal history. There is a near-unit genetic correlation between self-reported measures of parental AD and the incidence of late-onset AD. This suggests that parental history is a valid proxy for AD in genetic studies (85, 491, 492). However, the prevalent cases in this study consisted of healthy individuals, which might have precluded biomarkers from detecting age- or disease-related processes that underlie AD. There were 20 incident cases of AD during 14 years of follow up, and the study may have had insufficient power to identify associations between epigenetic ageing and AD. The relationship between DNAm GrimAge and AD should be tested further in large-scale studies with rich incidence data.

AgeAccelGrim predicted time-to-onset of COPD, CVD, type 2 diabetes independently of common lifestyle risk factors. The risk factors included alcohol consumption, body mass index, deprivation, education and smoking pack years. DNAm GrimAge might serve as a read-out for causal genomic and cellular mechanisms underlying peripheral age-related diseases. However, unmeasured confounders or measurement errors in known covariates could have biased the observed associations in this study. In summary, I have performed the largest and most comprehensive studies of DNAm GrimAge in reference to its associations with cognitive fitness and common disease states. This work sets out a template for future studies to examine associations between blood-based composite biomarkers and cognitive ageing or dementia.

10.3 Limitations

I have discussed limitations specific to each study in Chapters 5-9. In the following section, I outline general limitations that apply to the cohorts and methods used in this thesis.

10.3.1 Cohorts

The empirical work in this thesis was based on data from two Scottish cohorts: the LBC1936 (Chapters 5, 6 and 8) and GS (Chapters 7 and 9). The cohorts are intensively-phenotyped with genotyping, blood DNA methylation profiling, protein measurements, brain scans, cognitive assessments and detailed health and lifestyle questionnaire data (Chapter 4). A key strength of the LBC1936 cohort is the rare phenotype of lifetime cognitive change, marking it as an important resource for cognitive epidemiology. GS is the single largest published cohort with methylation data. The cohorts were well-suited to test the hypotheses in this thesis.

The majority of health and lifestyle questionnaire data in the cohorts relied on self-report, which might be affected by recall bias (493). The cohorts are also limited by selection bias. They are oversampled for individuals with higher educational attainment and socioeconomic status than the general population. Although GS is not fully representative of the general Scottish population, it contains participants from all socioeconomic status strata (389). The samples primarily include individuals of white British ancestry. Therefore, the results in this thesis are not necessarily generalisable to other populations. A key issue in epidemiological research is the lack of diverse cohort studies. Estimates for GrimAge acceleration are higher in African Americans when compared to European Americans (Cohen's d = 0.32, 95% CI = [0.24, 0.41]) (494). However, it is unknown whether relationships between AgeAccelGrim and cognitive function and peripheral disease states are present in non-European samples. The majority of pQTL and protein-CpG studies have been conducted in individuals with European ancestry. Recently, Zhang *et al.* (2021) analysed *cis*-genetic regulation of the plasma proteome in European and African Americans. Associations detected in European Americans replicated at a lower rate in African Americans than vice versa (59.7% vs. 74.1%). Elastic net regression models were used to predict protein levels based on all *cis* pQTLs within each population. The European model performed much worse in African Americans than the converse, further underscoring the need for more diverse cohorts in molecular epidemiology (495).

10.3.2 Methodological

Proteomic assays

The Olink Neurology and Inflammatory panels were used to measure plasma protein levels in Chapters 5 and 6, respectively. The SOMAscan platform was employed in Chapter 7. There are strengths and limitations to all high-throughput proteomic assays. SomaLogic technology allows for deeper coverage of the blood proteome than Olink. However, it has been estimated that 7% of SOMAmers bind to non-target proteins (204, 496). Olink technology allows for multiplexed immunoassays (92 proteins and 96 samples) without the extensive cross-reactivity that would occur in conventional ELISAs at this scale (191). The degree to which assay-based measurements of protein levels reflect *in viv*o abundances is uncertain. Further, few studies have assessed the agreement of protein measurements across high-throughput proteomic assays. Fifty-six out of 425 proteins were highly correlated across SOMAscan and Olink platforms (*r* > 0.7, n = 48 blood samples). The median correlation coefficient was 0.36 (497). This is in line with a median estimate of 0.38 in a larger study that analysed 871 proteins in 485 individuals (351). However, median correlation coefficients of 0.62 and 0.73 have been reported in samples from high-grade serous ovarian cancer patients (498, 499). Eight out of 35 proteins measured across Olink and MS-based methods showed a correlation coefficient above 0.5 across 173 plasma samples from a Southern German population-based cohort (500). The specificities of many aptamers have been confirmed with complementary MS-based proteomics; however, not all available SOMAmers have been tested (339, 501).

The identification of *cis* pQTLs or CpG associations provides evidence for specific antibodies and aptamers. However, pQTLs might induce differential binding effects including alterations to the epitope region. Therefore, some pQTL associations might reflect differential binding of reagents to targets rather than differences in protein abundances. These pQTLs may be platform-specific. Sun *et al.* (2018) demonstrated that 371/549 *cis* pQTLs (67.6%) for SOMAscan-measured proteins were unlikely to have been influenced by differential aptamer binding. The authors tested replication of pQTL associations in 4,998 individuals using Olink data. Effect sizes were strongly correlated between pQTLs detected with SOMAscan and Olink (*r* = 0.83), and 65% of pQTLs replicated (81% *cis* and 52% *trans*) (204). Pietzner *et al.* (2021) reported a similar estimate of 64% for the proportion of pQTLs shared across SomaLogic and Olink platforms (n = 10,708). However, correlations between effect sizes were lower (*cis r* = 0.41, *trans r* = 0.34). Variants were more likely to replicate when cross-platform correlations between protein measurements and SOMAmer binding affinities were higher. Protein altering variants were significantly associated with SOMAscan-specific pQTLs, but not with associations specific to the Olink platform. This finding might reflect the reliance of SOMAmers on the shape of their protein targets (351). In this thesis, I did not account for the presence of protein altering variants. I performed functional annotation of all pQTL associations and highlighted nonsynonymous polymorphisms that might affect protein structure. However, it is possible that other pQTLs were in LD with protein altering variants. I also applied existing evidence from the literature to identify potentially non-specific pQTL associations. Additional cross-platform studies will aid in deciphering the biological relevance of pQTL and protein-CpG associations.

DNA methylation assays

Microarrays were used to quantify genome-wide methylation in this thesis. At present, the arrays only cover up to 3% of CpGs across the genome. In relation to Chapters 5-7, important CpG loci that associate with protein levels could be missed owing to this limited coverage. The proportion of distal regulatory elements interrogated by Illumina microarrays is limited (115). Increased coverage of distal regulatory sites could reveal important functional relationships between the epigenome and proteome or biological ageing. However, it is unclear whether examining additional sites in the epigenome will add substantial value to epigenetic measures of ageing (306). Zhang *et al.* (2019) showed that a near-perfect predictor of chronological age can be generated using Illumina 450k data in a large training sample (n = 13,661) (291). Building a reliable epigenetic predictor of biological ageing is more challenging given that there is no clear definition of biological age.

Statistical methods

I used BayesR+ to estimate SNP-based heritabilities of plasma protein levels. Heritability estimates based on common SNPs are generally lower than those based on twin and family studies (502, 503). Pedigree studies assess heritability based on all causal variants regardless if they are common or rare. SNP-based studies depend on LD between genotyped SNPs and causal variants. Incomplete LD between tagged and unknown causal variants induces an underestimation of narrow-sense heritability. Liu *et al.* (2015) quantified the relative contributions of heritable and environmental factors towards plasma levels of 342 proteins measured using MS in 58 pairs of monozygotic and dizygotic twins. The most strongly heritable protein was apolipoprotein A (66%). The protein most affected by common environmental factors was immunoglobulin heavy constant alpha 2 (67%) (358). Using a family-based design, Enroth *et al.* (2014) demonstrated that 75% of proteins on the Olink Oncology I panel were significantly heritable. The most heritable protein was CCL24 with an estimate of 78%. This is higher than the estimate of 67% that I obtained using SNP data and BayesR+, which implicitly controls for relatedness (324). BayesR+ controls for intercorrelations between CpG probes (133). However, the variance in protein levels explained by epigenetic variation might have been underestimated due to the limited coverage of probes on state-of-the-art microarrays.

The majority of MR analyses in this thesis relied on single instruments. Estimates of causal effects in MR analyses were obtained from a Wald ratio, which is the ratio of the SNP-outcome effect and the SNP-exposure effect (504). The result is reliable if the SNP directly influences the outcome through the exposure (vertical pleiotropy). A causal effect might instead reflect horizontal pleiotropy where the SNP affects the outcome through pathways other than, or in addition to, the exposure. The association could arise if distinct causal variants for the exposure and outcome are in LD with one another (505). Genetic colocalisation techniques including coloc can help to rule out the possibility of distinct causal variants and highlight unreliable associations in MR (220). PVR and CSF3 levels were associated with AD risk in Chapters 5 and 7, respectively. Colocalisation analyses suggested the presence of separate causal variants for plasma protein levels and disease risk. There was strong evidence for colocalisation between TREM2 or TBCA levels and AD risk (Chapter 7). As discussed in Section 2.4, the presence of multiple conditionally independent causal SNPs in a given locus could lead to an incorrect interpretation of colocalisation (274, 506, 507). Future work is required to determine whether methods that account for single or multiple causal variants are best suited to pQTL studies. In any case, vertical pleiotropy cannot be inferred in single instrument MR tests. Triangulation of evidence across other experimental and epidemiological approaches will be necessary to determine if blood TREM2 and TBCA levels influence AD risk.

I could not infer causal directionality between epigenetic ageing and health outcomes in this thesis. A recent GWAS on epigenetic ageing measures detected causal effects of adiposity and smoking status on GrimAge acceleration (508). These associations might reflect the inclusion of DNAm-based proxies for plasma leptin levels and smoking pack years in DNAm GrimAge (509). There was consistent evidence across several MR methods for a causal link between higher educational attainment and lower GrimAge acceleration. Alcohol intake was not associated with GrimAge acceleration. DNAm GrimAge acceleration was not causally associated with 23 health and disease outcomes, including heart disease and type 2 diabetes (508). In Chapter 9, associations between AgeAccelGrim and COPD, heart disease and type 2 diabetes were independent of several lifestyle factors that exhibit causal influences on GrimAge acceleration. The biological and environmental factors that underlie associations between epigenetic ageing measures and health outcomes are yet unknown.

10.4 Recommendations for future research

The first objective of this thesis was to identify molecular determinants of plasma protein levels and explore their relationships with neurological disease risk. Looking to the future, meta-analyses and increased sample samples are required to identify additional pQTL and protein-CpG associations. There is also a need for multi-ethnic GWAS and EWAS on blood protein levels, which may increase power to detect associations and highlight loci that generalise across multiple populations (510, 511). Similarly, pooling proteomic data across different cohorts and assays could determine high-confidence sets of pQTL and CpG associations. SomaLogic and Olink report relative quantification of protein levels, whereas other methods use absolute quantification. Therefore, appropriate harmonisation and standardisation methods are required to account for possible systematic biases across data sources and ensure that protein measurements are comparable. Together, these advances will help to identify ancestry- and platform-specific associations and determine why they occur.

Tissue- or cell type-specific pQTL and protein-CpG studies are required to find shared associations between blood and distal tissues including the brain and CSF. In a recent multi-tissue GWAS on protein levels, over 70% of pQTLs identified in plasma and CSF were replicated in brain tissue (parietal cortex). These results suggest that plasma is informative for identifying genetic regulators of the brain proteome (342). EPIC methylation data averaged across CpG sites correlates strongly between blood and resected brain tissue (r = 0.86) (512). However, fewer than 10% of CpG sites show significant correlations between blood and brain samples (150, 151) (Section 1.5.3). EWAS on the brain proteome are needed to determine the proportion of protein-CpG associations that are shared with blood-based studies.

There is a limited understanding of how pQTLs affect protein abundances. Evidence from my studies and others show that some pQTLs influence protein levels through their effects on DNA methylation and gene expression. The role of other processes, such as histone acetylation and chromatin remodelling, in protein regulation should be examined using multi-omics methodologies. Further, a given protein may show many ‘proteoforms’ owing to sequence variations, splicing isoforms and post-translational modifications, such as glycosylation. Characterising genetic and epigenetic factors that underpin differential proteoform levels will refine our understanding of protein regulation. For instance, Klarić *et al.* (2020) performed GWAS on immunoglobulin G glycosylation using 8,090 plasma samples. The authors demonstrated that immunoglobulin G glycosylation is influenced by variation in key transcription factors and enzymes involved in glycan synthesis (513). Wahl *et al.* (2018) found 7 CpG correlates of immunoglobulin G glycosylation, including smoking-associated probes in *AHRR* and *F2RL3* (514). An important area of future research is the development of efficient statistical methods that can integrate several omics types to study the regulation of proteins and their relationships with disease states. These methods must account for correlation structure among high-dimensional multi-omics data. Summary association results and high-dimensional molecular data pose significant data access and storage burdens. Therefore, these challenges must be overcome to permit feasible workflows in multi-omics study designs.

The second objective of this thesis was to determine whether the blood-based epigenetic ageing biomarker DNAm GrimAge associated with cognitive function and incident disease. Further studies are needed to define the factors that explain associations between DNAm GrimAge and health outcomes. Epigenetic alterations are important markers of disease even if they are non-causal and reflect passive changes in the underlying biology (306). Recently, we generated DNAm-based predictors of 953 plasma protein levels via elastic net regression (Gadd*, Hillary* *et al*., (515)). Protein levels were adjusted for known pQTL effects. The DNAm proxy for acid sphingomyelinase (ASM) predicted the incidence of AD several years before diagnosis (515). Adapting epigenetic ageing biomarkers to include proxies for proteins associated with dementia, such as ASM, might improve blood-based prediction of dementia risk. Many epigenetic ageing measures rely on bulk-cell-derived data. Further improvements might be gained by single-cell analyses, which could reveal novel biological insights into ageing-related processes (516).

Methylation-based predictors of human traits work well at the population level but are not consistently accurate at the individual level. At present, their use as biomarkers relates to their research rather than clinical applications. I highlighted these limitations in an online, interactive platform for methylation-based health profiling termed ‘MethylDetectR’ (517). Future work is warranted to assess the translational potential of methylation-based health predictors and their utility across different clinical populations. The increasing accuracy of DNAm-based predictors for traits such as chronological age has led to their use in forensics (518-520). The potential application of epigenetic biomarkers in forensic contexts and insurance risk assessments poses significant ethical and legal considerations (521-523). Frameworks are needed to enable adequate informed consent in advance of DNAm-based trait estimations and the preservation of patient or individual autonomy (306). Further research is also required to rigorously assess the validity of epigenetic ageing biomarkers in different ethnicities, disease contexts and environmental conditions.

10.5 Final summary

The increasing application of high-throughput molecular phenotyping technologies to large-scale population biobanks provides a growing resource upon which the molecular mechanisms that underlie complex disease states can be interrogated. In this thesis, an integrative, multi-omics approach was employed to examine the molecular regulation of blood proteins and their relationships with dementia risk. First, a number of statistical approaches were applied to assess the genetic and epigenetic architectures of over 400 plasma proteins. Thirty-six novel genetic variants and 47 novel epigenetic loci were associated with plasma protein abundances. Causal modelling implicated blood proteins including TREM2 and TBCA in Alzheimer’s disease risk. Second, a blood-based predictor of all-cause mortality termed DNAm GrimAge, which integrates epigenetic and proteomic data, associated with cognitive function and pulmonary and cardiometabolic disease states. However, DNAm GrimAge did not predict incident Alzheimer’s disease. The body of work in this thesis provides evidence that blood-based multi-omics approaches can be used to track disease risk mechanisms in other tissues. Further progress in these research areas will uncover perturbed molecular systems that promote dementia risk and may enable non-invasive methods that aid in the prediction of cognitive decline.

11 References

1. Organization WH. Towards a dementia plan: a WHO guide. Geneva: World Health Organization. 2018;Licence: CCBY-NC-SA 3.0 IGO.

2. Murman DL. The Impact of Age on Cognition. Semin Hear. 2015;36(3):111-21.

3. Salthouse T. Consequences of age-related cognitive declines. Annual review of psychology. 2012;63:201-26.

4. Woolgar A, Parr A, Cusack R, Thompson R, Nimmo-Smith I, Torralva T, et al. Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. Proceedings of the National Academy of Sciences. 2010;107(33):14899-902.

5. Harada CN, Natelson Love MC, Triebel KL. Normal cognitive aging. Clinics in geriatric medicine. 2013;29(4):737-52.

6. Salat DH, Kaye JA, Janowsky JS. Prefrontal gray and white matter volumes in healthy aging and Alzheimer disease. Archives of neurology. 1999;56(3):338-44.

7. O'Sullivan M, Summers PE, Jones DK, Jarosz JM, Williams SC, Markus HS. Normal-appearing white matter in ischemic leukoaraiosis: a diffusion tensor MRI study. Neurology. 2001;57(12):2307-10.

8. Madden DJ, Spaniol J, Costello MC, Bucur B, White LE, Cabeza R, et al. Cerebral white matter integrity mediates adult age differences in cognitive performance. Journal of cognitive neuroscience. 2009;21(2):289-302.

9. Terry RD, Katzman R. Life span and synapses: will there be a primary senile dementia? Neurobiology of aging. 2001;22(3):347-8; discussion 53-4.

10. Pannese E. Morphological changes in nerve cells during normal aging. Brain structure & function. 2011;216(2):85-9.

11. Terry RD, Masliah E, Salmon DP, Butters N, DeTeresa R, Hill R, et al. Physical basis of cognitive alterations in Alzheimer's disease: synapse loss is the major correlate of cognitive impairment. Annals of neurology. 1991;30(4):572-80.

12. Stern Y. What is cognitive reserve? Theory and research application of the reserve concept. Journal of the International Neuropsychological Society : JINS. 2002;8(3):448-60.

13. Richards M, Deary IJ. A life course approach to cognitive reserve: a model for cognitive aging and development? Annals of neurology. 2005;58(4):617-22.

14. Hedden T, Gabrieli JDE. Insights into the ageing mind: a view from cognitive neuroscience. Nature Reviews Neuroscience. 2004;5(2):87-96.

15. Deary IJ, Corley J, Gow AJ, Harris SE, Houlihan LM, Marioni RE, et al. Age-associated cognitive decline. British medical bulletin. 2009;92:135-52.

16. Clouston SAP, Richards M, Cadar D, Hofer SM. Educational Inequalities in Health Behaviors at Midlife: Is There a Role for Early-life Cognition? J Health Soc Behav. 2015;56(3):323-40.

17. Crichton GE, Elias MF, Davey A, Alkerwi A, Dore GA. Higher Cognitive Performance Is Prospectively Associated with Healthy Dietary Choices: The Maine Syracuse Longitudinal Study. The journal of prevention of Alzheimer's disease. 2015;2(1):24-32.

18. Deary IJ. Looking for ‘System Integrity’ in Cognitive Epidemiology. Gerontology. 2012;58(6):545-53.

19. Goldberg TE, Harvey PD, Wesnes KA, Snyder PJ, Schneider LS. Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. Alzheimers Dement (Amst). 2015;1(1):103-11.

20. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011;7(3):270-9.

21. Elahi FM, Miller BL. A clinicopathological approach to the diagnosis of dementia. Nature reviews Neurology. 2017;13(8):457-76.

22. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. Jama. 1997;278(16):1349-56.

23. Anstey KJ, Ee N, Eramudugolla R, Jagger C, Peters R. A Systematic Review of Meta-Analyses that Evaluate Risk Factors for Dementia to Evaluate the Quantity, Quality, and Global Representativeness of Evidence. J Alzheimers Dis. 2019;70(s1):S165-S86.

24. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet. 2020;396(10248):413-46.

25. Peters R, Booth A, Rockwood K, Peters J, D'Este C, Anstey KJ. Combining modifiable risk factors and risk of dementia: a systematic review and meta-analysis. BMJ open. 2019;9(1):e022846.

26. Wortmann M. World Alzheimer report 2014: dementia and risk reduction. Alzheimer's & Dementia: The Journal of the Alzheimer's Association. 2015;7(11):P837.

27. Lincoln P, Fenton K, Alessi C, Prince M, Brayne C, Wortmann M, et al. The Blackfriars Consensus on brain health and dementia. The Lancet. 2014;383(9931):1805-6.

28. Canivez GL, Watkins MW. Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. Psychological Assessment. 2010;22(4):827-36.

29. Spearman C. 'General intelligence,' objectively determined and measured. The American Journal of Psychology. 1904;15(2):201-93.

30. Horn JL, Cattell RB. Age differences in fluid and crystallized intelligence. Acta psychologica. 1967;26(2):107-29.

31. Carroll JB. Human cognitive abilities: A survey of factor-analytic studies: Cambridge University Press; 1993.

32. Vernon PE. Ability factors and environmental influences. American Psychologist. 1965;20(9):723.

33. Johnson W, Bouchard Jr TJ. The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. Intelligence. 2005;33(4):393-416.

34. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research. 1975;12(3):189-98.

35. Molloy DW, Alemayehu E, Roberts R. Reliability of a standardized mini-mental state examination compared with the traditional mini-mental state examination. Am J Psychiatry. 1991;148(1):102-5.

36. Tsoi KK, Chan JY, Hirai HW, Wong SY, Kwok TC. Cognitive Tests to Detect Dementia: A Systematic Review and Meta-analysis. JAMA internal medicine. 2015;175(9):1450-8.

37. Bateman RJ, Xiong C, Benzinger TLS, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med. 2012;367(9):795-804.

38. Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, et al. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. Annals of neurology. 2004;55(3):306-19.

39. Sperling RA, Rentz DM, Johnson KA, Karlawish J, Donohue M, Salmon DP, et al. The A4 study: stopping AD before symptoms begin? Science translational medicine. 2014;6(228):228fs13.

40. Mormino EC, Betensky RA, Hedden T, Schultz AP, Ward A, Huijbers W, et al. Amyloid and APOE ε4 interact to influence short-term decline in preclinical Alzheimer disease. Neurology. 2014;82(20):1760-7.

41. Wirth M, Madison CM, Rabinovici GD, Oh H, Landau SM, Jagust WJ. Alzheimer's disease neurodegenerative biomarkers are associated with decreased cognitive function but not β-amyloid in cognitively normal older individuals. Journal of Neuroscience. 2013;33(13):5553-63.

42. Leal SL, Lockhart SN, Maass A, Bell RK, Jagust WJ. Subthreshold amyloid predicts tau deposition in aging. Journal of Neuroscience. 2018;38(19):4482-9.

43. Wittenberg R, Knapp M, Karagiannidou M, Dickson J, Schott J. Economic impacts of introducing diagnostics for mild cognitive impairment Alzheimer's disease patients. Alzheimers Dement (N Y). 2019;5:382-7.

44. Villemagne VL, Fodero-Tavoletti MT, Masters CL, Rowe CC. Tau imaging: early progress and future directions. The Lancet Neurology. 2015;14(1):114-24.

45. Heiss W-D, Rosenberg GA, Thiel A, Berlot R, de Reuck J. Neuroimaging in vascular cognitive impairment: a state-of-the-art review. BMC Med. 2016;14(1):174.

46. Khasawneh AH, Garling RJ, Harris CA. Cerebrospinal fluid circulation: What do we know and how do we know it? Brain Circ. 2018;4(1):14-8.

47. Duits FH, Martinez-Lage P, Paquet C, Engelborghs S, Lleo A, Hausner L, et al. Performance and complications of lumbar puncture in memory clinics: results of the multicenter lumbar puncture feasibility study. Alzheimer's & Dementia. 2016;12(2):154-63.

48. Engelborghs S, Niemantsverdriet E, Struyfs H, Blennow K, Brouns R, Comabella M, et al. Consensus guidelines for lumbar puncture in patients with neurological diseases. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2017;8:111-26.

49. Bozdagi O, Rich E, Tronel S, Sadahiro M, Patterson K, Shapiro ML, et al. The neurotrophin-inducible gene Vgf regulates hippocampal function and behavior through a brain-derived neurotrophic factor-dependent mechanism. Journal of Neuroscience. 2008;28(39):9857-69.

50. Pedrero-Prieto CM, García-Carpintero S, Frontiñán-Rubio J, Llanos-González E, Aguilera García C, Alcaín FJ, et al. A comprehensive systematic review of CSF proteins and peptides that define Alzheimer’s disease. Clinical Proteomics. 2020;17(1):21.

51. de Almeida SM, Shumaker SD, LeBlanc SK, Delaney P, Marquie-Beck J, Ueland S, et al. Incidence of post-dural puncture headache in research volunteers. Headache. 2011;51(10):1503-10.

52. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

53. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science (New York, NY). 2001;291(5507):1304-51.

54. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science (New York, NY). 1998;280(5366):1077-82.

55. Antonarakis S, Krawczak M, Cooper D. The Metabolic and Molecular bases of inherited disease. Toronto: MaGraq-Hill. 2001:356-8.

56. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

57. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics. 2006;7(2):85-97.

58. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics. 2009;10(4):241-51.

59. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science (New York, NY). 1996;273(5281):1516-7.

60. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. 2009;37(13):4181-93.

61. Consortium IH. A haplotype map of the human genome. Nature. 2005;437(7063):1299.

62. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-D12.

63. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.

64. Consortium IH. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52.

65. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 2010;11(7):499-511.

66. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997-1004.

67. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2001;20(1):4-16.

68. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904-9.

69. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2008;32(4):381-5.

70. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature Reviews Genetics. 2014;15(5):335-46.

71. Davies G, Lam M, Harris SE, Trampush JW, Luciano M, Hill WD, et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. Nature Communications. 2018;9(1):2098.

72. Olesen RH, Hyde TM, Kleinman JE, Smidt K, Rungby J, Larsen A. Obesity and age-related alterations in the gene expression of zinc-transporter proteins in the human brain. Transl Psychiatry. 2016;6(6):e838.

73. Tim-Aroon T, Jinawath N, Thammachote W, Sinpitak P, Limrungsikul A, Khongkhatithum C, et al. 1q21.3 deletion involving GATAD2B: An emerging recurrent microdeletion syndrome. American journal of medical genetics Part A. 2017;173(3):766-70.

74. Sniekers S, Stringer S, Watanabe K, Jansen PR, Coleman JRI, Krapohl E, et al. Genome-wide association meta-analysis of 78,308 individuals

identifies new loci and genes influencing human intelligence. Nat Genet. 2017;49(7):1107-12.

75. Trampush JW, Yang MLZ, Yu J, Knowles E, Davies G, Liewald DC, et al. GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. Molecular Psychiatry. 2017;22(3):336-45.

76. Davies G, Marioni RE, Liewald DC, Hill WD, Hagenaars SP, Harris SE, et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112,151). Mol Psychiatry. 2016;21(6):758-67.

77. Davies G, Armstrong N, Bis JC, Bressler J, Chouraki V, Giddaluru S, et al. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53,949). Mol Psychiatry. 2015;20(2):183-92.

78. Dixon RA, DeCarlo CA, MacDonald SWS, Vergote D, Jhamandas J, Westaway D. APOE and COMT polymorphisms are complementary biomarkers of status, stability, and transitions in normal aging and early mild cognitive impairment. Front Aging Neurosci. 2014;6:236.

79. Lin CH, Lin E, Lane HY. Genetic Biomarkers on Age-Related Cognitive Decline. Frontiers in psychiatry. 2017;8:247.

80. Davies G, Harris SE, Reynolds CA, Payton A, Knight HM, Liewald DC, et al. A genome-wide association study implicates the APOE locus in nonpathological cognitive ageing. Molecular Psychiatry. 2014;19(1):76-87.

81. Raj T, Chibnik LB, McCabe C, Wong A, Replogle JM, Yu L, et al. Genetic architecture of age-related cognitive decline in African Americans. Neurology Genetics. 2017;3(1):e125.

82. Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Rongve A, et al. Largest GWAS (N=1,126,563) of Alzheimer’s Disease Implicates Microglia and Immune Cells. medRxiv. 2020:2020.11.20.20235275.

83. Bellenguez C, Küçükali F, Jansen I, Andrade V, Moreno-Grau S, Amin N, et al. New insights on the genetic etiology of Alzheimer’s and related dementia. medRxiv. 2020:2020.10.01.20200659.

84. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019.

85. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry. 2018;8(1):99.

86. Kim Y, Kong M, Lee C. Association of intronic sequence variant in the gene encoding spleen tyrosine kinase with susceptibility to vascular dementia. The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry. 2013;14(3):220-6.

87. Schrijvers EM, Schürmann B, Koudstaal PJ, van den Bussche H, Van Duijn CM, Hentschel F, et al. Genome-wide association study of vascular dementia. Stroke. 2012;43(2):315-9.

88. Chia R, Sabir MS, Bandres-Ciga S, Saez-Atienzar S, Reynolds RH, Gustavsson E, et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. Nature Genetics. 2021;53(3):294-303.

89. Ferrari R, Hernandez DG, Nalls MA, Rohrer JD, Ramasamy A, Kwok JB, et al. Frontotemporal dementia and its subtypes: a genome-wide association study. The Lancet Neurology. 2014;13(7):686-99.

90. Russo V, Martienssen R, Riggs A. Epigenetic mechanisms of gene regulation. 1996. Plainview, NY: Cold Spring Harbor Laboratory Press, xii.

91. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biology. 2014;15(5):R73.

92. Lemire M, Zaidi SH, Ban M, Ge B, Aïssi D, Germain M, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. Nat Commun. 2015;6:6326.

93. Bell JT, Spector TD. DNA methylation studies using twins: what are they telling us? Genome Biology. 2012;13(10):172.

94. McCartney DL, Stevenson AJ, Hillary RF, Walker RM, Bermingham ML, Morris SW, et al. Epigenetic signatures of starting and stopping smoking. EBioMedicine. 2018;37:214-20.

95. Niculescu MD, Zeisel SH. Diet, methyl donors and DNA methylation: interactions between dietary folate, methionine and choline. The Journal of nutrition. 2002;132(8):2333S-5S.

96. Fang M, Chen D, Yang CS. Dietary polyphenols may affect DNA methylation. The Journal of nutrition. 2007;137(1):223S-8S.

97. Rider CF, Carlsten C. Air pollution and DNA methylation: effects of exposure in humans. Clinical Epigenetics. 2019;11(1):131.

98. Ben-Hattar J, Jiricny J. Methylation of single CpG dinucleotides within a promoter element of the Herpes simplex virus tk gene reduces its transcription in vivo. Gene. 1988;65(2):219-27.

99. Watt F, Molloy PL. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes & development. 1988;2(9):1136-43.

100. Iguchi-Ariga S, Schaffner W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes & development. 1989;3(5):612-9.

101. Ferguson-Smith AC, Sasaki H, Cattanach BM, Surani MA. Parental-origin-specific epigenetic modification of the mouse H19 gene. Nature. 1993;362(6422):751-5.

102. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. Nature. 1993;366(6453):362-5.

103. Bartolomei MS, Webber AL, Brunkow ME, Tilghman SM. Epigenetic mechanisms underlying the imprinting of the mouse H19 gene. Genes & development. 1993;7(9):1663-73.

104. Lock LF, Takagi N, Martin GR. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. Cell. 1987;48(1):39-46.

105. Bird A. DNA methylation patterns and epigenetic memory. Genes & development. 2002;16(1):6-21.

106. Laird CD, Pleasant ND, Clark AD, Sneeden JL, Hassan KA, Manley NC, et al. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. Proceedings of the National Academy of Sciences. 2004;101(1):204-9.

107. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nature Reviews Genetics. 2013;14(3):204-20.

108. Li E, Zhang Y. DNA methylation in mammals. Cold Spring Harbor perspectives in biology. 2014;6(5):a019133.

109. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;69(6):915-26.

110. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247-57.

111. Adusumalli S, Mohd Omar MF, Soong R, Benoukraf T. Methodological aspects of whole-genome bisulfite sequencing analysis. Briefings in bioinformatics. 2015;16(3):369-79.

112. Barros-Silva D, Marques CJ, Henrique R, Jerónimo C. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. Genes. 2018;9(9).

113. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. Journal of biotechnology. 2017;261:105-15.

114. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nature protocols. 2011;6(4):468-81.

115. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17(1):208.

116. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(30):10604-9.

117. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biology. 2014;15(2):R31.

118. Birney E, Smith GD, Greally JM. Epigenome-wide association studies and the interpretation of disease-omics. PLoS genetics. 2016;12(6):e1006105.

119. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aïssi D, Wahl S, et al. DNA methylation and body-mass index: a genome-wide analysis. Lancet (London, England). 2014;383(9933):1990-8.

120. Agha G, Houseman EA, Kelsey KT, Eaton CB, Buka SL, Loucks EB. Adiposity is associated with DNA methylation profile in adipose tissue. International journal of epidemiology. 2015;44(4):1277-87.

121. Huang T, Zheng Y, Qi Q, Xu M, Ley SH, Li Y, et al. DNA Methylation Variants at HIF3A Locus, B-Vitamin Intake, and Long-term Weight Change: Gene-Diet Interactions in Two U.S. Cohorts. Diabetes. 2015;64(9):3146-54.

122. Pan H, Lin X, Wu Y, Chen L, Teh AL, Soh SE, et al. HIF3A association with adiposity: the story begins before birth. Epigenomics. 2015;7(6):937-50.

123. Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. Human molecular genetics. 2015;24(15):4464-79.

124. Richmond RC, Sharp GC, Ward ME, Fraser A, Lyttleton O, McArdle WL, et al. DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework. Diabetes. 2016;65(5):1231-44.

125. Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, et al. Estimation of a significance threshold for epigenome-wide association studies. Genetic epidemiology. 2018;42(1):20-33.

126. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. BMC Genomics. 2019;20(1):366.

127. van Iterson M, van Zwet EW, Heijmans BT. Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution. Genome biology. 2017;18(1):1-13.

128. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):e161.

129. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539-52.

130. Caye K, Jumentier B, François O. LFMM 2.0: Latent factor models for confounder adjustment in genome and epigenome-wide association studies. bioRxiv. 2018:255893.

131. Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nature methods. 2016;13(5):443.

132. Zhang F, Chen W, Zhu Z, Zhang Q, Nabais MF, Qi T, et al. OSCA: a tool for omic-data-based complex trait analysis. Genome Biology. 2019;20(1):107.

133. Trejo Banos D, McCartney DL, Patxot M, Anchieri L, Battram T, Christiansen C, et al. Bayesian reassessment of the epigenetic architecture of complex traits. Nature Communications. 2020;11(1):2865.

134. Zhang Y-M, Mao Y, Xie C, Smith H, Luo L, Xu S. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (Zea mays L.). Genetics. 2005;169(4):2267-75.

135. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics. 2006;38(2):203-8.

136. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709-23.

137. Wang S-B, Feng J-Y, Ren W-L, Huang B, Zhou L, Wen Y-J, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Scientific reports. 2016;6(1):1-10.

138. Li G, Zhu H. Genetic studies: the linear mixed models in genome-wide association studies. The Open Bioinformatics Journal. 2013;7(1).

139. Fernando RL, Garrick D. Bayesian methods applied to GWAS. Methods in molecular biology (Clifton, NJ). 2013;1019:237-74.

140. Kärkkäinen HP, Sillanpää MJ. Robustness of Bayesian multilocus association models to cryptic relatedness. Annals of human genetics. 2012;76(6):510-23.

141. Marioni RE, McRae AF, Bressler J, Colicino E, Hannon E, Li S, et al. Meta-analysis of epigenome-wide association studies of cognitive abilities. Molecular Psychiatry. 2018;23(11):2133-44.

142. Starnawska A, Tan Q, McGue M, Mors O, Børglum AD, Christensen K, et al. Epigenome-Wide Association Study of Cognitive Functioning in Middle-Aged Monozygotic Twins. Front Aging Neurosci. 2017;9:413.

143. Hüls A, Robins C, Conneely KN, Edgar R, De Jager PL, Bennett DA, et al. Brain DNA Methylation Patterns in CLDN5 Associated With Cognitive Decline. Biological Psychiatry. 2021.

144. Smith RG, Pishva E, Shireby G, Smith AR, Roubroeks JAY, Hannon E, et al. A meta-analysis of epigenome-wide association studies in Alzheimer’s disease highlights novel differentially methylated loci across cortex. Nature Communications. 2021;12(1):3517.

145. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nature neuroscience. 2014;17(9):1156-63.

146. Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. Nature neuroscience. 2014;17(9):1164-70.

147. Sanchez-Mut JV, Heyn H, Vidal E, Moran S, Sayols S, Delgado-Morales R, et al. Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. Transl Psychiatry. 2016;6(1):e718.

148. Roubroeks JAY, Smith AR, Smith RG, Pishva E, Ibrahim Z, Sattlecker M, et al. An epigenome-wide association study of Alzheimer's disease blood highlights robust DNA hypermethylation in the HOXB6 gene. Neurobiology of aging. 2020;95:26-45.

149. Li Y, Chen JA, Sears RL, Gao F, Klein ED, Karydas A, et al. An epigenetic signature in peripheral blood associated with the haplotype on

17q21. 31, a risk factor for neurodegenerative tauopathy. PLoS Genet. 2014;10(3):e1004211.

150. Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. Epigenetics. 2015;10(11):1024-32.

151. Walton E, Hass J, Liu J, Roffman JL, Bernardoni F, Roessner V, et al. Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. Schizophrenia bulletin. 2016;42(2):406-14.

152. Toombs J, Zetterberg H. In the blood: biomarkers for amyloid pathology and neurodegeneration in Alzheimer's disease. Brain communications. 2020;2(1):fcaa054.

153. Hye A, Lynham S, Thambisetty M, Causevic M, Campbell J, Byers HL, et al. Proteome-based plasma biomarkers for Alzheimer's disease. Brain : a journal of neurology. 2006;129(Pt 11):3042-50.

154. Güntert A, Campbell J, Saleem M, O'Brien DP, Thompson AJ, Byers HL, et al. Plasma gelsolin is decreased and correlates with rate of decline in Alzheimer's disease. J Alzheimers Dis. 2010;21(2):585-96.

155. Sattlecker M, Kiddle SJ, Newhouse S, Proitsi P, Nelson S, Williams S, et al. Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. Alzheimers Dement. 2014;10(6):724-34.

156. Greco I, Day N, Riddoch-Contreras J, Reed J, Soininen H, Kłoszewska I, et al. Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. J Transl Med. 2012;10:217.

157. Hakobyan S, Harding K, Aiyaz M, Hye A, Dobson R, Baird A, et al. Complement Biomarkers as Predictors of Disease Progression in Alzheimer's Disease. J Alzheimers Dis. 2016;54(2):707-16.

158. Thambisetty M, Simmons A, Velayudhan L, Hye A, Campbell J, Zhang Y, et al. Association of plasma clusterin concentration with severity, pathology, and progression in Alzheimer disease. Archives of general psychiatry. 2010;67(7):739-48.

159. Thambisetty M, Simmons A, Hye A, Campbell J, Westman E, Zhang Y, et al. Plasma biomarkers of brain atrophy in Alzheimer's disease. PloS one. 2011;6(12):e28527.

160. Leung R, Proitsi P, Simmons A, Lunnon K, Güntert A, Kronenberg D, et al. Inflammatory proteins in plasma are associated with severity of Alzheimer's disease. PloS one. 2013;8(6):e64971.

161. Hye A, Riddoch-Contreras J, Baird AL, Ashton NJ, Bazenet C, Leung R, et al. Plasma proteins predict conversion to dementia from prodromal disease. Alzheimers Dement. 2014;10(6):799-807.e2.

162. Velayudhan L, Killick R, Hye A, Kinsey A, Güntert A, Lynham S, et al. Plasma transthyretin as a candidate marker for Alzheimer's disease. J Alzheimers Dis. 2012;28(2):369-75.

163. Sattlecker M, Khondoker M, Proitsi P, Williams S, Soininen H, Kłoszewska I, et al. Longitudinal Protein Changes in Blood Plasma Associated with the Rate of Cognitive Decline in Alzheimer's Disease. J Alzheimers Dis. 2016;49(4):1105-14.

164. Thambisetty M, Tripaldi R, Riddoch-Contreras J, Hye A, An Y, Campbell J, et al. Proteome-based plasma markers of brain amyloid-β deposition in non-demented older individuals. J Alzheimers Dis. 2010;22(4):1099-109.

165. Ashton NJ, Kiddle SJ, Graf J, Ward M, Baird AL, Hye A, et al. Blood protein predictors of brain amyloid for enrichment in clinical trials? Alzheimers Dement (Amst). 2015;1(1):48-60.

166. Westwood S, Leoni E, Hye A, Lynham S, Khondoker MR, Ashton NJ, et al. Blood-Based Biomarker Candidates of Cerebral Amyloid Using PiB PET in Non-Demented Elderly. J Alzheimers Dis. 2016;52(2):561-72.

167. Kiddle SJ, Thambisetty M, Simmons A, Riddoch-Contreras J, Hye A, Westman E, et al. Plasma based markers of [11C] PiB-PET brain amyloid burden. PloS one. 2012;7(9):e44260.

168. Voyle N, Baker D, Burnham SC, Covin A, Zhang Z, Sangurdekar DP, et al. Blood Protein Markers of Neocortical Amyloid-β Burden: A Candidate Study Using SOMAscan Technology. J Alzheimers Dis. 2015;46(4):947-61.

169. Westwood S, Liu B, Baird AL, Anand S, Nevado-Holgado AJ, Newby D, et al. The influence of insulin resistance on cerebrospinal fluid and plasma biomarkers of Alzheimer's pathology. Alzheimers Res Ther. 2017;9(1):31.

170. Tanaka T, Lavery R, Varma V, Fantoni G, Colpo M, Thambisetty M, et al. Plasma proteomic signatures predict dementia and cognitive impairment. Alzheimers Dement (N Y). 2020;6(1):e12018.

171. Shi L, Winchester LM, Liu BY, Killick R, Ribe EM, Westwood S, et al. Dickkopf-1 Overexpression in vitro Nominates Candidate Blood Biomarkers Relating to Alzheimer’s Disease Pathology. Journal of Alzheimer's Disease. 2020;77:1353-68.

172. Shi L, Westwood S, Baird AL, Winchester L, Dobricic V, Kilpert F, et al. Discovery and validation of plasma proteomic biomarkers relating to brain amyloid burden by SOMAscan assay. Alzheimers Dement. 2019;15(11):1478-88.

173. Whelan CD, Mattsson N, Nagle MW, Vijayaraghavan S, Hyde C, Janelidze S, et al. Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer’s disease. Acta Neuropathologica Communications. 2019;7(1):169.

174. Deniz K, Ho CCG, Malphrus KG, Reddy JS, Nguyen T, Carnwath TP, et al. Plasma Biomarkers of Alzheimer's Disease in African Americans. J Alzheimers Dis. 2020.

175. Ashton NJ, Nevado-Holgado AJ, Barber IS, Lynham S, Gupta V, Chatterjee P, et al. A plasma protein classifier for predicting amyloid burden for preclinical Alzheimer’s disease. Science Advances. 2019;5(2):eaau7220.

176. Park J-C, Han S-H, Lee H, Jeong H, Byun MS, Bae J, et al. Prognostic plasma protein panel for Aβ deposition in the brain in Alzheimer's disease. Progress in Neurobiology. 2019;183:101690.

177. C. S EKE, Jammeh E, Li X, Carroll C, Pearson S, Ifeachor E. Early Detection of Alzheimer's Disease with Blood Plasma Proteins using Support Vector Machines. IEEE Journal of Biomedical and Health Informatics. 2020:1.

178. Lindbohm JV, Mars N, Walker KA, Singh-Manoux A, Livingston G, Brunner EJ, 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. 2020:2020.11.18.20234070.

179. Begic E, Hadzidedic S, Kulaglic A, Ramic-Brkic B, Begic Z, Causevic M. SOMAscan-based proteomic measurements of plasma brain natriuretic peptide are decreased in mild cognitive impairment and in Alzheimer's dementia patients. PloS one. 2019;14(2):e0212261.

180. Kim SH, Weiß C, Hoffmann U, Borggrefe M, Akin I, Behnes M. Advantages and Limitations of Current Biomarker Research: From Experimental Research to Clinical Application. Current pharmaceutical biotechnology. 2017;18(6):445-55.

181. Shi L, Baird AL, Westwood S, Hye A, Dobson R, Thambisetty M, et al. A Decade of Blood Biomarkers for Alzheimer's Disease Research: An Evolving Field, Improving Study Designs, and the Challenge of Replication. J Alzheimers Dis. 2018;62(3):1181-98.

182. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Bio/technology (Nature Publishing Company). 1996;14(1):61-5.

183. Baker MS, Ahn SB, Mohamedali A, Islam MT, Cantor D, Verhaert PD, et al. Accelerating the search for the missing proteins in the human proteome. Nature Communications. 2017;8(1):14271.

184. Schwenk JM, Omenn GS, Sun Z, Campbell DS, Baker MS, Overall CM, et al. The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. J Proteome Res. 2017;16(12):4299-310.

185. Omenn GS, Menon R, Adamski M, Blackwell T, Haab BB, Gao W, et al. The Human Plasma and Serum Proteome. In: Thongboonkerd V, editor. Proteomics of Human Body Fluids: Principles, Methods, and Applications. Totowa, NJ: Humana Press; 2007. p. 195-224.

186. Jaros JAJ, Guest PC, Bahn S, Martins-de-Souza D. Affinity Depletion of Plasma and Serum for Mass Spectrometry-Based Proteome Analysis. In: Zhou M, Veenstra T, editors. Proteomics for Biomarker Discovery. Totowa, NJ: Humana Press; 2013. p. 1-11.

187. Kaur G, Poljak A, Ali SA, Zhong L, Raftery MJ, Sachdev P. Extending the Depth of Human Plasma Proteome Coverage Using Simple Fractionation Techniques. J Proteome Res. 2021;20(2):1261-79.

188. Smith JG, Gerszten RE. Emerging Affinity-Based Proteomic Technologies for Large-Scale Plasma Profiling in Cardiovascular Disease. Circulation. 2017;135(17):1651-64.

189. Gullberg M, Gústafsdóttir SM, Schallmeiner E, Jarvius J, Bjarnegård M, Betsholtz C, et al. Cytokine detection by antibody-based proximity ligation. Proceedings of the National Academy of Sciences. 2004;101(22):8420-4.

190. Lundberg M, Eriksson A, Tran B, Assarsson E, Fredriksson S. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. Nucleic Acids Res. 2011;39(15):e102.

191. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PloS one. 2014;9(4):e95192.

192. Landegren U, Al-Amin RA, Björkesten J. A myopic perspective on the future of protein diagnostics. New Biotechnology. 2018;45:14-8.

193. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. Nature. 1990;346(6287):818-22.

194. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science (New York, NY). 1990;249(4968):505-10.

195. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, et al. Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. PloS one. 2010;5(12):e15004.

196. Damerval C, Maurice A, Josse J, De Vienne D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics. 1994;137(1):289-301.

197. Ye Y, Zhang Z, Liu Y, Diao L, Han L. A multi-omics perspective of quantitative trait loci in precision medicine. Trends in Genetics. 2020;36(5):318-36.

198. Ghilardi N, Wiestner A, Skoda RC. Thrombopoietin production is inhibited by a translational mechanism. Blood, The Journal of the American Society of Hematology. 1998;92(11):4023-30.

199. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta analysis. bioRxiv. 2018:447367.

200. Ongen H, Dermitzakis ET. Alternative Splicing QTLs in European and African Populations. Am J Hum Genet. 2015;97(4):567-75.

201. Grubert F, Zaugg Judith B, Kasowski M, Ursu O, Spacek Damek V, Martin Alicia R, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 2015;162(5):1051-65.

202. Zheng Z, Huang D, Wang J, Zhao K, Zhou Y, Guo Z, et al. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Res. 2019;48(D1):D983-D91.

203. Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics & chromatin. 2015;8:57.

204. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73-9.

205. Yao C, Chen G, Song C, Keefe J, Mendelson M, Huan T, et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nature communications. 2018;9(1):1-11.

206. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.

207. Giral H, Landmesser U, Kratzer A. Into the Wild: GWAS Exploration of Non-coding RNAs. Front Cardiovasc Med. 2018;5:181.

208. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics. 2013;45(10):1238-43.

209. Jelinek GA. Determining Causation from Observational Studies: A Challenge for Modern Neuroepidemiology. Front Neurol. 2017;8:265.

210. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. International journal of epidemiology. 2004;33(1):30-42.

211. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601.

212. Pierce BL, Ahsan H, Vanderweele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. International journal of epidemiology. 2011;40(3):740-52.

213. Staiger D, Stock JH. Instrumental variables regression with weak instruments. Econometrica: journal of the Econometric Society. 1997:557-86.

214. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International journal of epidemiology. 2015;44(2):512-25.

215. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol. 2017;32(5):377-89.

216. Schmidt AF, Dudbridge F. Mendelian randomization with Egger pleiotropy correction and weakly informative Bayesian priors. International journal of epidemiology. 2018;47(4):1217-28.

217. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018;50(5):693-8.

218. Bowden J, Del Greco M F, Minelli C, Zhao Q, Lawlor DA, Sheehan NA, et al. Improving the accuracy of two-sample summary-data Mendelian

randomization: moving beyond the NOME assumption. International journal of epidemiology. 2019;48(3):728-42.

219. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. Annals of Statistics. 2020;48(3):1742-69.

220. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. Human molecular genetics. 2018;27(R2):R195-R208.

221. Thompson JR, Minelli C, Bowden J, Del Greco FM, Gill D, Jones EM, et al. Mendelian randomization incorporating uncertainty about pleiotropy. Statistics in medicine. 2017;36(29):4627-45.

222. Davies NM, Thomas KH, Taylor AE, Taylor GM, Martin RM, Munafò MR, et al. How to compare instrumental variable and conventional regression analyses using negative controls and bias plots. International journal of epidemiology. 2017;46(6):2067-77.

223. Jackson JW, Swanson SA. Toward a clearer portrayal of confounding bias in instrumental variable applications. Epidemiology (Cambridge, Mass). 2015;26(4):498.

224. Lyall DM, Celis-Morales C, Ward J, Iliodromiti S, Anderson JJ, Gill JMR, et al. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. JAMA Cardiol. 2017;2(8):882-9.

225. Swerdlow DI, Kuchenbaecker KB, Shah S, Sofat R, Holmes MV, White J, et al. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. International journal of epidemiology. 2016;45(5):1600-16.

226. Chen J, Alt FW. Gene rearrangement and B-cell development. Current Opinion in Immunology. 1993;5(2):194-200.

227. Hoffman ES, Passoni L, Crompton T, Leu T, Schatz DG, Koff A, et al. Productive T-cell receptor beta-chain gene rearrangement: coincident regulation of cell cycle and clonality during development in vivo. Genes & development. 1996;10(8):948-62.

228. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nature reviews Genetics. 2009;10(3):184-94.

229. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010;6(4):e1000895.

230. Yang T-P, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, et al. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics. 2010;26(19):2474-6.

231. Plagnol V, Smyth DJ, Todd JA, Clayton DG. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics. 2009;10(2):327-34.

232. Wallace C, Rotival M, Cooper JD, Rice CM, Yang JH, McNeill M, et al. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Human molecular genetics. 2012;21(12):2815-24.

233. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am J Hum Genet. 2013;92(5):667-80.

234. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.

235. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nature genetics. 2017;49(4):600-5.

236. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics. 2016;48(5):481-7.

237. Wallace C. Statistical testing of shared genetic control for potentially related traits. Genetic epidemiology. 2013;37(8):802-13.

238. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLOS Genetics. 2017;13(3):e1006646.

239. Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. bioRxiv. 2021:2021.02.23.432421.

240. Giambartolomei C, Zhenli Liu J, Zhang W, Hauberg M, Shi H, Boocock J, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics. 2018;34(15):2538-45.

241. Foley CN, Staley JR, Breen PG, Sun BB, Kirk PDW, Burgess S, et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. Nature Communications. 2021;12(1):764.

242. Deng Y, Pan W. A powerful and versatile colocalization test. PLoS Comput Biol. 2020;16(4):e1007778.

243. Gleason KJ, Yang F, Pierce BL, He X, Chen LS. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. Genome Biology. 2020;21(1):236.

244. Jansen R, Hottenga JJ, Nivard MG, Abdellaoui A, Laport B, de Geus EJ, et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Human molecular genetics. 2017;26(8):1444-51.

245. Hormozdiari F, Zhu A, Kichaev G, Ju CJ, Segrè AV, Joo JWJ, et al. Widespread Allelic Heterogeneity in Complex Traits. Am J Hum Genet. 2017;100(5):789-802.

246. Deng Y, Pan W. Significance Testing for Allelic Heterogeneity. Genetics. 2018;210(1):25-32.

247. He B, Shi J, Wang X, Jiang H, Zhu H-J. Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. BMC Biology. 2020;18(1):97.

248. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nature Communications. 2018;9(1):224.

249. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nature Genetics. 2016;48(7):709-17.

250. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99(6):1245-60.

251. Gong J, Wang F, Xiao B, Panjwani N, Lin F, Keenan K, et al. Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci. PLOS Genetics. 2019;15(2):e1008007.

252. Guo C, Sieber KB, Esparza-Gordillo J, Hurle MR, Song K, Yeo AJ, et al. Identification of putative effector genes across the GWAS Catalog using molecular quantitative trait loci from 68 tissues and cell types. bioRxiv. 2019:808444.

253. King EA, Dunbar F, Davis JW, Degner JF. Estimating colocalization probability from limited summary statistics. bioRxiv. 2020:2020.05.19.104927.

254. Pividori M, Rajagopal PS, Barbeira A, Liang Y, Melia O, Bastarache L, et al. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Science Advances. 2020;6(37):eaba2083.

255. Zhu A, Matoba N, Wilson EP, Tapia AL, Li Y, Ibrahim JG, et al. MRLocus: Identifying causal genes mediating a trait through Bayesian estimation of allelic heterogeneity. PLOS Genetics. 2021;17(4):e1009455.

256. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics & chromatin. 2018;11(1):37.

257. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011;479(7371):74-9.

258. Maunakea AK, Chepelev I, Cui K, Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell research. 2013;23(11):1256-69.

259. Ziller MJ, Ortega JA, Quinlan KA, Santos DP, Gu H, Martin EJ, et al. Dissecting the functional consequences of de novo DNA methylation dynamics in human motor neuron differentiation and physiology. Cell stem cell. 2018;22(4):559-74. e9.

260. Wu H, Coskun V, Tao J, Xie W, Ge W, Yoshikawa K, et al. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. Science (New York, NY). 2010;329(5990):444-8.

261. Halpern KB, Vana T, Walker MD. Paradoxical role of DNA methylation in activation of FoxA2 gene expression during endoderm development. Journal of Biological Chemistry. 2014;289(34):23882-92.

262. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253-7.

263. Teissandier A, Bourc'his D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. The EMBO journal. 2017;36(11):1471-3.

264. Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, et al. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell. 2005;123(4):581-92.

265. Aquino E, Benton M, Haupt L, Sutherland H, Griffiths L, Lea R. Current understanding of DNA methylation and age-related disease. OBM Genetics. 2018;2(2):Article number: 016 1-17.

266. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010;6(5):e1000952.

267. Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao Y, et al. Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. Am J Hum Genet. 2018;103(5):654-65.

268. Volkov P, Olsson AH, Gillberg L, Jørgensen SW, Brøns C, Eriksson K-F, et al. A genome-wide mQTL analysis in human adipose tissue identifies genetic variants associated with DNA methylation, gene expression and metabolic traits. PloS one. 2016;11(6):e0157776.

269. Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, et al. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. Nature neuroscience. 2016;19(1):48-54.

270. Quon G, Lippert C, Heckerman D, Listgarten J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. Nucleic Acids Res. 2013;41(4):2095-104.

271. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC genomics. 2014;15(1):1-11.

272. van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC genomics. 2012;13(1):1-13.

273. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016;167(5):1398-414.e24.

274. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61.

275. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. medRxiv. 2020:2020.09.01.20180406.

276. Tachmazidou I, Süveges D, Min JL, Ritchie GR, Steinberg J, Walter K, et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. The American Journal of Human Genetics. 2017;100(6):865-84.

277. Kato N, Loh M, Takeuchi F, Verweij N, Wang X, Zhang W, et al. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. Nature genetics. 2015;47(11):1282-93.

278. Battram T, Richmond RC, Baglietto L, Haycock PC, Perduca V, Bojesen SE, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. International journal of epidemiology. 2019;48(5):1493-504.

279. Richardson TG, Zheng J, Davey Smith G, Timpson NJ, Gaunt TR, Relton CL, et al. Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk. Am J Hum Genet. 2017;101(4):590-602.

280. Jamieson E, Korologou-Linden R, Wootton RE, Guyatt AL, Battram T, Burrows K, et al. Smoking, DNA Methylation, and Lung Function: a Mendelian Randomization Analysis to Investigate Causal Pathways. The American Journal of Human Genetics. 2020;106(3):315-26.

281. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. Genome Biology. 2016;17(1):176.

282. Hannon E, Dempster EL, Mansell G, Burrage J, Bass N, Bohlken MM, et al. DNA methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia. eLife. 2021;10.

283. Pierce BL, Tong L, Argos M, Demanelis K, Jasmine F, Rakibuz-Zaman M, et al. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. Nature Communications. 2018;9(1):804.

284. Dolinoy DC, Weidman JR, Jirtle RL. Epigenetic gene regulation: linking early developmental environment to adult disease. Reproductive toxicology. 2007;23(3):297-307.

285. Byun H-M, Nordio F, Coull BA, Tarantini L, Hou L, Bonzini M, et al. Temporal stability of epigenetic markers: sequence characteristics and predictors of short-term DNA methylation variations. PloS one. 2012;7(6):e39220.

286. Talens RP, Boomsma DI, Tobi EW, Kremer D, Jukema JW, Willemsen G, et al. Variation, patterns, and temporal stability of DNA methylation:

considerations for epigenetic epidemiology. The FASEB Journal. 2010;24(9):3135-44.

287. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome biology. 2018;19(1):1-11.

288. Hamilton OK, Zhang Q, McRae AF, Walker RM, Morris SW, Redmond P, et al. An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. International Journal of Obesity. 2019;43(9):1795-802.

289. Barbu MC, Shen X, Walker RM, Howard DM, Evans KL, Whalley HC, et al. Epigenetic prediction of major depressive disorder. Molecular Psychiatry. 2020:1-12.

290. Langdon RJ, Beynon RA, Ingarfield K, Marioni RE, McCartney DL, Martin RM, et al. Epigenetic prediction of complex traits and mortality in a cohort of individuals with oropharyngeal cancer. Clinical epigenetics. 2020;12:1-14.

291. Zhang Q, Vallerga CL, Walker RM, Lin T, Henders AK, Montgomery GW, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. Genome Medicine. 2019;11(1):54.

292. Zhang Y, Elgizouli M, Schöttker B, Holleczek B, Nieters A, Brenner H. Smoking-associated DNA methylation markers predict lung cancer incidence. Clinical epigenetics. 2016;8(1):1-12.

293. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33(1):1.

294. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

295. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67.

296. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301-20.

297. Gabay C, Kushner I. Acute-phase proteins and other systemic responses to inflammation. New England journal of medicine. 1999;340(6):448-54.

298. Sproston NR, Ashworth JJ. Role of C-reactive protein at sites of inflammation and infection. Frontiers in immunology. 2018;9:754.

299. Koenig W, Sund M, Fröhlich M, Löwel H, Hutchinson WL, Pepys MB. Refinement of the association of serum C-reactive protein concentration and coronary heart disease risk by correction for within-subject variation over time: the MONICA Augsburg studies, 1984 and 1987. American journal of epidemiology. 2003;158(4):357-64.

300. Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, et al. Characterisation of an inflammation-related epigenetic

score and its association with cognitive ability. Clinical epigenetics. 2020;12(1):1-11.

301. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biology. 2016;17(1):255.

302. Barker ED, Cecil CA, Walton E, Houtepen LC, O'Connor TG, Danese A, et al. Inflammation-related epigenetic risk and child and adolescent mental health: A prospective study from pregnancy to middle adolescence. Development and psychopathology. 2018;30(3):1145-56.

303. Green C, Shen X, Stevenson AJ, Conole ELS, Harris MA, Barbu MC, et al. Structural brain correlates of serum and epigenetic markers of inflammation in major depressive disorder. Brain, behavior, and immunity. 2021;92:39-48.

304. Stevenson AJ, Gadd DA, Hillary RF, McCartney DL, Campbell A, Walker RM, et al. Creating and validating a DNA methylation-based proxy for interleukin-6. The journals of gerontology Series A, Biological sciences and medical sciences. 2021.

305. Guerreiro R, Bras J. The age factor in Alzheimer's disease. Genome medicine. 2015;7:106.

306. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: challenges and recommendations. Genome biology. 2019;20(1):1-24.

307. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics. 2018;19(6):371-84.

308. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome biology. 2015;16(1):1-12.

309. McCartney DL, Stevenson AJ, Walker RM, Gibson J, Morris SW, Campbell A, et al. Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2018;10:429-37.

310. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. Clinical epigenetics. 2016;8(1):1-7.

311. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging (Albany NY). 2015;7(12):1130.

312. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai P-C, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). 2016;8(9):1844.

313. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019;11(2):303-27.

314. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115.

315. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell. 2013;49(2):359-67.

316. Klemera P, Doubal S. A new approach to the concept and computation of biological age. Mechanisms of ageing and development. 2006;127(3):240-8.

317. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging. 2018;10(4):573-91.

318. Lu AT, Seeboth A, Tsai P-C, Sun D, Quach A, Reiner AP, et al. DNA methylation-based estimator of telomere length. Aging. 2019;11(16):5895-923.

319. Brouilette S, Singh RK, Thompson JR, Goodall AH, Samani NJ. White cell telomere length and risk of premature myocardial infarction. Arteriosclerosis, thrombosis, and vascular biology. 2003;23(5):842-6.

320. Valdes AM, Andrew T, Gardner JP, Kimura M, Oelsner E, Cherkas LF, et al. Obesity, cigarette smoking, and telomere length in women. Lancet (London, England). 2005;366(9486):662-4.

321. Takubo K, Nakamura K, Izumiyama N, Furugori E, Sawabe M, Arai T, et al. Telomere shortening with aging in human liver. The journals of gerontology Series A, Biological sciences and medical sciences. 2000;55(11):B533-6.

322. Belsky DW, Caspi A, Houts R, Cohen HJ, Corcoran DL, Danese A, et al. Quantification of biological aging in young adults. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(30):E4104-E10.

323. Belsky DW, Caspi A, Arseneault L, Baccarelli A, Corcoran DL, Gao X, et al. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. eLife. 2020;9.

324. Enroth S, Johansson Å, Enroth SB, Gyllensten U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. Nature communications. 2014;5(1):1-11.

325. Deming Y, Xia J, Cai Y, Lord J, Del-Aguila JL, Fernandez MV, et al. Genetic studies of plasma analytes identify novel potential biomarkers for several complex traits. Scientific Reports. 2016;6(1):1-17.

326. Ahsan M, Ek WE, Rask-Andersen M, Karlsson T, Lind-Thomsen A, Enroth S, et al. The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. PLoS genetics. 2017;13(9):e1007005.

327. Bretherick AD, Canela-Xandri O, Joshi PK, Clark DW, Rawlik K, Boutin TS, et al. Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. PLoS genetics. 2020;16(7):e1008785.

328. Nath AP, Ritchie SC, Grinberg NF, Tang HH-F, Huang QQ, Teo SM, et al. Multivariate genome-wide association analysis of a cytokine network reveals variants with widespread immune, haematological, and cardiometabolic pleiotropy. The American Journal of Human Genetics. 2019;105(6):1076-90.

329. Zhernakova DV, Le TH, Kurilshikov A, Atanasovska B, Bonder MJ, Sanna S, et al. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. Nature genetics. 2018;50(11):1524-32.

330. Gudjonsson A, Gudmundsdottir V, Axelsson GT, Gudmundsson EF, Jonsson BG, Launer LJ, et al. A genome-wide association study of serum proteins reveals shared loci with common diseases. bioRxiv. 2021:2021.07.02.450858.

331. Lourdusamy A, Newhouse S, Lunnon K, Proitsi P, Powell J, Hodges A, et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Human molecular genetics. 2012;21(16):3719-26.

332. Folkersen L, Gustafsson S, Wang Q, Hansen DH, Hedman ÅK, Schork A, et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nature metabolism. 2020;2(10):1135-48.

333. Sliz E, Kalaoja M, Ahola-Olli A, Raitakari O, Perola M, Salomaa V, et al. Genome-wide association study identifies seven novel loci associating with circulating cytokines and cell adhesion molecules in Finns. Journal of medical genetics. 2019;56(9):607-16.

334. Emilsson V, Gudmundsdottir V, Gudjonsson A, Karim MA, Ilkov M, Staley JR, et al. Coding and regulatory variants affect serum protein levels and common disease. bioRxiv. 2021:2020.05.06.080440.

335. Di Narzo AF, Telesco SE, Brodmerkel C, Argmann C, Peters LA, Li K, et al. High-throughput characterization of blood serum proteomics of IBD patients with respect to aging and genetic factors. PLoS genetics. 2017;13(1):e1006565.

336. Carayol J, Chabert C, Di Cara A, Armenise C, Lefebvre G, Langin D, et al. Protein quantitative trait locus study in obesity during weight-loss identifies a leptin regulator. Nature communications. 2017;8(1):1-14.

337. Benson MD, Yang Q, Ngo D, Zhu Y, Shen D, Farrell LA, et al. Genetic architecture of the cardiovascular risk proteome. Circulation. 2018;137(11):1158-72.

338. Gurinovich A, Song Z, Zhang W, Federico A, Monti S, Andersen SL, et al. Effect of longevity genetic variants on the molecular aging rate. GeroScience. 2021.

339. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science (New York, NY). 2018;361(6404):769-73.

340. Gao P, Ye L, Cheng H, Li H. The Mechanistic Role of Bridging Integrator 1 (BIN1) in Alzheimer’s Disease. Cellular and Molecular Neurobiology. 2020.

341. Qin Q, Teng Z, Liu C, Li Q, Yin Y, Tang Y. TREM2, microglia, and Alzheimer's disease. Mechanisms of ageing and development. 2021;195:111438.

342. Yang C, Farias FHG, Ibanez L, Suhy A, Sadler B, Fernandez MV, et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. Nature neuroscience. 2021.

343. Pietzner M, Wheeler E, Carrasco-Zanini J, Raffler J, Kerrison ND, Oerton E, et al. Genetic architecture of host proteins involved in SARS-CoV-2 infection. Nat Commun. 2020;11(1):6397.

344. Zhang J, Dutta D, Köttgen A, Tin A, Schlosser P, Grams ME, et al. Large Bi-Ethnic Study of Plasma Proteome Leads to Comprehensive Mapping of cis-pQTL and Models for Proteome-wide Association Studies. bioRxiv. 2021:2021.03.15.435533.

345. Folkersen L, Fauman E, Sabater-Lleal M, Strawbridge RJ, Frånberg M, Sennblad B, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS genetics. 2017;13(4):e1006706.

346. Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, et al. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. Scientific reports. 2019;9(1):1-14.

347. Gilly A, Park Y-C, Png G, Barysenka A, Fischer I, Bjørnland T, et al. Whole-genome sequencing analysis of the cardiometabolic proteome. Nature communications. 2020;11(1):1-9.

348. Viñuela A, Brown AA, Fernandez J, Hong M-g, Brorsson CA, Koivula RW, et al. Genetic analysis of blood molecular phenotypes reveals regulatory networks affecting complex traits: a DIRECT study. medRxiv. 2021:2021.03.26.21254347.

349. Zhong W, Gummesson A, Tebani A, Karlsson MJ, Hong M-G, Schwenk JM, et al. Whole-genome sequence association analysis of blood proteins in a longitudinal wellness cohort. Genome Medicine. 2020;12(1):1-16.

350. Zhong W, Edfors F, Gummesson A, Bergström G, Fagerberg L, Uhlén M. Next generation plasma proteome profiling to monitor health and disease. Nature communications. 2021;12(1):1-12.

351. Pietzner M, Wheeler E, Carrasco-Zanini J, Kerrison ND, Oerton E, Koprulu M, et al. Cross-platform proteomics to advance genetic prioritisation strategies. bioRxiv. 2021:2021.03.18.435919.

352. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Statistical Science. 2009;24(4):451-71.

353. Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. Nature Reviews Genetics. 2013;14(1):1-2.

354. Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3: Genes, Genomes, Genetics. 2014;4(6):1027-46.

355. Heckerman D, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, et al. Linear mixed model for heritability estimation that explicitly addresses

environmental variation. Proceedings of the National Academy of Sciences. 2016;113(27):7377-82.

356. Lynch M, Walsh B. Genetics and analysis of quantitative traits: Sinauer Sunderland, MA; 1998.

357. Solomon T, Lapek Jr JD, Jensen SB, Greenwald WW, Hindberg K, Matsui H, et al. Identification of common and rare genetic variation associated with plasma protein levels using whole-exome sequencing and mass spectrometry. Circulation: Genomic and Precision Medicine. 2018;11(12):e002170.

358. Liu Y, Buil A, Collins BC, Gillet LC, Blum LC, Cheng LY, et al. Quantitative variability of 342 plasma proteins in a human twin population. Molecular systems biology. 2015;11(2):786.

359. Johansson Å, Enroth S, Palmblad M, Deelder AM, Bergquist J, Gyllensten U. Identification of genetic variants influencing the human plasma proteome. Proceedings of the National Academy of Sciences. 2013;110(12):4673-8.

360. Ruffieux H, Carayol J, Popescu R, Harper M-E, Dent R, Saris WH, et al. A fully joint Bayesian quantitative trait locus mapping of human protein abundance in plasma. PLoS Comput Biol. 2020;16(6):e1007882.

361. Melzer D, Perry JR, Hernandez D, Corsi A-M, Stevens K, Rafferty I, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008;4(5):e1000072.

362. Kim S, Swaminathan S, Inlow M, Risacher SL, Nho K, Shen L, et al. Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. PloS one. 2013;8(7):e70269.

363. Sun W, Kechris K, Jacobson S, Drummond MB, Hawkins GA, Yang J, et al. Common genetic polymorphisms influence blood biomarker measurements in COPD. PLoS genetics. 2016;12(8):e1006011.

364. Ahola-Olli AV, Würtz P, Havulinna AS, Aalto K, Pitkänen N, Lehtimäki T, et al. Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. The American Journal of Human Genetics. 2017;100(1):40-50.

365. de Vries PS, Yu B, Feofanova EV, Metcalf GA, Brown MR, Zeighami AL, et al. Whole-genome sequencing study of serum peptide levels: the Atherosclerosis Risk in Communities study. Human molecular genetics. 2017;26(17):3442-50.

366. Zaghlool SB, Kühnel B, Elhadad MA, Kader S, Halama A, Thareja G, et al. Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. Nature communications. 2020;11(1):1-12.

367. Davis BK, Roberts RA, Huang MT, Willingham SB, Conti BJ, Brickey WJ, et al. Cutting edge: NLRC5-dependent activation of the inflammasome. The Journal of Immunology. 2011;186(3):1333-7.

368. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. Circulation: cardiovascular genetics. 2016;9(5):436-47.

369. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS one. 2013;8(5):e63812.

370. Yang R, Wu GW, Verhoeven JE, Gautam A, Reus VI, Kang JI, et al. A DNA methylation clock associated with age-related illnesses and mortality is accelerated in men with combat PTSD. Molecular psychiatry. 2020:1-11.

371. Kuan P-F, Ren X, Clouston S, Yang X, Jonas K, Kotov R, et al. PTSD is associated with accelerated transcriptional aging in World Trade Center responders. Transl Psychiatry. 2021;11(1):311.

372. Katrinli S, Stevens J, Wani AH, Lori A, Kilaru V, van Rooij SJ, et al. Evaluating the impact of trauma and PTSD on epigenetic prediction of lifespan and neural integrity. Neuropsychopharmacology. 2020;45(10):1609-16.

373. McLachlan KJ, Cole JH, Harris SE, Marioni RE, Deary IJ, Gale CR. Attitudes to ageing, biomarkers of ageing and mortality: the Lothian Birth Cohort 1936. J Epidemiol Community Health. 2020;74(4):377-83.

374. Protsenko E, Yang R, Nier B, Reus V, Hammamieh R, Rampersaud R, et al. “GrimAge,” an epigenetic predictor of mortality, is accelerated in major depressive disorder. Transl Psychiatry. 2021;11(1):1-9.

375. Higgins-Chen AT, Boks MP, Vinkers CH, Kahn RS, Levine ME. Schizophrenia and Epigenetic Aging Biomarkers: Increased Mortality, Reduced Cancer Risk, and Unique Clozapine Effects. Biol Psychiatry. 2020;88(3):224-35.

376. Dugué P-A, Bassett JK, Wong EM, Joo JE, Li S, Yu C, et al. Biological Aging Measures Based on Blood DNA Methylation and Risk of Cancer: A Prospective Study. JNCI Cancer Spectr. 2020;5(1):pkaa109-pkaa.

377. Gào X, Zhang Y, Boakye D, Li X, Chang-Claude J, Hoffmeister M, et al. Whole blood DNA methylation aging markers predict colorectal cancer survival: a prospective cohort study. Clinical epigenetics. 2020;12(1):1-13.

378. Beynon RA, Ingle SM, Langdon R, May M, Ness A, Martin R, et al. Epigenetic biomarkers of ageing are predictive of mortality risk in a longitudinal clinical cohort of individuals diagnosed with oropharyngeal cancer. medRxiv. 2020:2020.02.04.20020198.

379. Ammous F, Zhao W, Ratliff SM, Mosley TH, Bielak LF, Zhou X, et al. Epigenetic age acceleration is associated with cardiometabolic risk factors and clinical cardiovascular disease risk scores in African Americans. Clinical epigenetics. 2021;13(1):1-13.

380. Lu AT, Narayan P, Grant MJ, Langfelder P, Wang N, Kwak S, et al. DNA methylation study of Huntington’s disease and motor progression in patients and in animal models. Nature communications. 2020;11(1):1-15.

381. Deary IJ, Gow AJ, Taylor MD, Corley J, Brett C, Wilson V, et al. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. BMC Geriatr. 2007;7:28.

382. Taylor AM, Pattie A, Deary IJ. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. International journal of epidemiology. 2018;47(4):1042-r.

383. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936. International journal of epidemiology. 2011;41(6):1576-84.

384. Houlihan LM, Davies G, Tenesa A, Harris SE, Luciano M, Gow AJ, et al. Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. Am J Hum Genet. 2010;86(4):626-31.

385. Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. Genome research. 2014;24(11):1725-33.

386. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363-9.

387. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC Medical Genetics. 2006;7(1):74.

388. Navrady LB, Wolters MK, MacIntyre DJ, Clarke T-K, Campbell AI, Murray AD, et al. Cohort Profile: Stratifying Resilience and Depression Longitudinally (STRADL): a questionnaire follow-up of Generation Scotland: Scottish Family Health Study (GS:SFHS). International journal of epidemiology. 2017;47(1):13-4g.

389. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. International journal of epidemiology. 2013;42(3):689-700.

390. Nagy R, Boutin TS, Marten J, Huffman JE, Kerr SM, Campbell A, et al. Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. Genome medicine. 2017;9(1):23.

391. Purcell S, Chang C. PLINK 1.9 package. 2016.

392. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742-015-0047-8.

393. Consortium GP. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061.

394. Amador C, Huffman J, Trochet H, Campbell A, Porteous D, Wilson JF, et al. Recent genomic heritage in Scotland. BMC Genomics. 2015;16(1):437.

395. Clarke T-K, Adams MJ, Howard DM, Xia C, Davies G, Hayward C, et al. Genetic and shared couple environmental contributions to smoking and alcohol use in the UK population. Molecular Psychiatry. 2019.

396. Howard DM, Hall LS, Hafferty JD, Zeng Y, Adams MJ, Clarke T-K, et al. Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank. Transl Psychiatry. 2017;7(11):1263.

397. Fortin J-P, Fertig E, Hansen K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Research. 2014;3.

398. Pidsley R, Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC genomics. 2013;14(1):1-10.

399. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genomics data. 2016;9:22-4.

400. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45(4):e22.

401. Walker RM, Vaher K, Bermingham ML, Morris SW, Bretherick AD, Zeng Y, et al. Identification of epigenome-wide DNA methylation differences between carriers of APOE ε4 and APOE ε2 alleles. Genome Medicine. 2021;13(1):1-14.

402. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018;34(23):3983-9.

403. Candia J, Cheung F, Kotliarov Y, Fantoni G, Sellers B, Griesman T, et al. Assessment of Variability in the SOMAscan Assay. Scientific Reports. 2017;7(1):14248.

404. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annual review of genomics and human genetics. 2009;10:387-406.

405. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology. 2010;34(8):816-34.

406. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

407. Therneau TM. coxme: Mixed Effects Cox Models. R package version 2.2-16. 2020.

408. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. 2018;7.

409. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013;41(7):e90.

410. Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. Nature Reviews Genetics. 2021;22(1):19-37.

411. Hillary RF, McCartney DL, Harris SE, Stevenson AJ, Seeboth A, Zhang Q, et al. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. Nature Communications. 2019;10(1):3160.

412. Dionisio-Santos DA, Olschowka JA, O'Banion MK. Exploiting microglial and peripheral immune cell crosstalk to treat Alzheimer's disease. Journal of Neuroinflammation. 2019;16(1):74.

413. Walker KA, Gottesman RF, Wu A, Knopman DS, Gross AL, Mosley TH, et al. Systemic inflammation during midlife and cognitive change over 20 years: The ARIC Study. Neurology. 2019;92(11):e1256-e67.

414. Walker KA, Hoogeveen RC, Folsom AR, Ballantyne CM, Knopman DS, Windham BG, et al. Midlife systemic inflammatory markers are associated with late-life brain volume: the ARIC study. Neurology. 2017;89(22):2262-70.

415. Lai KSP, Liu CS, Rau A, Lanctôt KL, Köhler CA, Pakosh M, et al. Peripheral inflammatory markers in Alzheimer's disease: a systematic review and meta-analysis of 175 studies. Journal of Neurology, Neurosurgery & Psychiatry. 2017;88(10):876-82.

416. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults. Genome Medicine. 2020;12(1):60.

417. 2020 GHE. Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. Geneva: World Health Organization. 2020.

418. DALYs GBD, Collaborators H. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet (London, England). 2017;390(10100):1260-344.

419. Nakamura A, Kaneko N, Villemagne VL, Kato T, Doecke J, Doré V, et al. High performance plasma amyloid-β biomarkers for Alzheimer's disease. Nature. 2018;554(7691):249-54.

420. Kiddle SJ, Voyle N, Dobson RJB. A Blood Test for Alzheimer's Disease: Progress, Challenges, and Recommendations. J Alzheimers Dis. 2018;64(s1):S289-S97.

421. Mielke MM, Hagen CE, Xu J, Chai X, Vemuri P, Lowe VJ, et al. Plasma phospho-tau181 increases with Alzheimer's disease clinical severity and is associated with tau- and amyloid-positron emission tomography. Alzheimers Dement. 2018;14(8):989-97.

422. Palmqvist S, Insel PS, Stomrud E, Janelidze S, Zetterberg H, Brix B, et al. Cerebrospinal fluid and plasma biomarker trajectories with increasing amyloid deposition in Alzheimer's disease. EMBO molecular medicine. 2019;11(12):e11170.

423. Olsson B, Lautner R, Andreasson U, Öhrfelt A, Portelius E, Bjerke M, et al. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. The Lancet Neurology. 2016;15(7):673-84.

424. Lista S, Faltraco F, Prvulovic D, Hampel H. Blood and plasma-based proteomic biomarker research in Alzheimer's disease. Prog Neurobiol. 2013;101-102:1-17.

425. Kibinge NK, Relton CL, Gaunt TR, Richardson TG. Characterizing the Causal Pathway for Genetic Variants Associated with Neurological Phenotypes Using Human Brain-Derived Proteome Data. The American Journal of Human Genetics. 2020;106(6):885-92.

426. Kiddle SJ, Steves CJ, Mehta M, Simmons A, Xu X, Newhouse S, et al. Plasma protein biomarkers of Alzheimer's disease endophenotypes in asymptomatic older twins: early cognitive decline and regional brain volumes. Transl Psychiatry. 2015;5(6):e584.

427. Begic E, Hadzidedic S, Obradovic S, Begic Z, Causevic M. Increased Levels of Coagulation Factor XI in Plasma Are Related to Alzheimer's Disease Diagnosis. J Alzheimers Dis. 2020;77(1):375-86.

428. Shi L, Winchester LM, Westwood S, Baird AL, Anand SN, Buckley NJ, et al. Replication study of plasma proteins relating to Alzheimer's pathology. Alzheimers Dement. 2021.

429. Walker KA, Chen J, Zhang J, Fornage M, Yang Y, Zhou L, et al. Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. Nature Aging. 2021;1(5):473-89.

430. Kiddle SJ, Sattlecker M, Proitsi P, Simmons A, Westman E, Bazenet C, et al. Candidate blood proteome markers of Alzheimer's disease onset and progression: a systematic review and replication study. J Alzheimers Dis. 2014;38(3):515-31.

431. Libby G, Smith A, McEwan NF, Chien PF, Greene SA, Forsyth JS, et al. The Walker Project: a longitudinal study of 48,000 children born 1952-1966 (aged 36-50 years in 2002) and their families. Paediatric and perinatal epidemiology. 2004;18(4):302-12.

432. Batty GD, Morton SM, Campbell D, Clark H, Smith GD, Hall M, et al. The Aberdeen Children of the 1950s cohort study: background, methods and follow-up information on a new resource for the study of life course and intergenerational influences on health. Paediatric and perinatal epidemiology. 2004;18(3):221-39.

433. Habota T, Sandu A, Waiter G, McNeil C, Steele J, Macfarlane J, et al. Cohort profile for the STratifying Resilience and Depression Longitudinally (STRADL) study: A depression-focused investigation of Generation Scotland, using detailed clinical, cognitive, and neuroimaging assessments [version 1; peer review: 1 approved, 1 not approved]. Wellcome Open Research. 2019;4(185).

434. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics. 2012;13:86.

435. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics. 2016;32(20):3207-9.

436. Deming Y, Li Z, Kapoor M, Harari O, Del-Aguila JL, Black K, et al. Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers. Acta Neuropathol. 2017;133(5):839-56.

437. Hong S, Prokopenko D, Dobricic V, Kilpert F, Bos I, Vos SJB, et al. Genome-wide association study of Alzheimer’s disease CSF biomarkers in the EMIF-AD Multimodal Biomarker Discovery dataset. Transl Psychiatry. 2020;10(1):403.

438. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17(1):122.

439. Innocenti F, Jiang C, Sibley AB, Etheridge AS, Hatch AJ, Denning S, et al. Genetic variation determines VEGF-A plasma levels in cancer patients. Scientific Reports. 2018;8(1):16332.

440. Ho JE, Mahajan A, Chen M-H, Larson MG, McCabe EL, Ghorbani A, et al. Clinical and genetic correlates of growth differentiation factor 15 in the community. Clin Chem. 2012;58(11):1582-91.

441. Jun G, Ibrahim-Verbaas CA, Vronskaya M, Lambert JC, Chung J, Naj AC, et al. A novel Alzheimer disease locus located near the gene encoding tau protein. Mol Psychiatry. 2016;21(1):108-17.

442. Cruchaga C, Kauwe JS, Harari O, Jin SC, Cai Y, Karch CM, et al. GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. Neuron. 2013;78(2):256-68.

443. Nazarian A, Yashin AI, Kulminski AM. Genome-wide analysis of genetic predisposition to Alzheimer’s disease and related sex disparities. Alzheimers Res Ther. 2019;11(1):5.

444. Philibert RA, Beach SR, Brody GH. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. Epigenetics. 2012;7(11):1331-8.

445. Kauwe JS, Bailey MH, Ridge PG, Perry R, Wadsworth ME, Hoyt KL, et al. Genome-wide association study of CSF levels of 59 alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. PLoS Genet. 2014;10(10):e1004758.

446. Liu C, Yu J. Genome-Wide Association Studies for Cerebrospinal Fluid Soluble TREM2 in Alzheimer's Disease. Front Aging Neurosci. 2019;11:297.

447. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, et al. Large-scale integration of the plasma proteome with genetics and disease. Nature Genetics. 2021.

448. Deming Y, Filipello F, Cignarella F, Cantoni C, Hsu S, Mikesell R, et al. The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. Science translational medicine. 2019;11(505).

449. Walker RM, Bermingham ML, Vaher K, Morris SW, Clarke T-K, Bretherick AD, et al. Epigenome-wide analyses identify DNA methylation signatures of dementia risk. Alzheimers Dement (Amst). 2020;12(1):e12078.

450. Riad A, Lengyel-Zhand Z, Zeng C, Weng C-C, Lee VMY, Trojanowski JQ, et al. The Sigma-2 Receptor/TMEM97, PGRMC1, and LDL Receptor

Complex Are Responsible for the Cellular Uptake of Aβ42 and Its Protein Aggregates. Molecular Neurobiology. 2020;57(9):3803-13.

451. Colom-Cadena M, Tulloch J, Rose J, Smith C, Spires-Jones T. TMEM97 is a potential amyloid beta receptor in human Alzheimer's disease synapses. Alzheimer's & Dementia. 2020;16(S2):e041782.

452. Sebastiani P, Monti S, Morris M, Gurinovich A, Toshiko T, Andersen SL, et al. A serum protein signature of APOE genotypes in centenarians. Aging cell. 2019;18(6):e13023.

453. Wolf EJ, Logue MW, Hayes JP, Sadeh N, Schichman SA, Stone A, et al. Accelerated DNA methylation age: associations with PTSD and neural integrity. Psychoneuroendocrinology. 2016;63:155-62.

454. Bressler J, Marioni RE, Walker RM, Xia R, Gottesman RF, Windham BG, et al. Epigenetic Age Acceleration and Cognitive Function in African American Adults in Midlife: The Atherosclerosis Risk in Communities Study. The journals of gerontology Series A, Biological sciences and medical sciences. 2020;75(3):473-80.

455. Chouliaras L, Pishva E, Haapakoski R, Zsoldos E, Mahmood A, Filippini N, et al. Peripheral DNA methylation, cognitive decline and brain aging: pilot findings from the Whitehall II imaging study. Epigenomics. 2018;10(5):585-95.

456. Marioni RE, Shah S, McRae AF, Ritchie SJ, Muniz-Terrera G, Harris SE, et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. International journal of epidemiology. 2015;44(4):1388-96.

457. Degerman S, Josefsson M, Adolfsson AN, Wennstedt S, Landfors M, Haider Z, et al. Maintained memory in aging is associated with young epigenetic age. Neurobiology of aging. 2017;55:167-71.

458. Belsky DW, Moffitt TE, Cohen AA, Corcoran DL, Levine ME, Prinz JA, et al. Eleven telomere, epigenetic clock, and biomarker-composite quantifications of biological aging: do they measure the same thing? American Journal of Epidemiology. 2018;187(6):1220-30.

459. Hillary RF, Stevenson AJ, Cox SR, McCartney DL, Harris SE, Seeboth A, et al. An epigenetic predictor of death captures multi-modal measures of brain health. Molecular Psychiatry. 2019.

460. Maddock J, Castillo-Fernandez J, Wong A, Cooper R, Richards M, Ong KK, et al. DNA methylation age and physical and cognitive aging. The Journals of Gerontology: Series A. 2020;75(3):504-11.

461. Rezwan FI, Imboden M, Amaral AF, Wielscher M, Jeong A, Triebner K, et al. Association of adult lung function with accelerated biological aging. Aging (Albany NY). 2020;12(1):518.

462. McCrory C, Fiorito G, Hernandez B, Polidoro S, O'Halloran AM, Hever A, et al. GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. The Journals of Gerontology: Series A. 2021;76(5):741-9.

463. Roshandel D, Chen Z, Canty AJ, Bull SB, Natarajan R, Paterson AD. DNA methylation age calculators reveal association with diabetic neuropathy in type 1 diabetes. Clinical epigenetics. 2020;12:1-16.

464. Kim K, Joyce B, Zheng Y, Schreiner PJ, Jacobs DR, Catov JM, et al. DNA methylation GrimAge and Incident Diabetes: The Coronary Artery Risk Development in Young Adults (CARDIA) Study. Diabetes. 2021.

465. Hillary RF, Stevenson AJ, McCartney DL, Campbell A, Walker RM, Howard DM, et al. Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. Clinical Epigenetics. 2020;12(1):115.

466. Cronjé HT, Nienaber-Rousseau C, Min JL, Green FR, Elliott HR, Pieters M. Comparison of DNA methylation clocks in Black South African men. Epigenomics. 2021;13(6):437-49.

467. Suhre K, Zaghlool S. Connecting the epigenome, metabolome and proteome for a deeper understanding of disease. Journal of internal medicine. 2021.

468. Madani R, Poirier R, Wolfer DP, Welzl H, Groscurth P, Lipp HP, et al. Lack of neprilysin suffices to generate murine amyloid-like deposits in the brain and behavioral deficit in vivo. Journal of neuroscience research. 2006;84(8):1871-8.

469. Poirier R, Wolfer DP, Welzl H, Tracy J, Galsworthy MJ, Nitsch RM, et al. Neuronal neprilysin overexpression is associated with attenuation of Abeta-related spatial memory deficit. Neurobiology of disease. 2006;24(3):475-83.

470. Shi X, Ohta Y, Liu X, Shang J, Morihara R, Nakano Y, et al. Acute Anti-Inflammatory Markers ITIH4 and AHSG in Mice Brain of a Novel Alzheimer's Disease Model. J Alzheimers Dis. 2019;68(4):1667-75.

471. Yang M-H, Yang Y-H, Lu C-Y, Jong S-B, Chen L-J, Lin Y-F, et al. Activity-dependent neuroprotector homeobox protein: A candidate protein identified in serum as diagnostic biomarker for Alzheimer's disease. Journal of Proteomics. 2012;75(12):3617-29.

472. Cervantes S, Samaranch L, Vidal-Taboada JM, Lamet I, Bullido MJ, Frank-García A, et al. Genetic variation in APOE cluster region and Alzheimer's disease risk. Neurobiology of aging. 2011;32(11):2107.e7-17.

473. Yu CE, Seltman H, Peskind ER, Galloway N, Zhou PX, Rosenthal E, et al. Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. Genomics. 2007;89(6):655-65.

474. Harwood JC, Leonenko G, Sims R, Escott-Price V, Williams J, Holmans P. Defining functional variants associated with Alzheimer’s disease in the induced immune response. Brain communications. 2021;3(2).

475. Wingo AP, Fan W, Duong DM, Gerasimov ES, Dammer EB, Liu Y, et al. Shared proteomic effects of cerebral atherosclerosis and Alzheimer’s disease on the human brain. Nature neuroscience. 2020;23(6):696-700.

476. Burgess S. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. International journal of epidemiology. 2014;43(3):922-9.

477. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, Snaedal J, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. New England Journal of Medicine. 2013;368(2):107-16.

478. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. New England Journal of Medicine. 2013;368(2):117-27.

479. Yuan P, Condello C, Keene CD, Wang Y, Bird TD, Paul SM, et al. TREM2 haplodeficiency in mice and humans impairs the microglia barrier function leading to decreased amyloid compaction and severe axonal dystrophy. Neuron. 2016;90(4):724-39.

480. Cheng-Hathaway PJ, Reed-Geaghan EG, Jay TR, Casali BT, Bemiller SM, Puntambekar SS, et al. The T rem 2 R47H variant confers loss-of-function-like phenotypes in Alzheimer’s disease. Molecular neurodegeneration. 2018;13(1):1-12.

481. Leyns CE, Gratuze M, Narasimhan S, Jain N, Koscal LJ, Jiang H, et al. TREM2 function impedes tau seeding in neuritic plaques. Nature neuroscience. 2019;22(8):1217-22.

482. Parhizkar S, Arzberger T, Brendel M, Kleinberger G, Deussing M, Focke C, et al. Loss of TREM2 function increases amyloid seeding but reduces plaque-associated ApoE. Nature neuroscience. 2019;22(2):191-204.

483. You S-F, He J, Filipello F, Brase L, Del-Aguila JL, Mihindukulasuriya KA, et al. Multiomics approaches reveal a link between the MS4A gene loci, TREM2, and microglia function. Alzheimer's & Dementia. 2020;16(S2):e043592.

484. Salama M, Shalash A, Magdy A, Makar M, Roushdy T, Elbalkimy M, et al. Tubulin and Tau: Possible targets for diagnosis of Parkinson’s and Alzheimer’s diseases. PloS one. 2018;13(5):e0196436.

485. Puig B, Ferrer I, Ludueña RF, Avila J. BetaII-tubulin and phospho-tau aggregates in Alzheimer's disease and Pick's disease. J Alzheimers Dis. 2005;7(3):213-20; discussion 55-62.

486. Yi B, Sahn JJ, Ardestani PM, Evans AK, Scott LL, Chan JZ, et al. Small molecule modulator of sigma 2 receptor is neuroprotective and reduces cognitive deficits and neuroinflammation in experimental models of Alzheimer's disease. Journal of neurochemistry. 2017;140(4):561-75.

487. Izzo NJ, Staniszewski A, To L, Fa M, Teich AF, Saeed F, et al. Alzheimer's therapeutics targeting amyloid beta 1-42 oligomers I: Abeta 42 oligomer binding to specific neuronal receptors is displaced by drug candidates that improve cognitive deficits. PloS one. 2014;9(11):e111898.

488. Kloppenborg RP, Nederkoorn PJ, Geerlings MI, van den Berg E. Presence and progression of white matter hyperintensities and cognition: a meta-analysis. Neurology. 2014;82(23):2127-38.

489. Prins ND, Scheltens P. White matter hyperintensities, cognitive impairment and dementia: an update. Nature reviews Neurology. 2015;11(3):157-65.

490. Niu H, Álvarez-Álvarez I, Guillén-Grima F, Aguinaga-Ontoso I. Prevalence and incidence of Alzheimer's disease in Europe: A meta-analysis. Neurologia (Barcelona, Spain). 2017;32(8):523-32.

491. Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature genetics. 2013;45(12):1452-8.

492. Liu JZ, Erlich Y, Pickrell JK. Case–control association mapping by proxy using family history of disease. Nature genetics. 2017;49(3):325-31.

493. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7.

494. Graf G, Crowe C, Kothari M, Kwon D, Manly J, Turney I, et al. Testing DNA-methylation and blood-chemistry measures of biological aging in models of Black-White disparities in healthspan characteristics. medRxiv. 2021:2021.03.02.21252685.

495. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am J Epidemiol. 1989;129(4):687-702.

496. Joshi A, Mayr M. In aptamers they trust: caveats of the SOMAscan biomarker discovery platform from SomaLogic. Am Heart Assoc; 2018.

497. Raffield LM, Dang H, Pratte KA, Jacobson S, Gillenwater LA, Ampleford E, et al. Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. Proteomics. 2020;20(12):e1900278.

498. Finkernagel F, Reinartz S, Schuldner M, Malz A, Jansen JM, Wagner U, et al. Dual-platform affinity proteomics identifies links between the recurrence of ovarian carcinoma and proteins released into the tumor microenvironment. Theranostics. 2019;9(22):6601-17.

499. Graumann J, Finkernagel F, Reinartz S, Stief T, Brödje D, Renz H, et al. Multi-platform Affinity Proteomics Identify Proteins Linked to Metastasis and Immune Suppression in Ovarian Cancer Plasma. Front Oncol. 2019;9:1150.

500. Petrera A, von Toerne C, Behler J, Huth C, Thorand B, Hilgendorff A, et al. Multiplatform approach for plasma proteomics: complementarity of olink proximity extension assay technology to mass spectrometry-based protein profiling. J Proteome Res. 2020;20(1):751-62.

501. Ngo D, Sinha S, Shen D, Kuhn EW, Keyes MJ, Shi X, et al. Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. Circulation. 2016;134(4):270-85.

502. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010;42(7):565-9.

503. Davies G, Tenesa A, Payton A, Yang J, Harris SE, Liewald D, et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular Psychiatry. 2011;16(10):996-1005.

504. Wald A. The fitting of straight lines if both variables are subject to error. The annals of mathematical statistics. 1940;11(3):284-300.

505. Fortune MD, Guo H, Burren O, Schofield E, Walker NM, Ban M, et al. Corrigendum: Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. Nat Genet. 2015;47(8):962.

506. Lloyd-Jones LR, Holloway A, McRae A, Yang J, Small K, Zhao J, et al. The Genetic Architecture of Gene Expression in Peripheral Blood. Am J Hum Genet. 2017;100(2):371.

507. Wood AR, Tuke MA, Nalls M, Hernandez D, Gibbs JR, Lin H, et al. Whole-genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes. Human molecular genetics. 2015;24(5):1504-12.

508. McCartney DL, Min JL, Richmond RC, Lu AT, Sobczyk MK, Davies G, et al. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. Genome biology. 2021;22(1):194.

509. Perakakis N, Farr OM, Mantzoros CS. Leptin in Leanness and Obesity: JACC State-of-the-Art Review. Journal of the American College of Cardiology. 2021;77(6):745-60.

510. Li YR, Keating BJ. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. 2014;6(10):91.

511. Breeze CE, Batorsky A, Lee MK, Szeto MD, Xu X, McCartney DL, et al. Epigenome-wide association study of kidney function identifies trans-ethnic and ethnic-specific loci. Genome Med. 2021;13(1):74.

512. Braun PR, Han S, Hing B, Nagahama Y, Gaul LN, Heinzman JT, et al. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. Transl Psychiatry. 2019;9(1):47.

513. Klarić L, Tsepilov YA, Stanton CM, Mangino M, Sikka TT, Esko T, et al. Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. Science Advances. 2020;6(8):eaax0301.

514. Wahl A, Kasela S, Carnero-Montoro E, van Iterson M, Štambuk J, Sharma S, et al. IgG glycosylation and DNA methylation are interconnected with smoking. Biochimica et biophysica acta General subjects. 2018;1862(3):637-48.

515. Gadd DA, Hillary RF, McCartney DL, Zaghlool SB, Stevenson AJ, Nangle C, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. bioRxiv. 2021:2020.12.01.404681.

516. Bahar R, Hartmann CH, Rodriguez KA, Denny AD, Busuttil RA, Dollé MET, et al. Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature. 2006;441(7096):1011-4.

517. Hillary R, Marioni R. MethylDetectR: a software for methylation-based health profiling [version 2; peer review: 2 approved]. Wellcome Open Research. 2021;5(283).

518. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome biology. 2017;18(1):1-13.

519. Yi SH, Xu LC, Mei K, Yang RZ, Huang DX. Isolation and identification of age-related DNA methylation markers for forensic age-prediction. Forensic Science International: Genetics. 2014;11:117-25.

520. Lee HY, Lee SD, Shin K-J. Forensic DNA methylation profiling from evidence material for investigative leads. BMB reports. 2016;49(7):359.

521. Abbott A. European scientists seek'epigenetic clock'to determine age of refugees. Nature. 2018;561(7721):15-6.

522. Dyke SO, Cheung WA, Joly Y, Ammerpohl O, Lutsik P, Rothstein MA, et al. Epigenome data release: a participant-centered approach to privacy protection. Genome biology. 2015;16(1):1-12.

523. Dyke SO, Saulnier KM, Dupras C, Webster AP, Maschke K, Rothstein M, et al. Points-to-consider on the return of results in epigenetic research. Genome medicine. 2019;11(1):1-9.

12 Appendix - Publications

First author publications

Published

Hillary RF, Marioni RE. MethylDetectR - a software for methylation-based health profiling. *Wellcome Open Research*. 2020; 5:283

Hillary RF, Stevenson AJ, McCartney DL, Campbell A, Walker RM, Howard DM, Ritchie CW, Horvath S, Hayward C, 4 authors, Marioni RE. Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. *Clinical Epigenetics.* 2020; 12(1):115

Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, Patxot M, Ojavee SE, Zhang Q, 9 authors, Marioni RE. Multi-method genome and epigenome wide studies of inflammatory protein levels in healthy older adults. *Genome Medicine*. 2020; 12(1):60

Hillary RF*, Stevenson AJ*, Cox SR, McCartney DL, Harris SE, Seeboth A, Higham J, Sproul D, Taylor AM, 13 authors, Marioni RE. An epigenetic predictor of death captures multi-modal measures of brain health. *Molecular Psychiatry*. 2019; no pagination

Hillary RF, McCartney DL, Harris SE, Stevenson AJ, Seeboth A, Zhang Q, Liewald DC, Evans KL, Ritchie CW, 5 authors, Marioni RE. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. *Nature Communications*. 2019; 10(1):3160

McCartney DL*, Hillary RF*, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, Morris SW, Bermingham ML, Campbell A, 11 authors, Marioni RE. Epigenetic prediction of complex traits and death. *Genome Biology*. 2018; 19, 1:136

Hillary RF, FitzGerald U. A lifetime of stress: ATF6 in development and homeostasis. *Journal of Biomedical Science*. 2018; 25, 1:48

Submitted

Gadd DA*, Hillary RF*, McCartney DL*, Zaghlool S*, Stevenson AJ, Nangle C, Campbell A, Flaig R, Harris SE, 17 authors, Marioni RE. Epigenetic scores for the circulating proteome as tools for disease prediction. *bioRxiv*. 2020; doi.org/10.1101/2020.12.01.404681

Hillary RF, Gadd DA, McCartney DL, Shi L, Campbell A, Walker RM, Ritchie CW, Deary IJ, Evans KL, 6 authors, Marioni RE. Genome and epigenome wide studies of plasma protein biomarkers for Alzheimer's disease implicate TBCA and TREM2 in disease risk. *medRxiv*. 2021; doi.org/10.1101/2021.06.07.21258457.

Middle author publications

Published

Fetit R, Hillary RF, Price DJ, Lawrie SM. The Neuropathology of Autism: A Systematic Review of Post-Mortem Studies of Autism and Related Disorders. *Neuroscience and Biobehavioral Reviews*. 2021; S0149-7634(21)00313-4.

Nabais MF, Laws SM, Tin L, Vallerga CL, Armstong NJ, Blair IP, Kwok JB, Mather KA, Mellick GD, 44 authors, Hillary RF, 17 authors. Meta-analysis of genome-wide DNA methylation identifies shared associations across neurodegenerative disorders. *Genome Biology*. 2021; 26, 22:90

Gadd DA, Stevenson AJ, Hillary RF, McCartney DL, Wrobel N, McCaffety S, Murphy L, Russ TC, Harris SE, 7 authors, Marioni RE. Epigenetic predictors of lifestyle traits applied to the blood and brain. *Brain Communications.* 2021; 19, 3(2)

Stevenson AJ, Gadd DA, Hillary RF, McCartney DL, Campbell A, Walker RM, Evans KL, Harris SE, Spires-Jones TL, 4 authors, Marioni RE. Creating and validating a DNA methylation-based proxy for interleukin-6. *The Journal of Gerontology, Series A: Biological Sciences and Medical Sciences*. 2021; glab046

Green C, Shen X, Stevenson AJ, Conole ELS, Harris MA, Barbu MC, Hawkins EL, Adams MJ, Hillary RF, 14 authors, Whalley HC. Structural brain correlates of serum and epigenetic markers of inflammation in major depressive disorder. *Brain, Behavior and Immunity.* 2020; no pagination

Madden RA, McCartney DL, Walker RM, Hillary RF, Bermingham ML, Rawlik K, Morris SW, Campbell A, Porteous DJ, 4 authors, Marioni RE. Birth weight associations with DNA methylation differences in an adult population. *Epigenetics*. 2020; 1-14

Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, Walker RM, Evans KL, Boutin TS, 6 authors, Marioni RE. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clinical Epigenetics*. 2020; 12, 1:113

Seeboth A, McCartney DL, Wang Y, Hillary RF, Stevenson AJ, Walker RM, Campbell A, Evans KL, McIntosh AM, 2 authors, Marioni RE. DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936. *Clinical Epigenetics*. 2020; 12, 1:49

McCartney DL, Zhang F, Hillary RF, Zhang Q, Stevenson AJ, Walker RM, Bermingham ML, Boutin T, Morris SW, 12 authors, Marioni RE. An epigenome-wide association study of sex-specific chronological ageing. *Genome Medicine*. 2019; 12, 1:1

Stevenson AJ, McCartney DL, Hillary RF, Redmond P, Taylor AM, Zhang Q, McRae AF, Spires-Jones TL, McIntosh AM, Deary IJ, Marioni RE. Childhood intelligence attenuates the association between biological ageing and health outcomes in later life. *Translational Psychiatry*. 2019; 9, 1:323

Gibson J, Russ TC, Clarke TK, Howard DM, Hillary RF, Evans KL, Walker RM, Bermingham ML, Morris SW, 8 authors, Marioni RE. A meta-analysis of genome-wide association studies of epigenetic age acceleration. *PLoS Genetics*. 2019; 15, 11:e1008104

Robertson NA, Hillary RF, McCartney DL, Terradas-Terradas M, Higham J, Sproul D, Deary IJ, Kirschner K, Marioni RE, Chandra T. Age-related clonal haemopoiesis is associated with increased epigenetic age. *Current Biology*. 2019; 29, 16:R786-R787

McCartney DL, Stevenson AJ, Hillary RF, Walker RM, Bermingham ML, Morris SW, Clarke TK, Campbell A, Murray AD, 6 authors, Marioni RE (2018). Epigenetic signatures of starting and stopping smoking. *EBioMedicine*. 2018; 37:214-220

Johnstone M, Hillary RF, St Clair D. Stem Cells to Inform the Neurobiology of Mental Illness. *Current Topics in Behavioral Neuroscience*. 2018; 40:13-43

Submitted

McCartney DL, Hillary RF, Conole ELS, Trejo-Banos D, Gadd DA, Walker RM, Nangle C, Flaig R, Campbell A, 14 authors, Marioni RE. Blood-based epigenome-wide analyses of cognitive abilities. *medRxiv*. 2021; doi.org/10.1101/2021.05.24.21257698

Stevenson AJ, McCartney DL, Shireby GL, Hillary RF, King D, Tzioras M, Wrobel N, McCafferty S, Murphy L, 12 authors, Spires-Jones TL. A comparison of blood and brain-derived ageing and inflammation-related DNA methylation signatures and their association with microglial burdens. *bioRxiv*. 2020; doi.org/10.1101/2020.11.30.404228

*, indicates joint first authorship