Epigenetic scores for the circulating proteome as tools for disease prediction

Danni A Gadd[1,*], Robert F Hillary[1,*], Daniel L McCartney[1,*], Shaza B Zaghlool[2,3*], Anna J Stevenson[1], Cliff Nangle[1], Archie Campbell[1], Robin Flaig[1], Sarah E Harris[4,5], Rosie M Walker[6], Liu Shi[7], Elliot M Tucker-Drob[8,9], Christian Gieger[10,11,12,13], Annette Peters[11,12,13], Melanie Waldenberger[10,11,12], Johannes Graumann[14,15], Allan F McRae[16], Ian J Deary[4,5], David J Porteous[1], Caroline Hayward[1,17], Peter M Visscher[16], Simon R Cox[4,5], Kathryn L Evans[1], Andrew M McIntosh[1,18], Karsten Suhre[2], Riccardo E Marioni[1,†]

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

[2] Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, PO 24144 Doha, Qatar

[3] Computer Engineering Department, Virginia Tech, Blacksburg, VA, USA.

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

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

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

[7] Department of Psychiatry, University of Oxford, UK

[8] Department of Psychology, The University of Texas at Austin, United States

[9] Population Research Center, The University of Texas at Austin, United States

[10] Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Bavaria, Germany

[11] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Bavaria, Germany

[12] German Center for Cardiovascular Research (DZHK), partner site Munich Heart Alliance, Munich, Germany

[13] German Center for Diabetes Research (DZD), Neuherberg, Germany

[14] Scientific Service Group Biomolecular Mass Spectrometry, Max Planck Institute for Heart and Lung Research, W.G. Kerckhoff Institute, Bad Nauheim, Germany

[15] German Centre for Cardiovascular Research (DZHK), Partner Site Rhine-Main, Max Planck Institute of Heart and Lung Research, Bad Nauheim, Germany

[16] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia

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

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

* These authors contributed equally

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

Abstract

Protein biomarkers have been identified across many age-related morbidities. However, characterising epigenetic influences could further inform disease predictions. Here, we leverage epigenome-wide data to study links between the DNAm signatures of the circulating proteome and incident diseases. Using data from four cohorts, we trained and tested epigenetic scores (EpiScores) for 953 plasma proteins, identifying 109 scores that explained between 1% and 58% of the variance in protein levels after adjusting for known protein quantitative trait loci (pQTL) genetic effects. By projecting these EpiScores into an independent sample, (Generation Scotland; n=9,537) and relating them to incident morbidities over a follow-up of 14 years, we uncovered 137 EpiScore – disease associations. These associations were largely independent of immune cell proportions, common lifestyle and health factors and biological aging. Notably, we found that our diabetes-associated EpiScores highlighted previous top biomarker associations from proteome-wide assessments of diabetes. These EpiScores for protein levels can therefore be a valuable resource for disease prediction and risk stratification.

Introduction

Chronic morbidities place longstanding burdens on our health as we age. Stratifying an individual's risk prior to symptom presentation is therefore critical (NHS England, 2016). Though complex morbidities are partially driven by genetic factors (Fuchsberger et al., 2016; Yao et al., 2018), epigenetic modifications have also been associated with disease (Lord & Cruchaga, 2014). DNA methylation (DNAm) encodes information on the epigenetic landscape of an individual and blood-based DNAm signatures have been found to predict all-cause

mortality and disease onset, providing strong evidence to suggest that methylation is an important measure of disease risk (Hillary, Stevenson, et al., 2020; Lu et al., 2019; Y. Zhang et al., 2017). DNAm can regulate gene transcription (Lea et al., 2018), and epigenetic differences can be reflected in the variability of the proteome (Hillary et al., 2019; Hillary, Trejo-Banos, et al., 2020; Zaghlool et al., 2020). Low-grade inflammation, which is thought to exacerbate many age-related morbidities, is particularly well-captured through DNAm studies of plasma protein levels (Zaghlool et al., 2020). Connecting the epigenome, proteome and time to disease onset may help to identify predictive biological signatures.

Epigenetic predictors have utilised DNAm from the blood to estimate a person's 'biological age' (Lu et al., 2019), measure their exposure to lifestyle and environmental exposures (McCartney, Hillary, et al., 2018; McCartney, Stevenson, Hillary, et al., 2018; Peters et al., 2021) and predict circulating levels of inflammatory proteins (A. Stevenson et al., 2020; A. J. Stevenson et al., 2021). A leading epigenetic predictor of biological aging, the GrimAge epigenetic clock incorporates methylation scores for seven proteins along with smoking and chronological age, and is associated with numerous incident disease outcomes (Hillary, Stevenson, et al., 2020; Lu et al., 2019). This suggests that there is predictive value in utilising DNAm relevant to protein levels for disease predictions. A portfolio of protein EpiScores across the circulating proteome may aid in the prediction of disease and may offer a complementary signal to that of composite scores. Generation of an extensive range of proteomic scores has not been attempted to date. The capability of specific protein scores to predict a range of morbidities has also not been tested. However, DNAm scores for Interleukin-6 and C-Reactive protein have been found to associate with a range of phenotypes independently of measured protein levels, show more stable longitudinal trajectories than

repeated protein measurements, and, in some cases, outperform blood-based proteomic associations with brain morphology (Conole et al., 2020; A. J. Stevenson et al., 2021). This is likely due to DNA methylation reflecting a more consistent profile of stress in the body than protein measurements.

Here, we report a comprehensive association study of blood-based DNAm with proteomics and disease (**Figure 1**). We trained epigenetic scores – referred to as EpiScores – for 953 plasma proteins (with sample size ranging from 725 – 944 individuals) and validated them using two independent cohorts with 778 and 162 participants. We regressed out known genetic pQTL effects from the protein levels prior to generating the EpiScores to preclude the signatures being driven by common SNP data that are invariant across the lifespan. Finally, we examined whether the most robust predictors (n=109 EpiScores) associated with the incidence of 12 major morbidities (**Table 1**), over a follow up period of up to 14 years in the Generation Scotland cohort (n = 9,537). We regressed out the effects of age on protein levels prior to training and testing; age was also included as a covariate in the time-to-event disease prediction models. We controlled for common risk factors for disease and assessed the capacity of EpiScores to identify previously reported protein-disease associations.

	Basic model			Fully-adjusted model		
Morbidity	**N cases**	**N Controls**	**Years to event (mean, sd)**	**N cases**	**N Controls**	**Years to event (mean, sd)**
Rheumatoid arthritis	65	9281	6.1 (3.5)	54	7736	6.4 (3.3)
Alzheimer's dementia	69	3764	8.3 (2.7)	52	3137	8.2 (2.7)
Bowel cancer	77	9398	6.4 (3.2)	65	7817	6.5 (3.2)
Depression	101	8306	3.9 (3.3)	80	6976	3.7 (3.3)
Breast cancer	129	5355	6 (3.4)	110	4401	5.9 (3.4)
Lung cancer	201	9265	5.2 (3.1)	172	7705	5.1 (3.1)
Inflammatory bowel disease	203	9083	5 (3.5)	163	7567	4.9 (3.5)
Stroke	317	9023	6.5 (3.4)	248	7546	6.4 (3.5)
COPD	346	8939	6.2 (3.4)	273	7459	6.1 (3.4)
Ischaemic heart disease	395	8646	5.8 (3.3)	309	7248	5.9 (3.3)
Diabetes	428	8756	5.7 (3.4)	322	7331	5.7 (3.4)
Pain	1494	5341	5.2 (3.5)	1221	4475	5.3 (3.5)

Table 1. Incident morbidities in the Generation Scotland cohort. Counts are provided for the number of cases and controls for each incident trait in the basic and fully-adjusted Cox models run in the Generation Scotland cohort (n=9,537). Mean time-to-event is summarised in years for each phenotype. Alzheimer's dementia cases and controls were restricted to those older than 65 years. Breast cancer cases and controls were restricted to females.

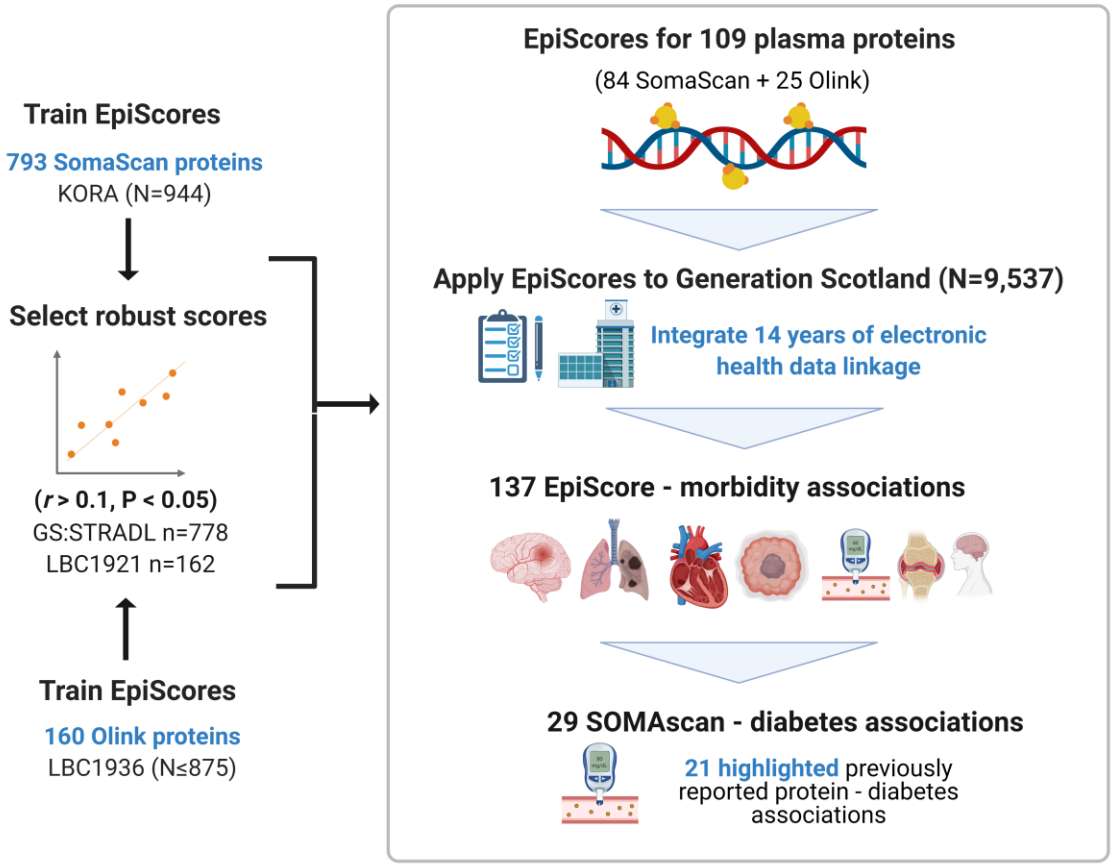

Figure 1. EpiScores for plasma proteins as tools for disease prediction study design. DNA methylation scores were trained on 953 circulating plasma protein levels in the KORA and LBC1936 cohorts. There were 109 EpiScores selected based on performance ($r > 0.1$, $P < 0.05$) in independent test sets. The selected EpiScores were projected into Generation Scotland, a cohort that has extensive data linkage to GP and hospital records. We tested whether levels of each EpiScore at baseline could predict the onset of 12 leading causes of morbidity, over a follow-up period of up to 14 years. 137 EpiScore – disease associations were identified, for 11 morbidities. We then assessed whether EpiScore associations reflected protein associations for diabetes, which is a trait that has been well-characterised using SOMAscan protein measurements. Of the 29 SOMAscan-derived EpiScore – diabetes associations, 21 reflected highlighted previously reported protein - diabetes associations.

Results

Selecting the most robust EpiScores for protein levels

To generate epigenetic scores for a comprehensive set of plasma proteins, we ran elastic net penalised regression models using protein measurements from the SOMAscan (aptamer-based) and Olink (antibody-based) platforms. We used two cohorts: the German population-based study KORA (n=944, mean age 59 years (SD 7.8), with 793 SOMAscan proteins) and the Scottish Lothian Birth Cohort 1936 (LBC1936) study (between 725 and 875 individuals in the training cohort, with a total of 160 Olink neurology and inflammatory panel proteins). The mean age of the LBC1936 participants at sampling was 70 (SD 0.8) for inflammatory and 73 (SD 0.7) for neurology proteins. Full demographic information is available for all cohorts in **Supplementary file 1A**.

Prior to running the elastic net models, we rank-based inverse normalised protein levels and adjusted for age, sex, cohort-specific variables and, where present, *cis* and *trans* pQTL effects identified from previous analyses (Hillary et al., 2019; Hillary, Trejo-Banos, et al., 2020; Suhre et al., 2017) (**Methods**). Of a possible 793 proteins in KORA, 84 EpiScores had Pearson $r >$ 0.1 and $P < 0.05$ when tested in an independent subset of Generation Scotland (The Stratifying Resilience and Depression Longitudinally [STRADL] study, n=778) (**Supplementary file 1B**). These EpiScores were selected for EpiScore-disease analyses. Of the 160 Olink proteins trained in LBC1936, there were 21 with $r > 0.1$ and $P < 0.05$ in independent test sets (STRADL, n=778, Lothian Birth Cohort 1921: LBC1921, n=162) (**Supplementary file 1C**). Independent test set data were not available for four Olink proteins. However, they were included based on

their performance ($r > 0.1$ and $P < 0.05$) in a holdout sample of 150 individuals who were left out of the training set. We then retrained these four predictors on the full training sample.

A total of 109 EpiScores (84 SOMAscan-based and 25 Olink-based) were brought forward ($r > 0.1$ and $P < 0.05$) to EpiScore-disease analyses (**Figure 2 and Supplementary file 1D)**. There were five EpiScores for proteins common to both Olink and SOMAscan panels, which had variable correlation strength (GZMA $r = 0.71$, MMP.1 $r = 0.46$, CXCL10 $r = 0.35$, NTRK3 $r = 0.26$, and CXCL11 $r = 0.09$). Predictor weights, positional information and *cis/trans* status for CpG sites contributing to these EpiScores are available in **Supplementary file 1E**. The number of CpG features selected for EpiScores ranged from one (Lyzozyme) to 395 (Aminoacylase-1), with a mean of 96 **Supplementary file 1F**). The most frequently selected CpG was the smoking-related site cg05575921 (mapping to the *AHRR* gene), which was included in 25 EpiScores. Counts for each CpG site are summarised in **Supplementary file 1G.** This table includes the set of protein EpiScores that each CpG contributes to, along with phenotypic annotations (traits) from the MRC-IEU EWAS catalog (MRC-IEU, 2021) for each CpG site having genome-wide significance ($P < 3.6 \times 10^{-8}$) (Saffari et al., 2017).

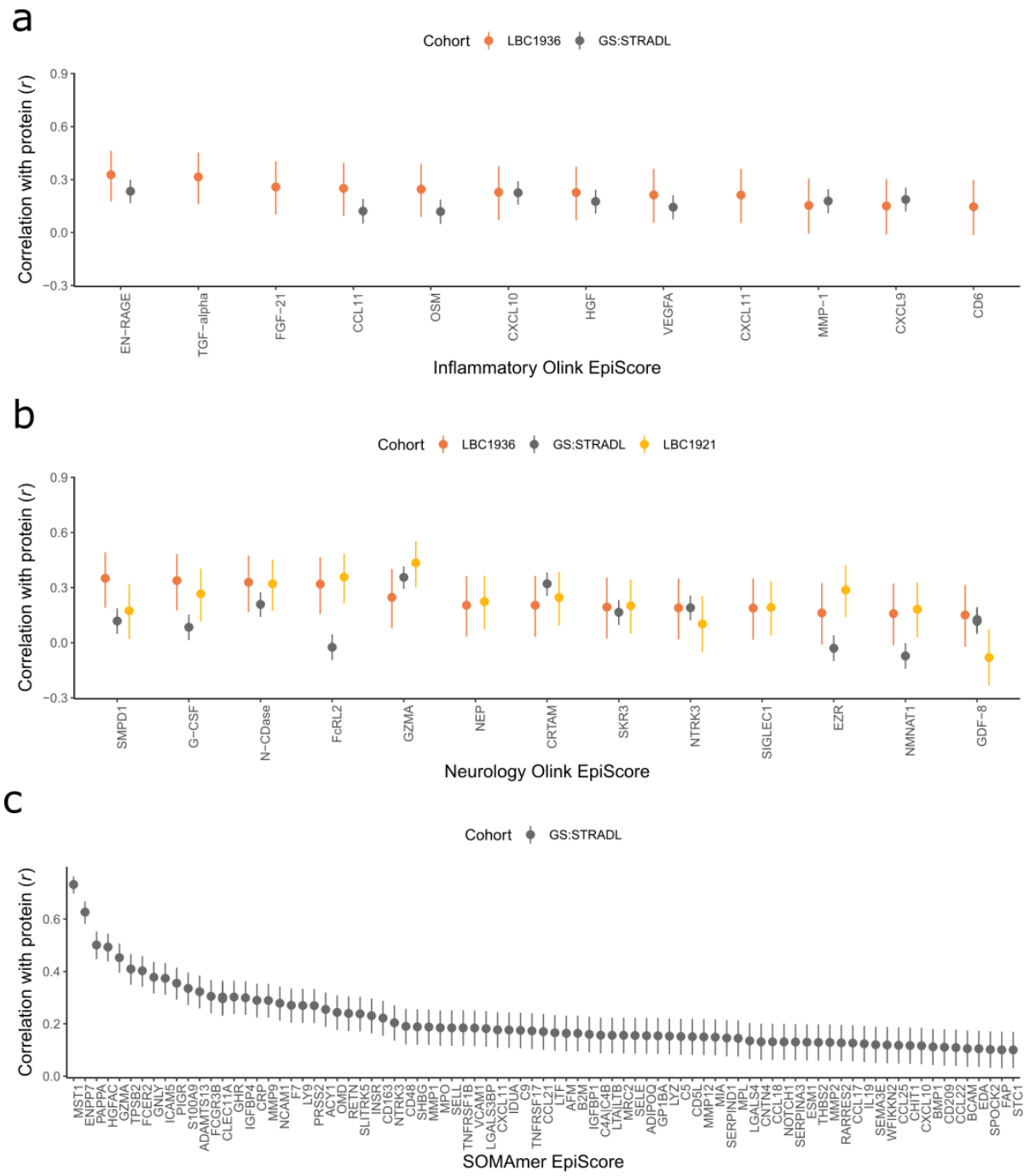

Figure 2. Test performance for the 109 selected protein EpiScores. Test set correlation coefficients for associations between protein EpiScores for (**a**) inflammatory Olink, (**b**) neurology Olink and (**c**) SOMAmer protein panel EpiScores and measured protein levels are plotted. Upper and lower confidence intervals are shown for each correlation. The 109 protein EpiScores shown achieved $r > 0.1$ and $P < 0.05$ either one or both of the GS:STRADL (n=778) and LBC1921 (n=162) test sets, wherever protein data was available for comparison. Data shown corresponds to the results included in **Supplementary files 1B-C**.

206 **EpiScore-disease associations in Generation Scotland**

207

208 The Generation Scotland dataset contains extensive electronic health data from GP and hospital
209 records available as well as DNA methylation data for 9,537 individuals. This makes it
210 uniquely positioned to test whether EpiScore signals can predict disease onset. We ran nested
211 mixed effects Cox proportional hazards models (**Figure 3**) to determine whether the levels of
212 each EpiScore at baseline associated with the incidence of 12 morbidities over a maximum of
213 14 years of follow up. The correlation structures for the 109 EpiScore measures used for Cox
214 modelling are presented in **Supplementary file 2A**.

215 The Cox proportional hazard assumption dictates that hazard ratios for EpiScore – disease
216 associations should remain constant over time. We correlated the Schoenfeld residuals from
217 the models with time to test this. Two associations in the basic model adjusting for age and sex
218 failed to satisfy the global assumption (across all covariates) and were excluded. There were
219 294 remaining EpiScore-disease associations with a False Discovery Rate (FDR)-adjusted $P <$
220 0.05 in the basic model. After further adjustment for common risk factor covariates (smoking,
221 social deprivation status, educational attainment, body mass index (BMI) and alcohol
222 consumption), 137 of the 294 EpiScore-disease associations from the basic model had $P < 0.05$
223 in the fully-adjusted model (**Supplementary files 1H-I**). Eleven of the 137 fully-adjusted
224 associations failed the Cox proportional hazards assumption for the EpiScore variable ($P < 0.05$
225 for the association between the Schoenfeld residuals and time; **Supplementary file 1J**). When
226 we restricted the time-to-event/censor period by each year of possible follow-up, there were
227 minimal differences in the EpiScore - disease hazard ratios between follow-up periods that did
228 not violate the assumption and those that did (**Supplementary file 1K**). The 137 associations
229 were therefore retained as the primary results.

The 137 associations found in the fully-adjusted model comprised 78 unique EpiScores that were related to the incidence of 11 of the 12 morbidities studied. Diabetes and chronic obstructive pulmonary disease (COPD) had the greatest number of associations, with 33 and 41, respectively. **Figure 4** presents the EpiScore-disease relationships for COPD and the remaining nine morbidities: stroke, lung cancer, ischaemic heart disease, inflammatory bowel disease, rheumatoid arthritis, depression, bowel cancer, pain and Alzheimer's dementia. There were 13 EpiScores that associated with the onset of three or more morbidities. **Figure 5** presents relationships for these 13 EpiScores in the fully-adjusted Cox model results. Of note is the EpiScore for Complement 5 (C5), which associated with five outcomes: stroke, diabetes, ischaemic heart disease, rheumatoid arthritis and COPD. Of the 29 SOMAscan-derived EpiScore associations with incident diabetes, 21 replicated previously reported protein associations (Elhadad et al., 2020; Gudmundsdottir et al., 2020) with incident or prevalent diabetes in one or more cohorts (**Figure 6 and Supplementary file 1L**).

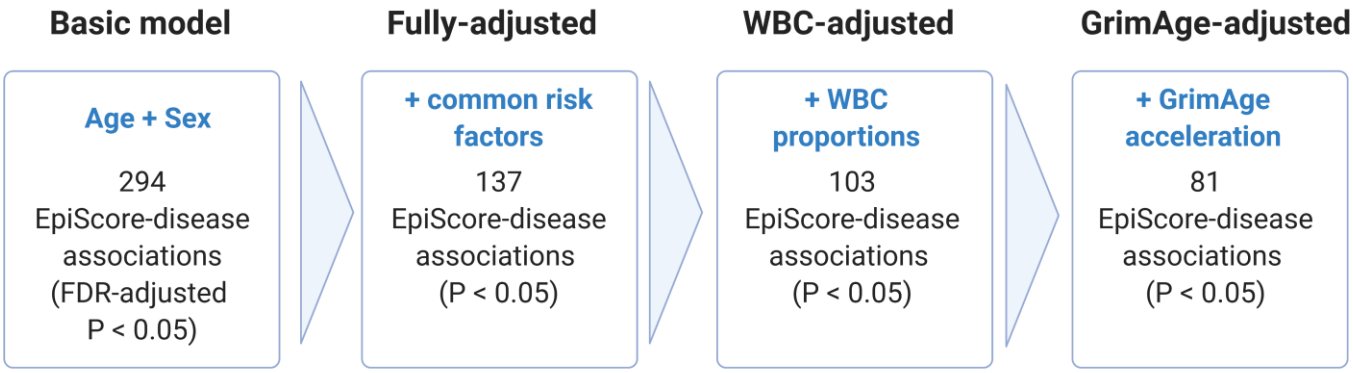

Figure 3. Nested Cox proportional hazards assessment of EpiScore-disease prediction. Mixed effects Cox proportional hazards analyses in Generation Scotland (n = 9,537) tested the relationships between each of the 109 selected EpiScores and the incidence of 12 leading causes of morbidity **(Supplementary files 1H-I)**. The basic model was adjusted for age and sex and yielded 294 associations between EpiScores and disease diagnoses, with FDR-adjusted $P < 0.05$. In the fully-adjusted model, which included common risk factors as additional covariates (smoking, deprivation, educational attainment, BMI and alcohol consumption) 137 of the basic model associations remained significant with $P < 0.05$. In a sensitivity analysis, the addition of estimated White Blood Cells (WBCs) to the fully-adjusted models led to the attenuation of 34 of the 137 associations. In a further sensitivity analysis, 81 associations remained after adjustment for both immune cell proportions and GrimAge acceleration.

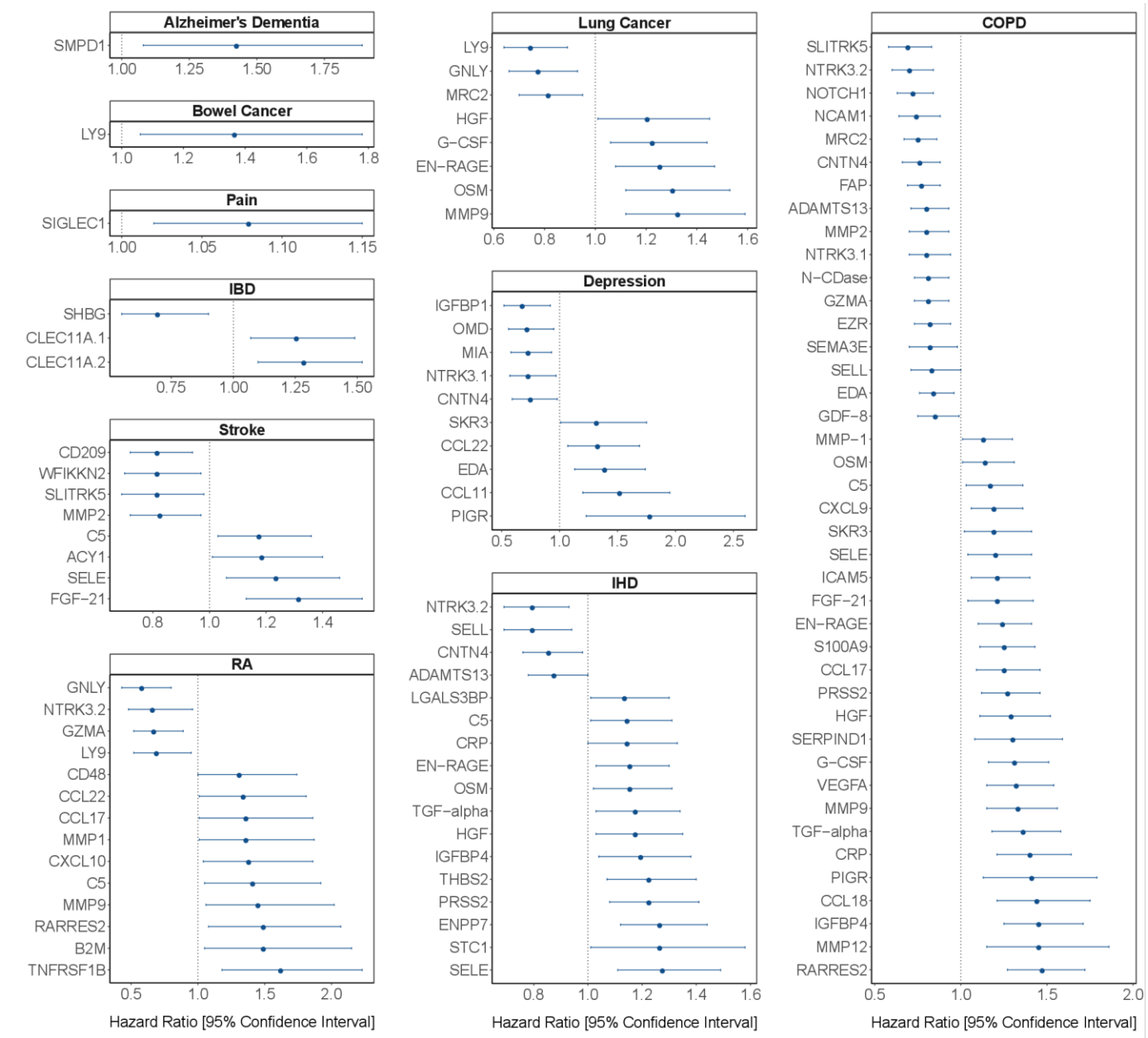

Figure 4. EpiScore associations with incident disease. EpiScore-disease associations for ten of the eleven morbidities with associations where $P < 0.05$ in the fully-adjusted mixed effects Cox proportional hazards models in Generation Scotland (n=9,537). Hazard ratios are presented with confidence intervals for 104 of the 137 EpiScore – incident disease associations reported. Models were adjusted for age, sex and common risk factors (smoking, BMI, alcohol consumption, deprivation and educational attainment). IBD: inflammatory bowel disease. IHD: ischaemic heart disease. COPD: chronic obstructive pulmonary disease. For EpiScore - diabetes associations, see **Figure 6**. Data shown corresponds to the results included in **Supplementary file 1I**.

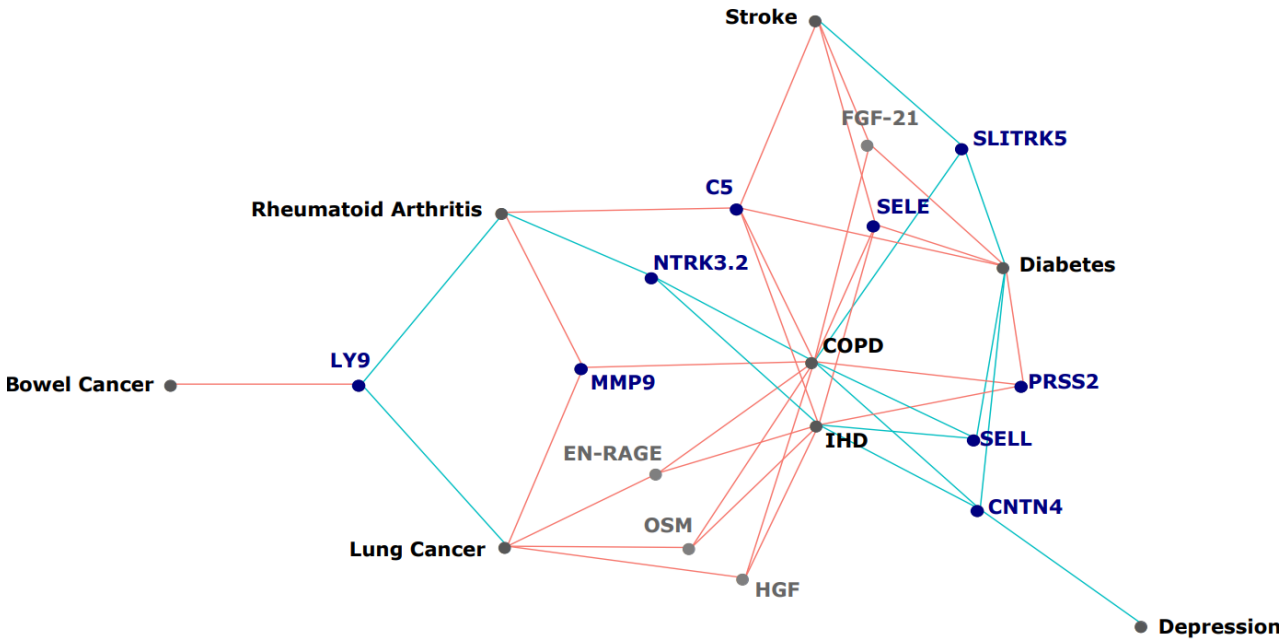

Figure 5. EpiScores that associated with the greatest number of morbidities. EpiScores with a minimum of three relationships with incident morbidities in the fully-adjusted Cox models. The network includes 13 EpiScores as dark blue (SOMAscan) and grey (Olink) nodes, with disease outcomes in black. EpiScore-disease associations with hazard ratios < 1 are shown as blue connections, whereas hazard ratios > 1 are shown in red. COPD: chronic obstructive pulmonary disease. IHD: ischaemic heart disease. Data shown corresponds to the results included in **Supplementary file 1I**.

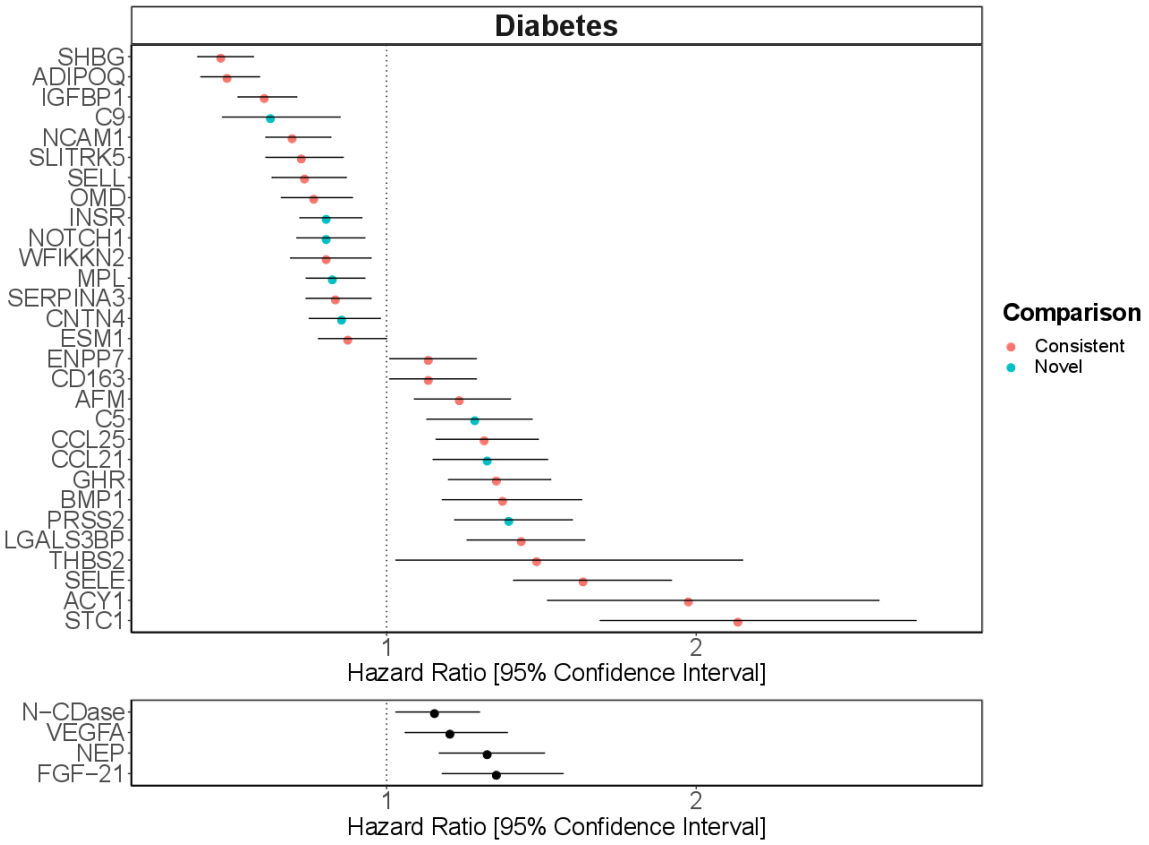

Figure 6. Replication of known protein-diabetes associations with EpiScores. EpiScore – incident diabetes associations in Generation Scotland (n=9,537). The 29 SOMAscan (top panel) and four Olink (bottom panel) associations shown with $P < 0.05$ in fully-adjusted mixed effects Cox proportional hazards models. Of the 29 SOMAscan-derived EpiScores, 21 associations were consistent with protein – diabetes associations (pink) in one or more of the four comparison cohorts that used SOMAscan protein levels. Eight associations were novel (blue). Data shown corresponds to the results included in **Supplementary files 1I and 1L** .

306 **Immune cell and GrimAge sensitivity analyses**

307

308 Correlations of the 109 EpiScores with covariates suggested interlinked relationships with both
309 immune cells and GrimAge acceleration (**Supplementary file 2B)**. These covariates were
310 therefore added incrementally to the fully-adjusted Cox models (**Figure 3**). There were 103
311 associations that remained statistically significant (FDR $P < 0.05$ in the basic model and $P <$
312 0.05 in the fully-adjusted model) after adjustment for immune cell proportions, of which 81
313 remained significant when GrimAge acceleration scores were added to this model
314 **(Supplementary file 1I)**. In a further sensitivity analysis, relationships between both estimated
315 White Blood Cell (WBC) proportions and GrimAge acceleration scores with incident diseases
316 were assessed in the Cox model structure independently of EpiScores. Of the 60 possible
317 relationships between WBC measures and the morbidities assessed, four were statistically
318 significant (FDR-adjusted $P < 0.05$) in the basic model and remained significant with $P < 0.05$
319 in the fully-adjusted model **(Supplementary file 1M**). A higher proportion of Natural Killer
320 cells was linked to decreased risk of incident COPD, rheumatoid arthritis, diabetes and pain.
321 The GrimAge acceleration composite score was associated with COPD, IHD, Diabetes and
322 Pain in the fully-adjusted models ($P < 0.05$) (**Supplementary file 1N**). The magnitude of the
323 GrimAge effect sizes were comparable to the EpiScore findings.

324

325

326

327

Discussion

Here, we report a comprehensive DNA methylation scoring study of 953 circulating proteins. We define 109 robust EpiScores for plasma protein levels that are independent of known pQTL effects. By projecting these EpiScores into a large cohort with extant data linkage, we show that 78 EpiScores associate with the incidence of 11 leading causes of morbidity (137 EpiScore – disease associations in total). Finally, we show that EpiScore - disease associations highlight previously measured protein - disease relationships. The bulk of EpiScore-disease associations are independent of common lifestyle and health factors, differences in immune cell composition and GrimAge acceleration. EpiScores therefore provide methylation-proteomic signatures for disease prediction and risk stratification.

The consistency between our EpiScore – diabetes associations and previously identified protein – diabetes relationships (Elhadad et al., 2020; Gudmundsdottir et al., 2020) suggests that epigenetic scores may identify candidate disease-protein pathways. In addition to the comprehensive lookup of SOMAscan proteins with diabetes, several of the markers we identified for COPD and IHD also reflect previous associations with measured proteins (Ganz et al., 2016; Serban et al., 2021). The two studies used for the diabetes comparison represent the largest candidate protein characterisations of diabetes to date and the top markers identified included aminoacylase-1 (ACY-1), sex hormone binding globulin (SHBG), growth hormone receptor (GHR) and Insulin-like growth factor-binding protein 2 (IGFBP-2) (Elhadad et al., 2020; Gudmundsdottir et al., 2020). Our EpiScores for these top markers are also associated with diabetes, in addition to EpiScores for several other protein markers reported in these

studies. A growing body of evidence suggests that type 2 diabetes is mediated by genetic and epigenetic regulators (Kwak & Park, 2016) and proteins such as ACY-1 and GHR are thought to influence a range of diabetes-associated metabolic mechanisms (Kim & Park, 2017; Pérez-Pérez et al., 2012). In the case of diabetes, EpiScores may therefore be used as disease-relevant risk biomarkers, many years prior to onset. Validation should be tested when sufficient data become available for the remaining morbidities.

With modest test set performances (for example, SHBG $r = 0.18$ and ACY-1 $r = 0.25$), it is perhaps surprising that such strong synergy is observed between EpiScores for proteins that associated with diabetes and the trends seen with measured proteins. Nonetheless, DNA methylation scores for CRP and IL6 have been shown to perform modestly in test sets ($r \sim 0.2$, equivalent to ~ 4% explained variance in protein level), but augment and often outperform the measured protein related to a range of phenotypes (A. Stevenson et al., 2020; A. J. Stevenson et al., 2021). Upper bounds for DNAm prediction of complex traits, such as proteins, can be estimated by variance components analyses (Hillary, Trejo-Banos, et al., 2020; Trejo Banos et al., 2020; F. Zhang et al., 2019).

Compared to epigenetic clocks like GrimAge, EpiScores enable the granular study of individual protein predictor signatures with disease outcomes. For example, levels of the acid sphingomyelinase (ASM) EpiScore predicted onset of Alzheimer's dementia, several years prior to diagnosis. ASM (encoded by *SMPD1*) has been discussed as a therapeutic candidate for Alzheimer's disease (Cataldo et al., 2004; Kamil et al., 2016; Lee et al., 2014; Park et al., 2020) and has been shown to disrupt autophagic protein degradation and associate with

374 accumulation of amyloid-beta in murine models of Alzheimer's pathology (Lee et al., 2014;
375 Park et al., 2020). The EpiScore for Complement Component 5 (C5) was associated with the
376 onset of five morbidities, the highest number for any EpiScore. Elevated levels of C5 peptides
377 have been associated with severe inflammatory, autoimmune and neurodegenerative states (Ma
378 et al., 2019; Mantovani et al., 2014; Morgan & Harris, 2015) and a range of C5-targetting
379 therapeutic approaches are in development (Alawieh et al., 2018; Brandolini et al., 2019;
380 Hawksworth et al., 2017; Hernandez et al., 2017; Morgan & Harris, 2015; Ort et al., 2020).
381 EpiScores that occupy central hubs in the disease-prediction framework may therefore provide
382 evidence of early methylation signatures common to the onset of multiple diseases. Our large-
383 scale assessment of EpiScores provides a platform for future studies, as composite predictors
384 may be created using our EpiScore database. These should be tested in incident disease
385 predictions when sufficient case data are available.

386

387 This study has several limitations. First, like with protein – disease association studies, we
388 cannot infer causality from our EpiScore – disease models. However, both protein levels and
389 EpiScores may have utility in risk prediction – future studies where both modalities are
390 available should assess paired protein and CpG contributions to traits. This should entail the
391 direct measurement of proteins, as inference from EpiScores alone, while useful for disease
392 risk stratification, is not sufficient to determine mechanisms. Second, the epitope nature of the
393 protein measurement in the SOMAscan panel may incur probe cross-reactivity and non-
394 specific binding; there may also be differences in how certain proteins are measured across
395 panels (Pietzner et al., 2020; Sun et al., 2018). Comparisons of both protein measurement
396 technologies on the same samples would help to explore this in more detail. Third, there may
397 also be pQTLs with small effect sizes that were not regressed from the proteins prior to

398 generating the EpiScores. Finally, associations present between EpiScore measures and disease
399 incidence may have been influenced by external factors such as prescription medications for
400 comorbid conditions and comorbid disease prevalence.

401

402 We have shown that EpiScores for circulating protein levels predict the incidence of multiple
403 diseases, up to 14 years prior to diagnosis. Our findings suggest that DNA methylation
404 phenotyping approaches and data linkage to electronic health records in large, population-
405 based studies have the potential to (1) Capture inter-individual variability in protein levels; (2)
406 Augment risk prediction many years prior to morbidity onset; and (3) highlight candidate
407 protein – disease associations for further exploration. The EpiScore weights are publicly
408 available, enabling any cohort with Illumina DNAm data to generate them and to relate them
409 to various outcomes. Given the increasingly widespread assessment of DNAm in cohort studies
410 (McCartney et al., 2020; Min et al., 2020), EpiScores offer an affordable and consistent (i.e.
411 array-based) way to utilise these signatures. This information is likely to be important in risk
412 stratification and prevention of age-related morbidities.

413

414 # Materials and Methods

415

416 ## The KORA sample population

417

418 The KORA F4 study includes 3,080 participants who reside in Southern Germany. Individuals
419 were between 32 and 81 years of age when recruited to the study from 2006 and 2008. In the

current study, there were 944 individuals with methylation, proteomics and genetic data available. The Infinium HumanMethylation450 BeadChip platform was used to generate DNA methylation data for these individuals. The Affymetrix Axiom array was used to generate genotyping data and the SOMAscan platform was used to generate proteomic measurements in the sample.

DNA methylation in KORA

Methylation data were generated for 1,814 individuals(Petersen et al., 2014); 944 also had protein and genotype measurements available. During preprocessing, 65 SNP probes were excluded and background correction was performed in minfi (Aryee et al., 2014). Samples with a detection rate of less than 95% were excluded. Next, the minfi R package was used to perform normalization on the intensity methylation measures (Aryee et al., 2014), with a method consistent with the Lumi:QN +BMIQ pipeline. After excluding non-cg sites and CpGs on sex chromosomes or with fewer than 100 measures available, 470,837 CpGs were available for analyses.

Proteomics in KORA

The SOMAscan platform (V3.2) (Gold et al., 2010) was used to quantify protein levels in undepleted plasma for 1129 SOMAmer probes (Suhre et al., 2017). Of the 1,000 samples provided for analysis, two samples were excluded due to errors in bio-bank sampling and one based on quality control measures. Of the 997 samples available, there were 944 individuals

443 with methylation and genotypic data. Of the 1,129 probes available, five failed the QC, leaving
444 a total of 1,124 probes for the subsequent analysis. Protein measurements were transformed
445 by rank-based inverse normalisation and regressed onto age, sex, known pQTLs and 20
446 genetic principal components of ancestry derived from the Affymetrix Axiom Array to control
447 for population structure. pQTLs for each protein were taken from a previous GWAS in the
448 sample (Suhre et al., 2017).

449

450 **The LBC1936 and LBC1921 sample populations**

451

452 The Lothian Birth Cohorts of 1921 (LBC1921; N = 550) and 1936 (LBC1936; N = 1091) are
453 longitudinal studies of aging in individuals who reside in Scotland (Deary et al., 2012; Taylor
454 et al., 2018). Participants completed an intelligence test at age 11 years and were recruited
455 for these cohorts at mean ages of 79 (LBC1921) and 70 (LBC1936). They have been
456 followed up triennially for a series of cognitive, clinical, physical and social data, along with
457 blood donations that have been used for genetic, epigenetic, and proteomic measurement.
458 DNAm, proteomic (Olink® platform) and genetic data for up to 875 individuals from Waves
459 1 and 2 of the LBC1936 (at mean ages 70 and 73 years) and 162 individuals at Wave 3 of the
460 LBC1921 (at mean age 87 years).

461

462 **DNAm in LBC1936 and LBC1921**

463

464 DNA from whole blood was assessed using the Illumina 450 K methylation array. Details of
465 quality control have been described elsewhere (Shah et al., 2014; Q. Zhang et al., 2018). Raw

intensity data were background-corrected and normalised using internal controls. Manual inspection resulted in the removal of low quality samples that presented issues related to bisulphite conversion, staining signal, inadequate hybridisation or nucleotide extension. Probes with low detection rate <95% at $P < 0.01$ and samples with low call rates (<450,000 probes detected at $P < 0.01$) were removed. Samples were also removed if they had a poor match between genotype and SNP control probes, or incorrect DNA methylation-predicted sex.

Proteomics in LBC1936 and LBC1921

Plasma samples were analysed using either the Olink® neurology 92-plex or the Olink® inflammation 92-plex proximity extension assays (Olink® Bioscience, Uppsala Sweden). One inflammatory panel protein (BDNF) failed quality control and was removed. A further 21 proteins were removed, as over 40% of samples fell below the lowest limit of detection. Two neurology proteins, MAPT and HAGH, were excluded due to >40% of observations being below the lower limit of detection. This resulted in 90 neurology (LBC1936 Wave 2) and 70 inflammatory (LBC1936 Wave 1) proteins in LBC1936 and 92 neurology proteins available in LBC1921. Protein levels were rank-based inverse normalised and regressed onto age, sex, four genetic components of ancestry derived from multidimensional scaling of the Illumina 610-Quadv1 genotype array and Olink® array plate. In LBC1936, pQTLs were adjusted for, through reference to GWAS in the samples (Hillary et al., 2019; Hillary, Trejo-Banos, et al., 2020).

Generation Scotland and STRADL sample populations

Generation Scotland: the Scottish Family Health Study (GS) is a large, family-structured, population-based cohort study of >24,000 individuals from Scotland (mean age 48 years) (Smith et al., 2006). Recruitment took place between 2006 and 2011 with a clinical visit where detailed health, cognitive, and lifestyle information was collected along with biological samples (blood, urine, saliva). In GS, there were 9,537 individuals with DNAm and phenotypic information available. The Stratifying Resilience and Depression Longitudinally (STRADL) cohort is a subset of 1,188 individuals from the GS cohort who undertook additional assessments approximately five years after the study baseline (Navrady et al., 2018).

DNA methylation in Generation Scotland and STRADL

In the GS cohort, blood-based DNA methylation was generated in two sets using the Illumina EPIC array. Set 1 comprised 5,190 related individuals whereas Set 2 comprised 4,583 individuals, unrelated to each other and to those in Set 1. During quality control, probes were removed based on visual outlier inspection, bead count <3 in over 5% of samples and samples with detection P value below adequate thresholds (McCartney, Stevenson, Walker, et al., 2018; Seeboth et al., 2020). Samples were removed based on sex mismatches, low detection P values for CpGs and saliva samples and genetic outliers (Amador et al., 2015). The quality-controlled dataset comprised 9,537 individuals (n_{Set1}=5,087, n_{Set2}=4,450). The same steps were also applied to process DNAm in STRADL.

511 **Proteomics in STRADL**

512

513 Measurements for 4,235 proteins in 1,065 individuals from the STRADL cohort were recorded
514 using the SOMAscan® technology. 793 epitopes matched between the KORA and STRADL
515 cohorts and were included for training in KORA and testing in STRADL. There were 778
516 individuals with proteomics data and DNAm data in STRADL. Protein measurements were
517 transformed by rank-based inverse normalisation and regressed onto age, sex and 20 genetic
518 principal components (derived from multidimensional scaling of genotype data from the
519 Illumina 610-Quadv1 array).

520

521 **Electronic health data linkage in Generation Scotland**

522

523 Over 98% of GS participants consented to allow access to electronic health records via data
524 linkage to GP records (Read 2 codes) and hospital records (ICD codes). Data are available
525 prospectively from the time of blood draw, yielding up to 14 years of linkage. We considered
526 incident data for 12 morbidities (**Supplementary file 3A**). Prevalent cases (ascertained via
527 retrospective ICD and Read 2 codes or self-report from a baseline questionnaire) were
528 excluded. For inflammatory bowel disease (IBD) prevalent cases were excluded based on data
529 linkage alone. Included and excluded terms can be found in **Supplementary files 4A-L**.
530 Alzheimer's dementia was limited to cases/controls with age of event/censoring ≥ 65 years.
531 Breast cancer was restricted to females only. Recurrent, major and moderate episodes of
532 depression were included in depression. Diabetes was comprised of type 2 diabetes and more

533 general diabetes codes such as diabetic retinopathy and diabetes mellitus with renal
534 manifestation. Type 1 and juvenile diabetes cases were excluded.

535

536 **Elastic net protein EpiScores**

537

538 Penalised regression models were generated for 160 proteins in LBC1936 and 793 proteins in
539 KORA using Glmnet (Version 4.0-2) (J et al., 2010) in R (Version 4.0) (R, 2020). Protein levels
540 were the outcome and there were 428,489 CpG features per model in the LBC1936 training
541 and 397,630 in the KORA training. An elastic net penalty was specified (alpha=0.5) and cross
542 validation was applied. DNAm and protein measurements were scaled to have a mean of zero
543 and variance of one.

544 In the KORA analyses, 10-fold cross validation was applied and EpiScores were tested in
545 STRADL (n=778). Of 480 EpiScores that generated ≥1 CpG features, 84 had Pearson $r > 0.1$
546 and $P < 0.05$ in STRADL. As test set comparisons were not available for every protein in the
547 LBC1936 analyses, a holdout sample was defined, with two folds set aside as test data and 10-
548 fold cross validation carried out on the remaining data (n_{train}=576, n_{test}=130 for neurology and
549 n_{train}=725, n_{test}=150 for inflammatory proteins). We retained 36 EpiScores with Pearson $r > 0.1$
550 and $P < 0.05$. New predictors for these 36 proteins were then generated using 12-fold cross
551 validation and tested externally in STRADL (n=778) and LBC1921 (n=162, for the neurology
552 panel). 21 EpiScores had $r > 0.1$ and $P < 0.05$ in at least one of the external test sets. Four
553 EpiScores did not have external comparisons and were included based on holdout performance.

554 The 109 selected EPiScores were then applied to Generation Scotland (n=9,537). DNAm at
555 each CpG site was scaled to have a mean of zero and variance of one, with scaling performed
556 separately for GS Sets.

557

558 **Associations with health linkage phenotypes in Generation Scotland**

559

560 Mixed effects Cox proportional hazards regression models adjusting for age, sex, and
561 methylation set were used to assess the relationship between 109 EpiScores and 12 morbidities
562 in Generation Scotland. Models were run using coxme (Therneau, 2020b) (Version 2.2-16)
563 with a kinship matrix accounting for relatedness in Set 1. Cases included those diagnosed after
564 baseline who had died, in addition to those who received a diagnosis and remained alive.
565 Controls were censored if disease free at time of death, or at the end of the follow-up period.
566 EpiScore levels were rank-base inverse normalised. Fully-adjusted models included: the
567 following additional covariates measured at baseline: alcohol consumption (units consumed in
568 the previous week); deprivation (assessed by the Scottish Index of Multiple Deprivation
569 (GovScot, 2016)); body mass index (kg/m^2); educational attainment (an 11-category ordinal
570 variable) and a DNAm-based score for smoking status (Bollepalli et al., 2019). A false
571 discovery rate multiple testing correction $P < 0.05$ was applied to the 1306 EpiScore-disease
572 associations (109 EpiScores by 12 incident disease traits, with 2 associations excluded for
573 failing the global proportional hazards assumption). Proportional hazards assumptions were
574 checked through Schoenfeld residuals (global test and a test for the protein-EpiScore variable)
575 using the coxph and cox.zph functions from the survival package (Therneau, 2020a) (Version
576 3.2-7). For each association failing to meet the assumption (Schoenfeld residuals $P < 0.05$), a
577 sensitivity analysis was run across yearly follow-up intervals.

578 Fully-adjusted Cox proportional hazards models were run with Houseman-estimated White
579 Blood Cell (WBC) proportions as covariates (Houseman et al., 2012). A further sensitivity
580 analyses added GrimAge acceleration (Lu et al., 2019) as an additional covariate. Basic and
581 fully-adjusted Cox models were also run with estimated Monocyte, Bcell, CD4T, CD8T and
582 Natural Killer cell proportions as predictors.

583 Correlation structures for EpiScores, DNAm-based white cell proportions and phenotypic
584 information were assessed using Pearson correlations and pheatmap (Kolde, 2019) (Version
585 1.0.12) and ggcorrplot packages (Version 0.1.3) (Kassambara, 2019). The psych package
586 (Version 2.0.9) (Revelle, 2020) was used to perform principal components analysis on
587 EpiScores. A network visualisation was produced using the ggraph package (Version 2.0.5)
588 (Pedersen, 2021). Figures 1 and 2 were created with BioRender.com.

589

590 **Consistency of disease associations between EpiScores and measured proteins**

591

592 Comparisons were conducted between EpiScore – diabetes associations and diabetes
593 associations with measured proteins using two previous large-scale proteomic studies (Elhadad
594 et al., 2020; Gudmundsdottir et al., 2020). In both studies, two cohorts were included (Study 1:
595 KORA n= 993, HUNT n= 940 (Elhadad et al., 2020), Study 2: AGES-Reykjavik n=5,438 and
596 QMDiab n=356 (Gudmundsdottir et al., 2020)). Study 1 included the KORA dataset, which we
597 use in this study to generate SOMAscan EpiScores. We characterised which SOMAscan-based
598 EpiScore – diabetes associations from our fully-adjusted results reflected those observed with
599 measured protein levels. We included basic (nominal $P < 0.05$) and fully adjusted results (with

600 either FDR or Bonferroni-corrected $P < 0.05$), wherever available, across the four cohorts
601 (**Supplementary file 1L**).

602

603

604 Acknowledgements

605

606 We are grateful to all study participants of KORA, LBC1936, LBC1921 and GS for their
607 invaluable contributions to this study.

608

609 Funding

610

611 **This research was funded in whole, or in part, by the Wellcome Trust [104036/Z/14/Z,**
612 **108890/Z/15/Z, 203771/Z/16/Z]. For the purpose of open access, the author has applied a**
613 **CC BY public copyright licence to any Author Accepted Manuscript version arising from**
614 **this submission.** The LBC1921 was supported by the UK's Biotechnology and Biological
615 Sciences Research Council (BBSRC), a Royal Society–Wolfson Research Merit Award to
616 I.J.D., and the Chief Scientist Office (CSO) of the Scottish Government's Health Directorates.
617 The LBC1936 is supported by Age UK (Disconnected Mind project, which supports S.E.H.),
618 the Medical Research Council (G0701120, G1001245, MR/M013111/1, MR/R024065/1,
619 which supports S.R.C.), and the University of Edinburgh. Genotyping was supported by the
620 Biotechnology and Biological Sciences Research Council (BB/F019394/1). Methylation
621 typing in both the LBC1921 and LBC1936 was supported by Centre for Cognitive Ageing and

622 Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional
623 Strategic Support Fund, The University of Edinburgh, and The University of Queensland.
624 Proteomic analyses in LBC1936 and LBC1921 were supported for by the Age UK grant and
625 NIH Grants R01AG054628 and R01AG05462802S1. This work was conducted in the Centre
626 for Cognitive Ageing and Cognitive Epidemiology, which was supported by the Medical
627 Research Council and Biotechnology and Biological Sciences Research Council
628 (MR/K026992/1) and which supports I.J.D. We acknowledge Grant P2CHD042849 for
629 supporting the Population Research Center at the University of Texas. This research was
630 supported by Australian National Health and Medical Research Council (grants 1010374,
631 1046880 and 1113400) and by the Australian Research Council (DP160102400). P.M.V. is
632 supported by the NHMRC Fellowship Scheme (1078037, 1078901). A.F.M. is supported by
633 the Australian Research Council Fellowship (FT200100837). P.M.V. was also funded by the
634 Australian Research Council (DP160102400 and FL180100072). Lothian Birth Cohort 1921
635 and 1936 proteomic analyses were supported by a National Institutes of Health research grant
636 (R01AG054628) which also supports E.M.T-D and S.R.C.

637 Generation Scotland received core support from the Chief Scientist Office of the Scottish
638 Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006).
639 Genotyping and DNA methylation profiling of the GS samples was carried out by the Genetics
640 Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland and was
641 funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust
642 Strategic Award "STratifying Resilience and Depression Longitudinally" ([STRADL;
643 Reference 104036/Z/14/Z]). Proteomic analyses in STRADL were supported by Dementias
644 Platform UK (DPUK). DPUK funded this work through core grant support from the Medical
645 Research Council [MR/L023784/2]. C.H. is supported by an MRC University Unit Programme

646 Grant MC_UU_00007/10 (QTL in Health and Disease). L.S. is funded by DPUK through MRC
647 (grant no. MR/L023784/2) and the UK Medical Research Council Award to the University of
648 Oxford (grant no. MC_PC_17215). L.S. also received support from the NIHR Biomedical
649 Research Centre at Oxford Health NHS Foundation Trust.

650 K.S. and S.B.Z. are supported by the Biomedical Research Program at Weill Cornell Medicine
651 in Qatar, a program funded by the Qatar Foundation. K.S. is also supported by QNRF
652 grant NPRP11C-0115-180010.

653 The KORA study was initiated and financed by the Helmholtz Zentrum München—German
654 Research Center for Environmental Health, which is funded by the German Federal Ministry
655 of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research
656 was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-
657 Maximilians-Universität, as part of LMUinnovativ. The KORA study is funded by the
658 Bavarian State Ministry of Health and Care through the research project DigiMed Bayern
659 (www.digimed-bayern.de). We gratefully acknowledge the contribution of all field staff
660 conducting the KORA F4 study.

661 D.A.G., R.F.H. and A.J.S. are supported by funding from the Wellcome Trust 4-year PhD in
662 Translational Neuroscience–training the next generation of basic neuroscientists to embrace
663 clinical research [D.A.G. and R.F.H.: 108890/Z/15/Z; A.J.S: 203771/Z/16/Z]. D.L.Mc.C. and
664 R.E.M. are supported by Alzheimer's Research UK major project grant ARUK-PG2017B−10.

665

666 ## Author contributions

667

R.E.M., D.A.G., R.F.H., D.L.Mc.C., S. B. Z., and K. S. were responsible for the conception and design of the study. D.A.G., R.F.H., D.L.Mc.C., and S. B. Z. carried out the data analyses. R.E.M. and D.A.G. drafted the article. C.N., and A.C., facilitated data linkage. R.M.W., S.E.H., R.F., L.S., E.M.T.D., A.F.M., I.J.D., C.G., A.P., M.W., D.J.P., C.H., P.M.V., S.R.C., K.L.E., and A.M.M., contributed to data collection and preparation. All authors read and approved the final manuscript.

Ethics

All KORA participants have given written informed consent and the study was approved by the Ethics Committee of the Bavarian Medical Association.

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

Ethical approval for the LBC1921 and LBC1936 studies 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; LREC/2003/2/29). In both studies, all participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration.

690 ## Availability of data and materials

691

692 The datasets generated and/or analysed during the current study are not publicly available.

693 The informed consent given by the KORA study participants does not cover posting of
694 participant level phenotype and genotype data in public databases. However, data are
695 available upon request from KORA-gen (http://epi.helmholtz-muenchen.de/kora-gen).
696 Requests are submitted online and are subject to approval by the KORA board.

697 Lothian Birth Cohort data are available on request from the Lothian Birth Cohort Study,
698 University of Edinburgh (simon.cox@ed.ac.uk). Lothian Birth Cohort data are not publicly
699 available due to them containing information that could compromise participant consent and
700 confidentiality.

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

704 Code is available with open access at the following Gitlab repository:
705 https://gitlab.com/dannigadd/episcores-for-protein-levels.

706

707 ## Competing interests

708

709 R.E.M has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock
710 Development Foundation. All other authors declare no competing interests.

References

Alawieh, A., Farris Langley, E., & Tomlinson, S. (2018). Targeted complement inhibition salvages stressed neurons and inhibits neuroinflammation after stroke in mice. *Science Translational Medicine*, *10*(441), 6459. https://doi.org/10.1126/scitranslmed.aao6459

Amador, C., Huffman, J., Trochet, H., Campbell, A., Porteous, D., Wilson, J. F., Hastie, N., Vitart, V., Hayward, C., Navarro, P., Haley, C. S., & Generation Scotland. (2015). Recent genomic heritage in Scotland. *BMC Genomics*, *16*(1), 437. https://doi.org/10.1186/s12864-015-1605-2

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, *30*(10), 1363–1369. https://doi.org/10.1093/bioinformatics/btu049

Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S., & Ollikainen, M. (2019). EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. *Epigenomics*, *11*(13), 1469–1486. https://doi.org/10.2217/epi-2019-0206

Brandolini, L., Grannonico, M., Bianchini, G., Colanardi, A., Sebastiani, P., Paladini, A., Piroli, A., Allegretti, M., Varrassi, G., & Di Loreto, S. (2019). The Novel C5aR Antagonist DF3016A Protects Neurons Against Ischemic Neuroinflammatory Injury. *Neurotoxicity Research*, *36*(1), 163–174. https://doi.org/10.1007/s12640-019-00026-w

Cataldo, A. M., Petanceska, S., Terio, N. B., Peterhoff, C. M., Durham, R., Mercken, M., Mehta, P. D., Buxbaum, J., Haroutunian, V., & Nixon, R. A. (2004). Aβ localization in abnormal endosomes: Association with earliest Aβ elevations in AD and Down syndrome. *Neurobiology of Aging*, *25*(10), 1263–1272.

https://doi.org/10.1016/j.neurobiolaging.2004.02.027

Conole, E. L. S., Stevenson, A. J., Green, C., Harris, S. E., Muñoz, S., Valdés-hernández, M. C., Harris, M. A., Bastin, M. E., & Joanna, M. (2020). *An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. medRxiv.* https://doi.org/https://doi.org/10.1101/2020.10.08.20205245

Deary, I. J., Gow, A. J., Pattie, A., & Starr, J. M. (2012). Cohort profile: The lothian birth cohorts of 1921 and 1936. *International Journal of Epidemiology*, *41*(6), 1576–1584.

Elhadad, M. A., Jonasson, C., Huth, C., Wilson, R., Gieger, C., Matias, P., Grallert, H., Graumann, J., Gailus-Durner, V., Rathmann, W., Toerne, C. von, Hauck, S. M., Koenig, W., Sinner, M. F., Oprea, T. I., Suhre, K., Thorand, B., Hveem, K., Peters, A., & Waldenberger, M. (2020). Deciphering the plasma proteome of type 2 diabetes. *Diabetes*, *69*(12), 2766–2778. https://doi.org/10.2337/db20-0296

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M. A., B Perry, J. R., Sim, X., Blackwell, T. W., Robertson, N. R., William Rayner, N., Cingolani, P., Locke, A. E., Fernandez Tajes, J., … McCarthy, M. I. (2016). The genetic architecture of type 2 diabetes. *Nature Publishing Group*, *536*, 41–47.

Ganz, P., Heidecker, B., Hveem, K., Jonasson, C., Kato, S., Segal, M. R., Sterling, D. G., & Williams, S. A. (2016). Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA*, *315*(23), 2532–2541. https://doi.org/10.1001/jama.2016.5951

Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E. N., Carter, J., Dalby, A. B., Eaton, B. E., Fitzwater, T., Flather, D., Forbes, A., Foreman, T., Fowler, C., Gawande, B., Goss, M., Gunn, M., Gupta, S., Halladay, D., … Zichi, D. (2010). Aptamer-based

759 multiplexed proteomic technology for biomarker discovery. *PLoS ONE*, *5*(12), e15004.
760 https://doi.org/10.1371/journal.pone.0015004
761 GovScot. (2016). *Scottish Government. The Scottish Index of Multiple Deprivation (SIMD);*
762 *1-20. (2016). Available from: http://www.gov.scot/ Resource/0050/00504809.pdf.*
763 *Accessed April 2021.*
764 Gudmundsdottir, V., Zaghlool, S. B., Emilsson, V., Aspelund, T., Ilkov, M., Gudmundsson,
765 E. F., Jonsson, S. M., Zilhão, N. R., Lamb, J. R., Suhre, K., Jennings, L. L., &
766 Gudnason, V. (2020). Circulating protein signatures and causal candidates for type 2
767 diabetes. *Diabetes*, *69*(8), 1843–1853. https://doi.org/10.2337/db19-1070
768 Hawksworth, O. A., Li, X. X., Coulthard, L. G., Wolvetang, E. J., & Woodruff, T. M. (2017).
769 New concepts on the therapeutic control of complement anaphylatoxin receptors.
770 *Molecular Immunology*, *89*, 36–43. https://doi.org/10.1016/j.molimm.2017.05.015
771 Hernandez, M. X., Jiang, S., Cole, T. A., Chu, S. H., Fonseca, M. I., Fang, M. J., Hohsfield,
772 L. A., Torres, M. D., Green, K. N., Wetsel, R. A., Mortazavi, A., & Tenner, A. J. (2017).
773 Prevention of C5aR1 signaling delays microglial inflammatory polarization, favors
774 clearance pathways and suppresses cognitive loss. *Molecular Neurodegeneration*, *12*(1),
775 66. https://doi.org/10.1186/s13024-017-0210-z
776 Hillary, R. F., McCartney, D. L., Harris, S. E., Stevenson, A. J., Seeboth, A., Zhang, Q.,
777 Liewald, D. C., Evans, K. L., Ritchie, C. W., Tucker-Drob, E. M., Wray, N. R., McRae,
778 A. F., Visscher, P. M., Deary, I. J., & Marioni, R. E. (2019). Genome and epigenome
779 wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936.
780 *Nature Communications*, *10*(1), 3160. https://doi.org/10.1038/s41467-019-11177-x
781 Hillary, R. F., Stevenson, A. J., McCartney, D. L., Campbell, A., Walker, R. M., Howard, D.
782 M., Ritchie, C. W., Horvath, S., Hayward, C., McIntosh, A. M., Porteous, D. J., Deary,

I. J., Evans, K. L., & Marioni, R. E. (2020). Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. *Clin Epigenet*, *12*(115). https://doi.org/10.1186/s13148-020-00905-6

Hillary, R. F., Trejo-Banos, D., Kousathanas, A., McCartney, D. L., Harris, S. E., Stevenson, A. J., Patxot, M., Ojavee, S. E., Zhang, Q., Liewald, D. C., Ritchie, C. W., Evans, K. L., Tucker-Drob, E. M., Wray, N. R., McRae, A. F., Visscher, P. M., Deary, I. J., Robinson, M. R., & Marioni, R. E. (2020). Multi-method genome- And epigenome-wide studies of inflammatory protein levels in healthy older adults. *Genome Medicine*, *12*, 60. https://doi.org/10.1186/s13073-020-00754-1

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., & Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, *13*, 86. https://doi.org/10.1186/1471-2105-13-86

J, F., T, H., & R, T. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Kamil, B., Anna, F., Anna, S., Sławomir, P., & Halina, C. (2016). Regulation of sphingomyelin metabolism. *Pharmacological Reports*, *68*(3), 570–581. https://doi.org/10.1016/j.pharep.2015.12.008

Kassambara, A. (2019). *ggcorrplot: Visualization of a Correlation Matrix using "ggplot2". R package (Version 0.1.3). https://cran.r-project.org/web/packages/ggcorrplot/ggcorrplot.pdf. Accessed April 2021*.

Kim, S.-H., & Park, M.-J. (2017). Effects of growth hormone on glucose metabolism and insulin resistance in human. *Annals of Pediatric Endocrinology & Metabolism*, *22*(3), 145–152. https://doi.org/10.6065/apem.2017.22.3.145

807 Kolde, R. (2019). *Pheatmap : Pretty Heatmaps. R package version 1.0.12. https://cran.r-*
808 *project.org/web/packages/pheatmap/index.html. Accessed April 2021*.
809 Kwak, S. H., & Park, K. S. (2016). Recent progress in genetic and epigenetic research on
810 type 2 diabetes. *Experimental & Molecular Medicine*, *48*(3), e220.
811 Lea, A. J., Vockley, C. M., Johnston, R. A., Del Carpio, C. A., Barreiro, L. B., Reddy, T. E.,
812 & Tung, J. (2018). Genome-wide quantification of the effects of DNA methylation on
813 human gene regulation. *ELife*, eLife 2018;7:e37513. https://doi.org/10.7554/eLife.37513
814 Lee, J. K., Jin, H. K., Park, M. H., Kim, B. R., Lee, P. H., Nakauchi, H., Carter, J. E., He, X.,
815 Schuchman, E. H., & Bae, J. S. (2014). Acid sphingomyelinase modulates the
816 autophagic process by controlling lysosomal biogenesis in Alzheimer's disease. *Journal*
817 *of Experimental Medicine*, *211*(8), 1551–1570. https://doi.org/10.1084/jem.20132451
818 Lord, J., & Cruchaga, C. (2014). *The epigenetic landscape of Alzheimer's disease*. *9*, 1921.
819 Lu, A. T., Quach, A., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., Hou, L., Baccarelli, A.
820 A., Li, Y., Stewart, J. D., Whitsel, E. A., Assimes, T. L., Ferrucci, L., & Horvath, S.
821 (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*,
822 *11*(2), 303–327. https://doi.org/10.18632/aging.101684
823 Ma, Y., Liu, Y., Zhang, Z., & Yang, G. Y. (2019). Significance of complement system in
824 ischemic stroke: A comprehensive review. *Aging and Disease*, *10*(2), 429–462.
825 https://doi.org/10.14336/AD.2019.0119
826 Mantovani, S., Gordon, R., Macmaw, J. K., Pfluger, C. M. M., Henderson, R. D., Noakes, P.
827 G., McCombe, P. A., & Woodruff, T. M. (2014). Elevation of the terminal complement
828 activation products C5a and C5b-9 in ALS patient blood. *Journal of Neuroimmunology*,
829 *276*(1–2), 213–218. https://doi.org/10.1016/j.jneuroim.2014.09.005
830 McCartney, D. L., Hillary, R. F., Stevenson, A. J., Ritchie, S. J., Walker, R. M., Zhang, Q.,

831 Morris, S. W., Bermingham, M. L., Campbell, A., Murray, A. D., Whalley, H. C., Gale,
832 C. R., Porteous, D. J., Haley, C. S., McRae, A. F., Wray, N. R., Visscher, P. M.,
833 McIntosh, A. M., Evans, K. L., … Marioni, R. E. (2018). Epigenetic prediction of
834 complex traits and death. *Genome Biology*, *19*(1), 136.
835 McCartney, D. L., Min, J. L., Richmond, R. C., Lu, A. T., Sobczyk, M. K., Davies, G., Broer,
836 L., Guo, X., Jeong, A., Jung, J., Kasela, S., Katrinli, S., Kuo, P. L., Matias-Garcia, P. R.,
837 Mishra, P. P., Nygaard, M., Palviainen, T., Patki, A., Raffield, L. M., … Marioni, R. E.
838 (2020). Genome-wide association studies identify 137 loci for DNA methylation
839 biomarkers of ageing. *BioRxiv*. https://doi.org/10.1101/2020.06.29.133702
840 McCartney, D. L., Stevenson, A. J., Hillary, R. F., Walker, R. M., Bermingham, M. L.,
841 Morris, S. W., Clarke, T. K., Campbell, A., Murray, A. D., Whalley, H. C., Porteous, D.
842 J., Visscher, P. M., McIntosh, A. M., Evans, K. L., Deary, I. J., & Marioni, R. E. (2018).
843 Epigenetic signatures of starting and stopping smoking. *EBioMedicine*, *37*, 214–220.
844 https://doi.org/10.1016/j.ebiom.2018.10.051
845 McCartney, D. L., Stevenson, A. J., Walker, R. M., Gibson, J., Morris, S. W., Campbell, A.,
846 Murray, A. D., Whalley, H. C., Porteous, D. J., McIntosh, A. M., Evans, K. L., Deary, I.
847 J., & Marioni, R. E. (2018). Investigating the relationship between DNA methylation
848 age acceleration and risk factors for Alzheimer's disease. *Alzheimer's and Dementia:*
849 *Diagnosis, Assessment and Disease Monitoring*, *10*, 429–437.
850 https://doi.org/10.1016/j.dadm.2018.05.006
851 Min, J. L., Hemani, G., Hannon, E., Dekkers, K. F., Castillo-Fernandez, J., Luijk, R.,
852 Carnero-Montoro, E., Lawson, D. J., Burrows, K., Suderman, M., Bretherick, A. D.,
853 Richardson, T. G., Klughammer, J., Iotchkova, V., Sharp, G., Khleifat, A. Al, Shatunov,
854 A., Iacoangeli, A., McArdle, W. L., … Relton, C. L. (2020). Genomic and phenomic

insights from an atlas of genetic effects on DNA methylation. *MedRxiv*. https://doi.org/10.1101/2020.09.01.20180406

Morgan, B. P., & Harris, C. L. (2015). Complement, a target for therapy in inflammatory and degenerative diseases. *Nature Reviews Drug Discovery*, *14*, 857–877. https://doi.org/10.1038/nrd4657

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

Navrady, L. B., Wolters, M. K., MacIntyre, D. J., Clarke, T. K., Campbell, A. I., Murray, A. D., Evans, K. L., Seck, J., Haley, C., Milburn, K., Wardlaw, J. M., Porteous, D. J., Deary, I. J., & McIntosh, A. M. (2018). 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*, *47*(1), 13-14g. https://doi.org/10.1093/ije/dyx115

NHS England. (2016). *Improving Outcomes Through Personalised Medicine. https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf. Accessed April 2021.*

Ort, M., Dingemanse, J., van den Anker, J., & Kaufmann, P. (2020). Treatment of Rare Inflammatory Kidney Diseases: Drugs Targeting the Terminal Complement Pathway. In *Frontiers in Immunology* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2020.599417

Park, M. H., Jin, H. K., & Bae, J. sung. (2020). Potential therapeutic target for aging and age-related neurodegenerative diseases: the role of acid sphingomyelinase. *Experimental and Molecular Medicine*, *52*, 380–389. https://doi.org/10.1038/s12276-020-0399-8

Pedersen, T. L. (2021). *ggraph: An Implementation of Grammar of Graphics for Graphs and*

879 *Networks. R package version 2.0.5. https://CRAN.R-project.org/package=ggraph.*
880 *Accessed April 2021*.
881 Pérez-Pérez, R., García-Santos, E., Ortega-Delgado, F. J., López, J. A., Camafeita, E., Ricart,
882 W., Fernández-Real, J. M., & Peral, B. (2012). Attenuated metabolism is a hallmark of
883 obesity as revealed by comparative proteomic analysis of human omental adipose tissue.
884 *Journal of Proteomics*, *75*(3), 783–795. https://doi.org/10.1016/j.jprot.2011.09.016
885 Peters, A., Nawrot, T. S., & Baccarelli, A. A. (2021). Hallmarks of environmental insults.
886 *Cell*, *184*(6), 1455–1468. https://doi.org/10.1016/j.cell.2021.01.043
887 Petersen, A. K., Zeilinger, S., Kastenmüller, G., Werner, R. M., Brugger, M., Peters, A.,
888 Meisinger, C., Strauch, K., Hengstenberg, C., Pagel, P., Huber, F., Mohney, R. P.,
889 Grallert, H., Illig, T., Adamski, J., Waldenberger, M., Gieger, C., & Suhre, K. (2014).
890 Epigenetics meets metabolomics: An epigenome-wide association study with blood
891 serum metabolic traits. *Human Molecular Genetics*, *23*(2), 534–545.
892 https://doi.org/10.1093/hmg/ddt430
893 Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Raffler, J., Kerrison, N. D., Oerton, E.,
894 Auyeung, V. P. W., Luan, J., Finan, C., Casas, J. P., Ostroff, R., Williams, S. A.,
895 Kastenmüller, G., Ralser, M., Gamazon, E. R., Wareham, N. J., Hingorani, A. D., &
896 Langenberg, C. (2020). Genetic architecture of host proteins involved in SARS-CoV-2
897 infection. *Nature Communications*, *11*, 6397. https://doi.org/10.1038/s41467-020-
898 19996-z
899 R. (2020). *R Core Team. R: A language and environment for statistical computing. R*
900 *Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*
901 *Accessed April 2021*.
902 Revelle, W. (2020). *psych: Procedures for Personality and Psychological Research,*

Northwestern University, Evanston, Illinois, USA. (Verison 2.0.9). https://CRAN.R-project.org/package=psych. Accessed April 2021.

Saffari, A., Silver, M. J., Zavattari, P., Moi, L., Columbano, A., Meaburn, E. L., & Dudbridge, F. (2017). *Estimation of a significance threshold for epigenome-wide association studies*. *42*(1), 22–23.

Seeboth, A., McCartney, D. L., Wang, Y., Hillary, R. F., Stevenson, A. J., Walker, R. M., Campbell, A., Evans, K. L., McIntosh, A. M., Hägg, S., Deary, I. J., & Marioni, R. E. (2020). DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936. *Clinical Epigenetics*, *12*, 49. https://doi.org/10.1186/s13148-020-00838-0

Serban, K. A., Pratte, K. A., & Bowler, R. P. (2021). Protein biomarkers for COPD outcomes. *Chest*, ISSN 0012-3692. https://doi.org/10.1016/j.chest.2021.01.004

Shah, S., McRae, A. F., Marioni, R. E., Harris, S. E., Gibson, J., Henders, A. K., Redmond, P., Cox, S. R., Pattie, A., Corley, J., Murphy, L., Martin, N. G., Montgomery, G. W., Starr, J. M., Wray, N. R., Deary, I. J., & Visscher, P. M. (2014). Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Research*, *24*(11), 1725–1733.

Smith, B. H., Campbell, H., Blackwood, D., Connell, J., Connor, M., Deary, I. J., Dominiczak, A. F., Fitzpatrick, B., Ford, I., Jackson, C., Haddow, G., Kerr, S., Lindsay, R., McGilchrist, M., Morton, R., Murray, G., Palmer, C. N. A., Pell, J. P., Ralston, S. H., … Morris, A. D. (2006). Generation Scotland: The Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Medical Genetics*, *7*, 74.

Stevenson, A. J., Gadd, D. A., Hillary, R. F., McCartney, D. L., Campbell, A., Walker, R. M., Evans, K. L., Harris, S. E., Spires-Jones, T. L., McRae, A. F., Visscher, P. M.,

McIntosh, A. M., Deary, I. J., & Marioni, R. E. (2021). Creating and validating a DNA methylation-based proxy for interleukin-6. *Gerontol A Biol Sci Med Sci.*, 17:glab046. https://doi.org/10.1093/gerona/glab046.

Stevenson, A., McCartney, D., Hillary, R., Campbell, A., Morris, S., Bermingham, M., Walker, R., Evans, K., Boutin, T., Hayward, C., McRae, A., McColl, B., Spires-Jones, T., McIntosh, A., Deary, I., & Marioni, R. (2020). Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clinical Epigenetics*, *12*, 113. https://doi.org/10.1101/802009

Suhre, K., Arnold, M., Bhagwat, A. M., Cotton, R. J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., Delisle, R. K., Gold, L., Pezer, M., Lauc, G., Selim, M. A. E. D., Mook-Kanamori, D. O., Al-Dous, E. K., Mohamoud, Y. A., Malek, J., Strauch, K., … Graumann, J. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications*, *8*, 14357. https://doi.org/10.1038/ncomms14357

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M. A., Prins, B. P., Wilcox, S. K., Zimmerman, E. S., Chi, A., Bansal, N., Spain, S. L., Wood, A. M., … Butterworth, A. S. (2018). Genomic atlas of the human plasma proteome. *Nature*, *558*, 73–79. https://doi.org/10.1038/s41586-018-0175-2

Taylor, A. M., Pattie, A., & Deary, I. J. (2018). Cohort profile update: The Lothian birth cohorts of 1921 and 1936. *International Journal of Epidemiology*, *47*(4), 1042r.

Therneau, T. M. (2020a). *A Package for Survival Analysis in R. R package version 3.2-7, https://CRAN.R-project.org/package=survival. Accessed April 2021*.

Therneau, T. M. (2020b). *coxme: Mixed Effects Cox Models. R package version 2.2-16.*

https://CRAN.R-project.org/package=coxme. Accessed April 2021.

Trejo Banos, D., McCartney, D. L., Patxot, M., Anchieri, L., Battram, T., Christiansen, C., Costeira, R., Walker, R. M., Morris, S. W., Campbell, A., Zhang, Q., Porteous, D. J., McRae, A. F., Wray, N. R., Visscher, P. M., Haley, C. S., Evans, K. L., Deary, I. J., McIntosh, A. M., … Robinson, M. R. (2020). Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications*, *11*(2865).

Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., Sun, B. B., Laser, A., Maranville, J. C., Wu, H., Ho, J. E., Courchesne, P., Lyass, A., Larson, M. G., Gieger, C., Graumann, J., Johnson, A. D., Danesh, J., Runz, H., … Levy, D. (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, *9*, 3268.

Zaghlool, S. B., Kühnel, B., Elhadad, M. A., Kader, S., Halama, A., Thareja, G., Engelke, R., Sarwath, H., Al-Dous, E. K., Mohamoud, Y. A., Meitinger, T., Wilson, R., Strauch, K., Peters, A., Mook-Kanamori, D. O., Graumann, J., Malek, J. A., Gieger, C., Waldenberger, M., & Suhre, K. (2020). Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nature Communications*, *11*(1), 15. https://doi.org/10.1038/s41467-019-13831-w

Zhang, F., Chen, W., Zhu, Z., Zhang, Q., Nabais, M. F., Qi, T., Deary, I. J., Wray, N. R., Visscher, P. M., McRae, A. F., & Yang, J. (2019). OSCA: A tool for omic-data-based complex trait analysis. *Genome Biology*, *20*(107). https://doi.org/10.1186/s13059-019-1718-z

Zhang, Q., Marioni, R. E., Robinson, M. R., Higham, J., Sproul, D., Wray, N. R., Deary, I. J., McRae, A. F., & Visscher, P. M. (2018). Genotype effects contribute to variation in longitudinal methylome patterns in older people. *Genome Medicine*, *10*, 75.

975 Zhang, Y., Wilson, R., Heiss, J., Breitling, L. P., Saum, K. U., Schöttker, B., Holleczek, B.,
976 Waldenberger, M., Peters, A., & Brenner, H. (2017). DNA methylation signatures in
977 peripheral blood strongly predict all-cause mortality. *Nature Communications*, *6*, 10192.

978

979 # Additional files

980 **Supplementary file 1**

981 (A) Demographic and array information for the cohorts and samples used in the study.

982 (B) SomaScan panel performance in the STRADL test set.

983 (C) Performance of Olink protein EpiScores in holdout, STRADL and LBC1921 test sets.

984 (D) Annotations for the proteins corresponding to the 109 selected EpiScores.

985 (E) Predictor weights for the 109 selected EpiScores.

986 (F) CpG feature counts for the 109 selected EpiScores.

987 (G) Frequency of CpG sites selected for EpiScores with EWAS catalogue annotations to
988 phenotypic traits.

989 (H) Basic Cox model results in Generation Scotland.

990 (I) Fully-adjusted and sensitivity analyses results for Cox models in Generation Scotland.

991 (J) Schoenfeld residual Cox sensitivity analyses.

992 (K) Schoenfeld residual Cox sensitivity analyses split by year of follow-up.

993 (L) SOMAscan – EpiScore diabetes association lookup against two large-scale plasma
994 protein – diabetes studies.

995 (M) White blood cell sensitivity analyses.

996 (N) GrimAge sensitivity analyses.

997 **Supplementary file 2**

998 (A) Correlation structures for the 109 selected EpiScores.

999 (B) Correlation structures for the 109 selected EpiScores in relation to common covariates.

1000 **Supplementary file 3**

1001 (A) Summary of the rationale for including each of the 12 morbidities in this study.

1002 **Supplementary file 4**

1003 (A-L) Primary and secondary health codes for each of the 12 morbidities in this study that

1004 were used to assign case/control status of participants.

1005