RESEARCH ARTICLE

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], Yipeng Cheng[1], Chloe Fawns-Ritchie[1,4], 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, United Kingdom; [2]Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, Doha, Qatar; [3]Computer Engineering Department, Virginia Tech, Blacksburg, United States; [4]Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom; [5]Lothian Birth Cohorts, University of Edinburgh, Edinburgh, United Kingdom; [6]Centre for Clinical Brain Sciences, Chancellor's Building, University of Edinburgh, Edinburgh, United Kingdom; [7]Department of Psychiatry, University of Oxford, Oxford, United Kingdom; [8]Department of Psychology, The University of Texas at Austin, Austin, United States; [9]Population Research Center, The University of Texas at Austin, Austin, United States; [10]Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [11]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, 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, Australia; [17]Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom; [18]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, United Kingdom

***For correspondence:** Riccardo.Marioni@ed.ac.uk

[†]These authors contributed equally to this work

Competing interest: See page 16

Funding: See page 16

Preprinted: 02 December 2020
Received: 30 June 2021
Accepted: 11 January 2022
Published: 13 January 2022

Reviewing Editor: YM Dennis Lo, The Chinese University of Hong Kong, Hong Kong

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 DNA methylation (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 = 9537) and relating them to incident morbidities over a follow-up of 14 years, we uncovered

130 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.

Editor's evaluation

This is an important study that demonstrates the potential utility of the circulating proteome 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 and 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 et al., 2020a***; ***Lu et al., 2019***; ***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 et al., 2020b***; ***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***). As proteins are the primary effectors of disease, connecting the epigenome, proteome, and time to disease onset may help to resolve 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 factors (***McCartney et al., 2018c***; ***McCartney et al., 2018a***; ***Peters et al., 2021***), and predict circulating levels of inflammatory proteins (***Stevenson et al., 2020***; ***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 independently of smoking (***Hillary et al., 2020a***; ***Lu et al., 2019***). This suggests there is predictive value gained in utilising DNAm scores relevant to protein levels as intermediaries for predictions. Methylation scores also point towards the pathways that may act on health beyond the protein biomarker that they are trained on. A portfolio of methylation scores for proteins across the circulating proteome could therefore aid in the prediction of disease and offer a different, but additive signal beyond methylation or protein data alone. Generation of an extensive range of epigenetic scores for protein levels 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 (IL-6) and C-reactive protein (CRP) 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 (***Stevenson et al., 2021***; ***Conole et al., 2021***). This is likely due to DNAm representing the accumulation of more sustained effects over a longer period of time than protein measurements, which have often been shown to be variable in their levels when measured at multiple time points (***Koenig et al., 2003***; ***Liu et al., 2015***; ***Moldoveanu et al., 2000***; ***Shah et al., 2014***). DNAm scores for proteins could therefore be used to alert clinicians to individuals with high-risk biological signatures, many years prior to disease onset.

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 706 to 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. We then 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

eLife digest Although our genetic code does not change throughout our lives, our genes can be turned on and off as a result of epigenetics. Epigenetics can track how the environment and even certain behaviors add or remove small chemical markers to the DNA that makes up the genome. The type and location of these markers may affect whether genes are active or silent, this is, whether the protein coded for by that gene is being produced or not. One common epigenetic marker is known as DNA methylation. DNA methylation has been linked to the levels of a range of proteins in our cells and the risk people have of developing chronic diseases.

Blood samples can be used to determine the epigenetic markers a person has on their genome and to study the abundance of many proteins. Gadd, Hillary, McCartney, Zaghlool et al. studied the relationships between DNA methylation and the abundance of 953 different proteins in blood samples from individuals in the German KORA cohort and the Scottish Lothian Birth Cohort 1936. They then used machine learning to analyze the relationship between epigenetic markers found in people's blood and the abundance of proteins, obtaining epigenetic scores or 'EpiScores' for each protein. They found 109 proteins for which DNA methylation patterns explained between at least 1% and up to 58% of the variation in protein levels.

Integrating the 'EpiScores' with 14 years of medical records for more than 9000 individuals from the Generation Scotland study revealed 130 connections between EpiScores for proteins and a future diagnosis of common adverse health outcomes. These included diabetes, stroke, depression, various cancers, and inflammatory conditions such as rheumatoid arthritis and inflammatory bowel disease.

Age-related chronic diseases are a growing issue worldwide and place pressure on healthcare systems. They also severely reduce quality of life for individuals over many years. This work shows how epigenetic scores based on protein levels in the blood could predict a person's risk of several of these diseases. In the case of type 2 diabetes, the EpiScore results replicated previous research linking protein levels in the blood to future diagnosis of diabetes. Protein EpiScores could therefore allow researchers to identify people with the highest risk of disease, making it possible to intervene early and prevent these people from developing chronic conditions as they age.

up to 14 years in the Generation Scotland cohort (n = 9537). We also tested for associations between EpiScore levels and COVID-19 disease outcomes. We regressed out the effects of age on protein levels prior to training and testing; age was also included as a covariate in disease prediction models. We controlled for common risk factors for disease and assessed the capacity of EpiScores to identify previously reported protein-disease associations.

Our MethylDetectR shiny app (***Hillary and Marioni, 2020***) has CpG weights for the 109 EpiScores integrated such that it automates the process of score generation for any DNAm dataset and is available at: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr. A video on how to use the MethylDetectR shiny app to generate EpiScores is available at: https://youtu.be/65Y2Rv-4tPU.

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 706 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 et al., 2020b***; ***Suhre et al., 2017***)

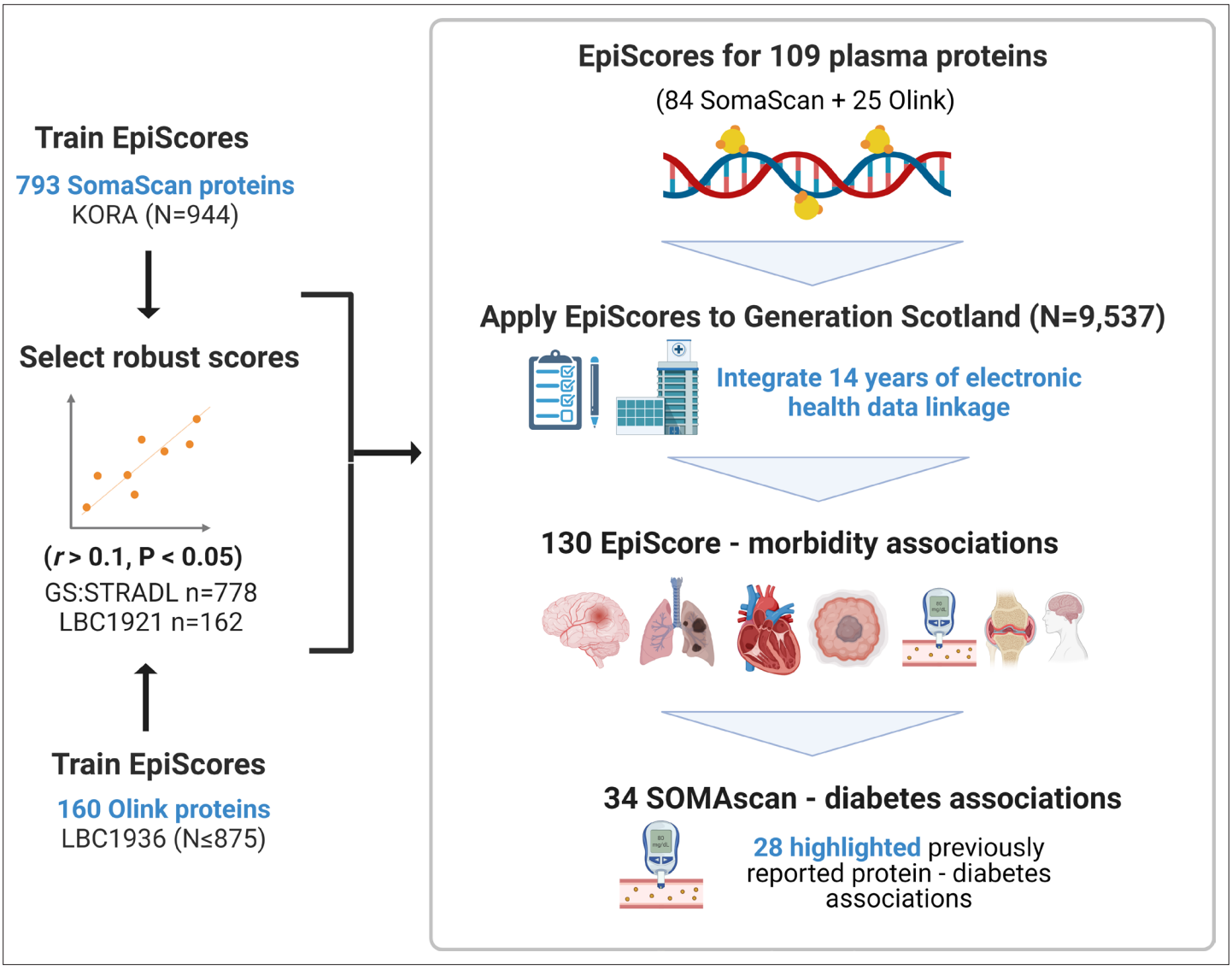

Figure 1. EpiScores for plasma proteins as tools for disease prediction study design. DNA methylation scores were trained on 953 circulating plasma protein levels in the KORA and LBC1936 cohorts. There were 109 EpiScores selected based on performance ($r > 0.1$, $p < 0.05$) in independent test sets. 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; 130 EpiScore-disease associations were identified, for 10 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 34 SOMAscan-derived EpiScore-diabetes associations, 28 highlighted previously reported protein-diabetes associations.

(Materials and 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 all selected predictors on the full training samples.

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

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 = 9537). 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.

	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	63	9289	5.6 (3.5)	52	7742	6.1 (3.3)
Alzheimer's dementia	69	3764	7.7 (3)	52	3137	7.6 (3.1)
Bowel cancer	78	9398	6.4 (3.2)	66	7817	6.5 (3.2)
Depression	95	8317	4 (3.2)	75	6984	3.8 (3.2)
Lung cancer	100	9433	5.6 (3.2)	78	7850	5.6 (3.1)
Breast cancer	131	5356	6.1 (3.4)	111	4402	5.9 (3.4)
Inflammatory bowel disease	194	9114	5 (3.6)	155	7592	4.8 (3.6)
Stroke	313	9026	6.4 (3.4)	246	7547	6.3 (3.5)
COPD	322	8960	5.5 (3.4)	253	7476	5.5 (3.5)
Ischaemic heart disease	385	8649	5.6 (3.4)	302	7251	5.7 (3.4)
Diabetes	429	8757	5.6 (3.4)	322	7332	5.5 (3.4)
Pain	1329	5480	4.8 (3.5)	1081	4593	4.9 (3.5)

COPD: chronic obstructive pulmonary disease.

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 1 (lyzozyme) to 395 (aminoacylase-1 [ACY-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., 2018***). GeneSet enrichment analysis of the original proteins used to train the 109 EpiScores highlighted pathways associated with immune response and cell remodelling, adhesion, and extracellular matrix function (***Supplementary file 1H***).

EpiScore-disease associations in Generation Scotland

The Generation Scotland dataset contains extensive electronic health data from GP and hospital records as well as DNAm data for 9537 individuals. This makes it uniquely positioned to test whether EpiScore signals can predict disease onset. We ran nested mixed effects Cox proportional hazards models (***Figure 3***) to determine whether the levels of each EpiScore at baseline associated with the incidence of 12 morbidities over a maximum of 14 years of follow-up. The correlation structures for the 109 EpiScore measures used for Cox modelling are presented in ***Figure 2—figure supplement 1***.

There were 286 EpiScore-disease associations with a false discovery rate (FDR)-adjusted $p < 0.05$ in the basic model. After further adjustment for common risk factor covariates (smoking, social deprivation status, educational attainment, body mass index [BMI], and alcohol consumption), 130 of the 286 EpiScore-disease associations from the basic model had $p < 0.05$ in the fully adjusted model (***Supplementary file 1I-J***). Ten of the 130 fully adjusted associations failed the Cox proportional hazards assumption for the EpiScore variable ($p < 0.05$ for the association between the Schoenfeld residuals and time; ***Supplementary file 1K***). When we restricted the time-to-event/censor period by each year

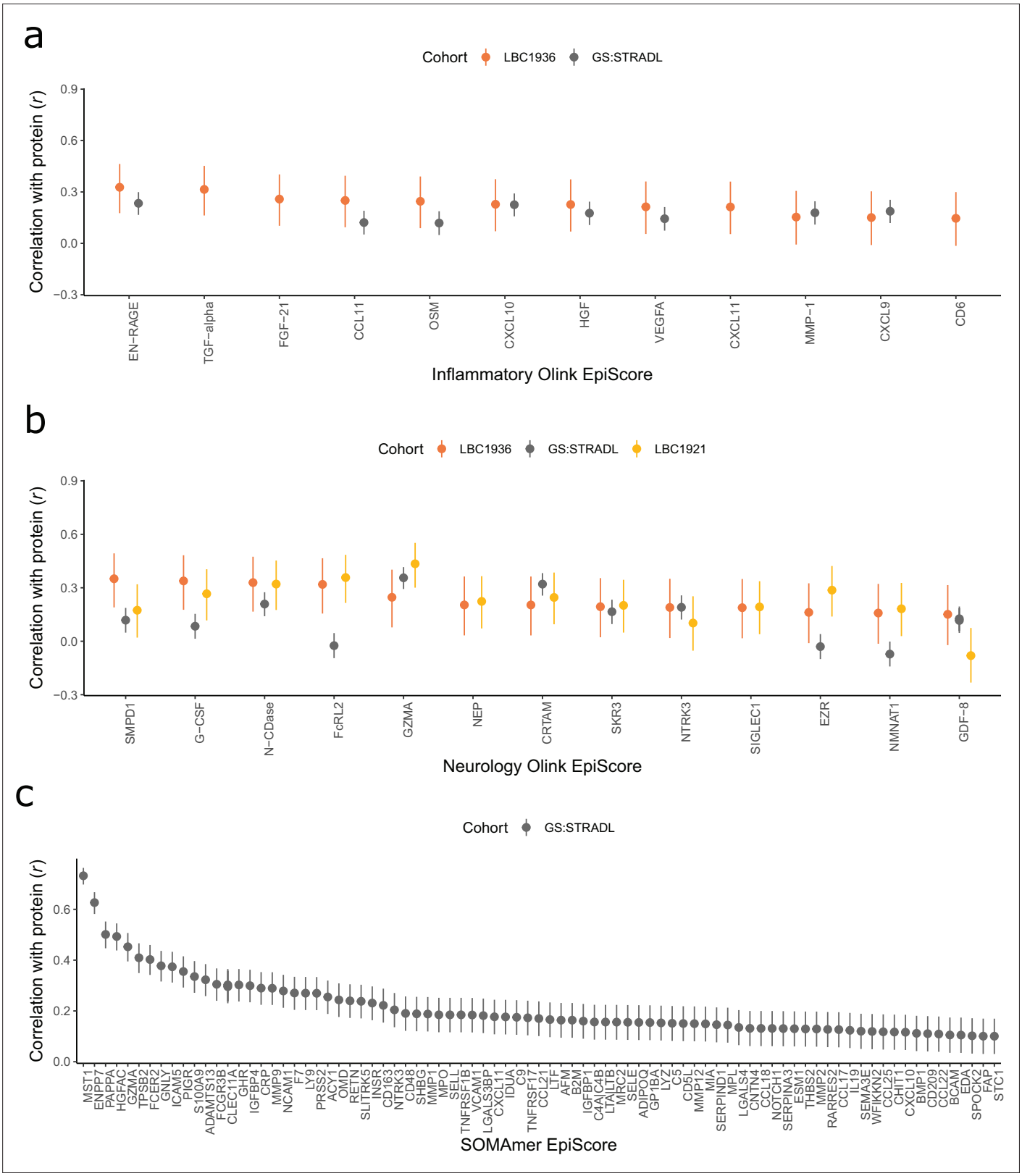

Figure 2. Test performance for the 109 selected protein EpiScores. Test set correlation coefficients for associations between protein EpiScores for (**a**) inflammatory Olink, (**b**) neurology Olink, and (**c**) SOMAmer protein panel EpiScores and measured protein levels are plotted. 95% confidence intervals are shown for each correlation. The 109 protein EpiScores shown had $r > 0.1$ and $p < 0.05$ in 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 file***

Figure 2 continued on next page

Figure 2 continued

1B-C. Correlation heatmaps between the 109 EpiScore measures (***Figure 2—figure supplement 1***) are provided, along with a summary of the most enriched functional pathways for the genes of the 109 proteins used to train EpiScores (***Figure 2—figure supplement 2***).

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Correlation heatmap for protein EpiScore measures in Generation Scotland.

Figure supplement 2. GeneSet enrichment of canonical pathways common to the genes encoding proteins that were used to train the 109 selected EpiScores.

of possible follow-up, there were minimal differences in the EpiScore-disease hazard ratios between follow-up periods that did not violate the assumption and those that did (***Supplementary file 1L***). The 130 associations were therefore retained as the primary results.

The 130 associations found in the fully adjusted model comprised 70 unique EpiScores that were related to the incidence of 10 of the 12 morbidities studied. Diabetes and chronic obstructive pulmonary disease (COPD) had the greatest number of associations, with 38 and 37, respectively. ***Figure 4*** presents the EpiScore-disease relationships for COPD and the remaining nine morbidities: stroke, lung cancer, ischaemic heart disease (IHD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), depression, bowel cancer and pain (back/neck). There were 16 EpiScores that associated with the onset of three or more morbidities. ***Figure 5*** presents relationships for these 16 EpiScores in the fully adjusted Cox model results. Of note is the EpiScore for Complement 5 (C5), which associated with four outcomes: stroke, diabetes, RA and COPD. Of the 34 SOMAscan-derived EpiScore associations with incident diabetes, 28 replicated previously reported SOMAscan protein associations (***Elhadad et al., 2020***; ***Gudmundsdottir et al., 2020***; ***Ngo et al., 2021***) with incident or prevalent diabetes in one or more cohorts (***Figure 6*** and ***Supplementary file 1M***).

Immune cell and GrimAge sensitivity analyses

Correlations of the 70 EpiScores that were associated with incident disease ($P < 0.05$ in the fully-adjusted cox proportional hazards models) with covariates suggested interlinked relationships with both estimated white blood cell proportions and GrimAge acceleration (***Figure 3—figure supplement 1***). These covariates were therefore added incrementally to the fully-adjusted Cox models (***Figure 3***). There were 99 associations that remained statistically significant (FDR $p < 0.05$ in the basic model

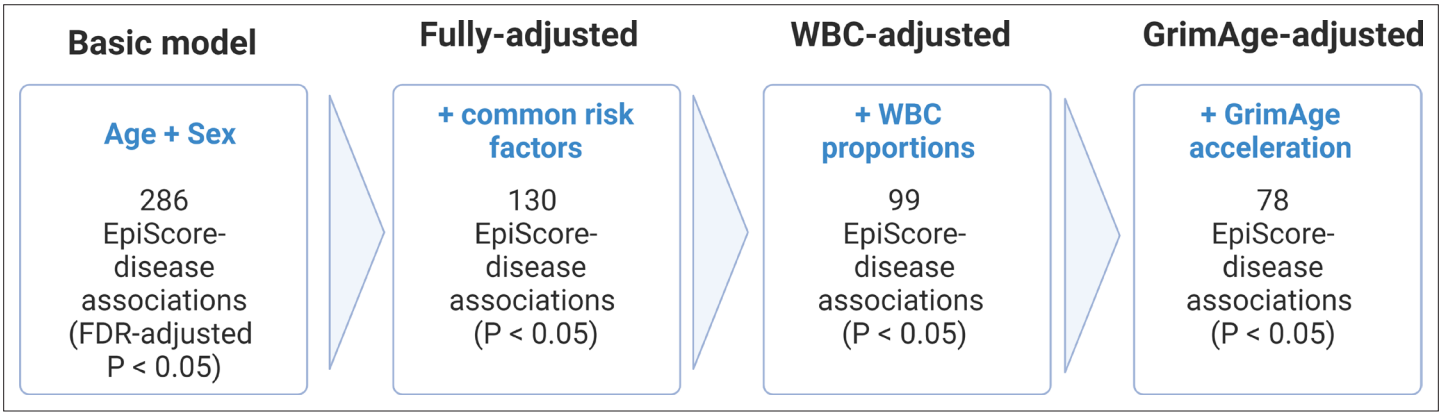

Figure 3. Nested Cox proportional hazards assessment of protein EpiScore-disease prediction. Mixed effects Cox proportional hazards analyses in Generation Scotland (n = 9537) tested the relationships between each of the 109 selected EpiScores and the incidence of 12 leading causes of morbidity (***Supplementary file 1I-J***). The basic model was adjusted for age and sex and yielded 286 associations between EpiScores and disease diagnoses, with false discovery rate (FDR)-adjusted $p < 0.05$. In the fully adjusted model, which included common risk factors as additional covariates (smoking, deprivation, educational attainment, body mass index (BMI), and alcohol consumption), 130 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 31 of the 130 associations. In a further sensitivity analysis, 78 associations remained after adjustment for both immune cell proportions and GrimAge acceleration.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Phenotypic trait and estimated white blood cell proportion correlations with EpiScores.

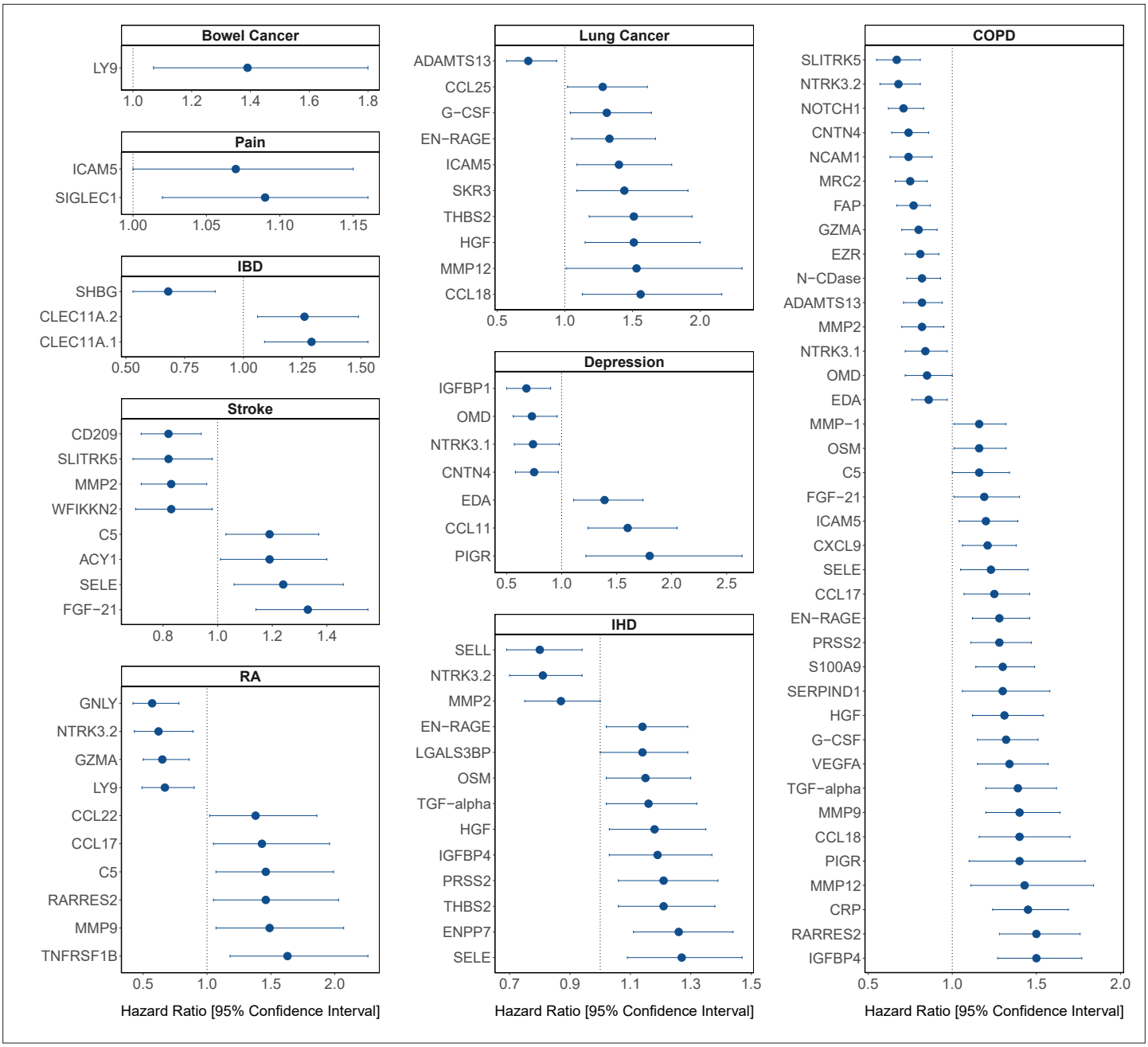

Figure 4. Protein EpiScore associations with incident disease. EpiScore-disease associations for 9 of the 11 morbidities with associations where $p < 0.05$ in the fully adjusted mixed effects Cox proportional hazards models in Generation Scotland (n = 9537). Hazard ratios are presented with confidence intervals for 92 of the 130 EpiScore-incident disease associations reported. Models were adjusted for age, sex, and common risk factors (smoking, body mass index (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 1J***.

and $p < 0.05$ in the fully adjusted model) after adjustment for immune cell proportions, of which 78 remained significant when GrimAge acceleration scores were added to this model (***Supplementary file 1J***). In a further sensitivity analysis, relationships between both estimated white blood cell (WBC) proportions and GrimAge acceleration scores with incident diseases were assessed in the Cox model structure independently of EpiScores. Of the 60 possible relationships between WBC measures and the morbidities assessed, three were statistically significant (FDR-adjusted $p < 0.05$) in the basic model and remained significant with $p < 0.05$ in the fully adjusted model (***Supplementary file 1N***). A higher

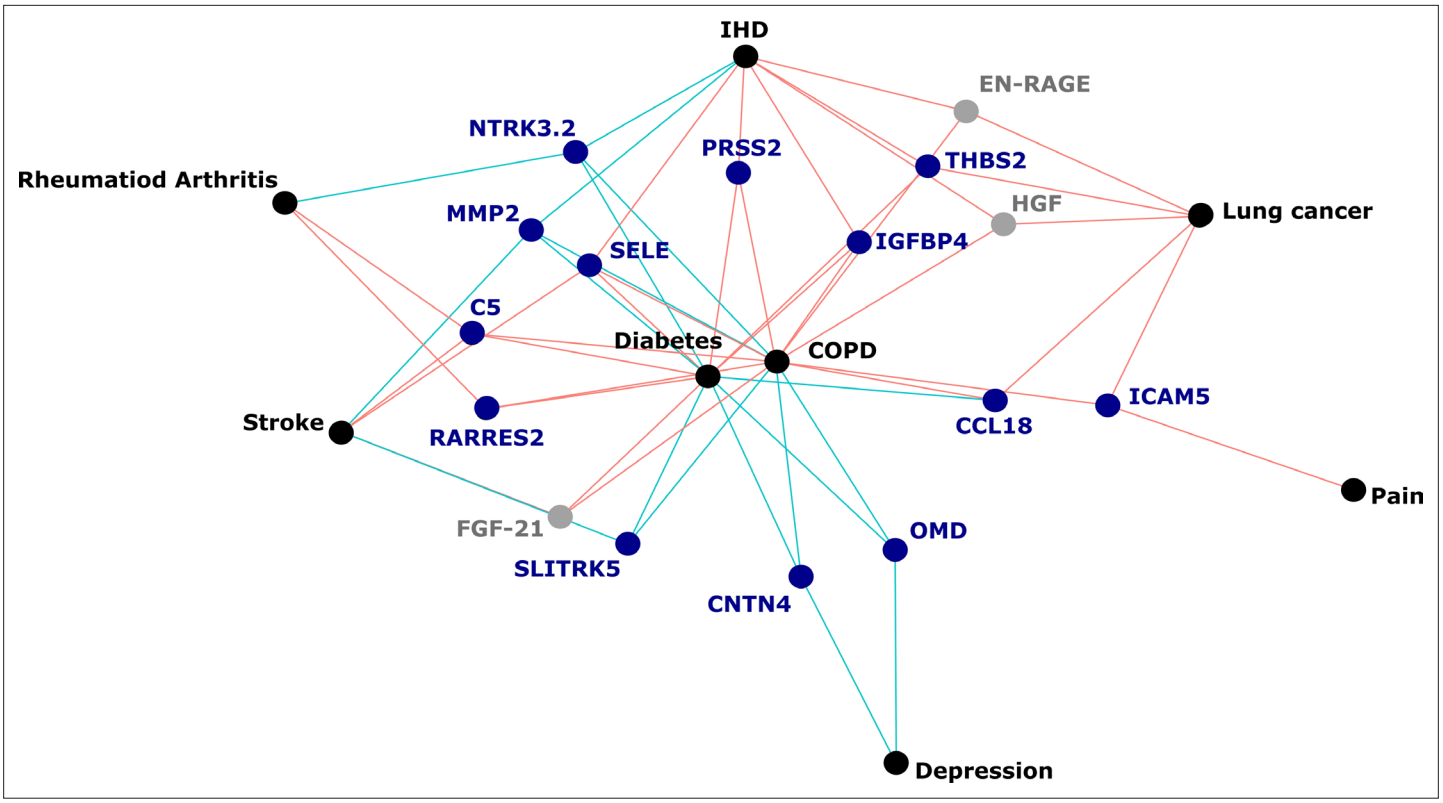

Figure 5. Protein EpiScores that associated with the greatest number of morbidities. EpiScores with a minimum of three relationships with incident morbidities in the fully adjusted Cox models. The network includes 16 EpiScores as dark blue (SOMAscan) and grey (Olink) nodes, with disease outcomes in black. 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 1J***.

proportion of natural killer cells was linked to decreased risk of incident COPD, RA and diabetes. The GrimAge acceleration composite score was associated with COPD, lung cancer, IBD, diabetes and RA in the fully adjusted models ($p < 0.05$) (***Supplementary file 1O***). The magnitude of the GrimAge effect sizes was comparable to the EpiScore findings.

Relationship between EpiScores and subsequent COVID-19

Two previous studies including pilot proteomic measurements from the Generation Scotland cohort (N = 199 controls) as part of wider analyses found that several proteins corresponding to our EpiScores were associated with COVID-19 outcomes (***Demichev et al., 2021***; ***Messner et al., 2020***). These included proteins such as CRP, C9, SELL, and SHBG, all of which were associated with one or more incident diseases in this study. Two subsets (N = 268 and N = 173) of the Generation Scotland sample who contracted COVID-19 were therefore used to test the hypothesis that EpiScores would associate with COVID-19 outcomes (acquired >9 years after the blood draw for DNAm analyses). No significant associations were identified that delineated differences between the development of long-covid (duration >4 weeks) or hospitalisation from COVID-19 (associations that had $p < 0.05$ did not withstand Bonferroni adjustment for multiple testing) (***Supplementary file 1P***).

Discussion

Here, we report a comprehensive DNAm 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 70 EpiScores associate with the incidence of 10 leading causes of morbidity (130 EpiScore-disease associations in total), but do not associate with COVID-19 outcomes. Finally, we show that EpiScore-diabetes associations highlight previously measured protein-diabetes relationships. The bulk of EpiScore-disease associations are independent of common

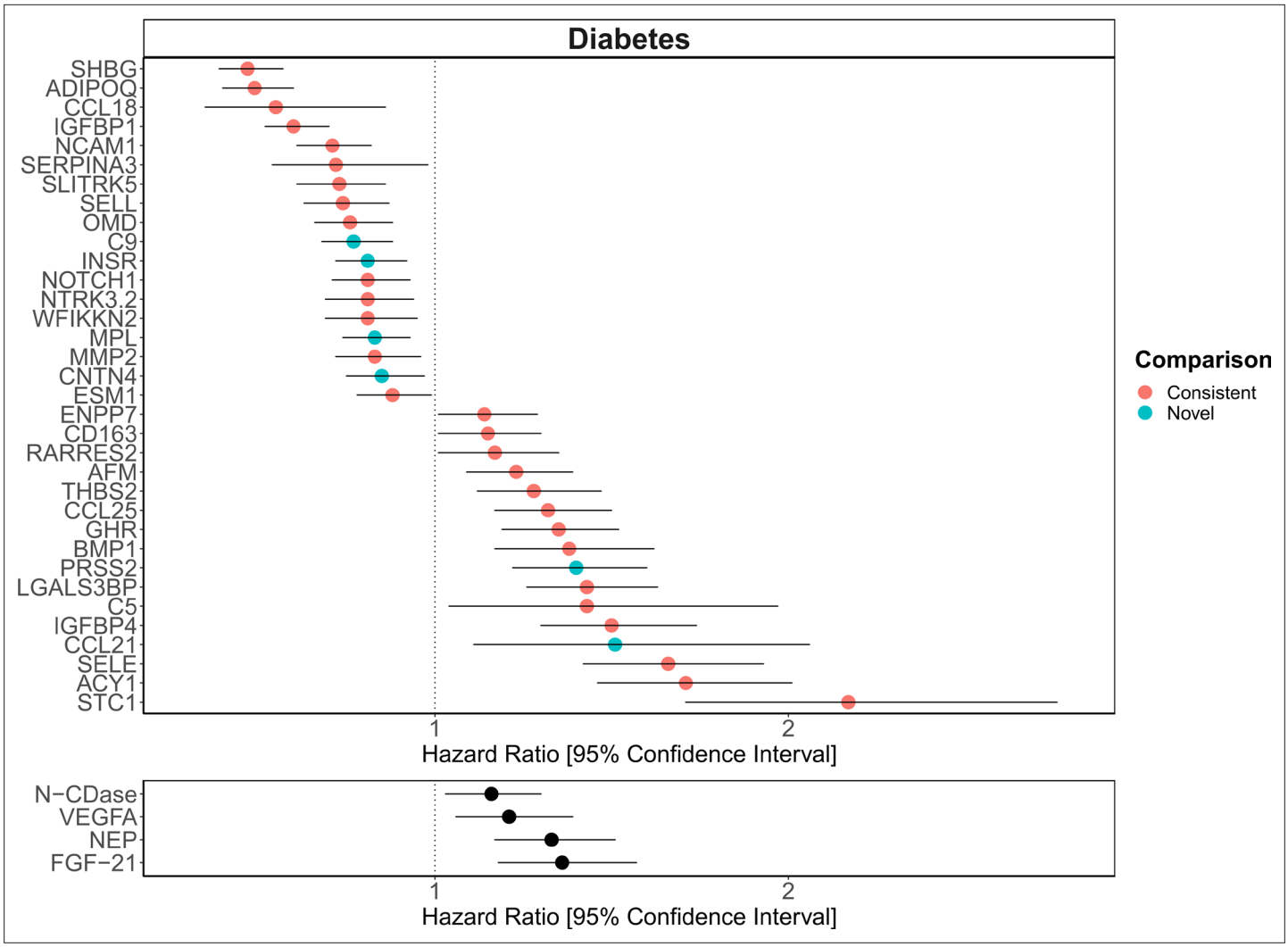

Figure 6. Replication of known protein-diabetes associations with protein EpiScores. EpiScore-incident diabetes associations in Generation Scotland (n = 9537). The 34 SOMAscan (top panel) and four Olink (bottom panel) associations shown with $p < 0.05$ in fully adjusted mixed effects Cox proportional hazards models. Of the 34 SOMAscan-derived EpiScores, 28 associations were consistent with protein-diabetes associations (pink) in one or more of the comparison studies that used SOMAscan protein levels. Six associations were novel (blue). Data shown corresponds to the results included in ***Supplementary files 1J and M***.

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***; ***Ngo et al., 2021***) suggests that epigenetic scores identify disease-relevant biological signals. 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 three studies used for the diabetes comparison represent the largest candidate protein characterisations of type 2 diabetes to date and the top markers identified included aminoacylase-1 (ACY-1), sex hormone-binding globulin (SHBG) and growth hormone receptor (GHR) (***Elhadad et al., 2020***; ***Gudmundsdottir et al., 2020***; ***Ngo et al., 2021***). Our EpiScores for these top markers were 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 and Park, 2016***) and proteins such as ACY-1 and GHR are thought to influence a range of diabetes-associated metabolic mechanisms (***Kim and Park, 2017***; ***Pérez-Pérez et al., 2012***). Proteins that we identify through EpiScore associations, such as NTR domain-containing protein 2 (WFIKKN2), have also been causally implicated in type 2 diabetes onset

through Mendelian randomisation analysis (**Ngo et al., 2021**). 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 (e.g., 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, DNAm scores for CRP and IL-6 have previously been shown to perform modestly in test sets (r ~ 0.2, equivalent to ~4% explained variance in protein level), but augment and often outperform the measured protein related to a range of phenotypes (**Stevenson et al., 2020**; **Stevenson et al., 2021**). Compared to scores utilising DNAm for the prediction of singular diseases, our EpiScores enable the granular study of individual protein predictor signatures with clinical outcomes.

Our large-scale assessment of EpiScores provides a platform for future studies, as composite predictors for traits may be created using our EpiScore database. These should be tested in incident disease predictions when sufficient case data are available. Our results indicated that the set of 109 EpiScores are likely to be heavily enriched for inflammatory, complement system and innate immune system pathways, in addition to extracellular matrix, cell remodelling, and cell adhesion pathways. This reinforces previous work linking chronic inflammation and the epigenome (**Zaghlool et al., 2020**). It also suggests that EpiScores could be useful in the prediction of morbidities that are characterised by differential inflammatory states. An example of this is the EpiScore for Complement Component 5 (C5), which was associated with the onset of four morbidities (**Figure 5**). The EpiScore for C5 is likely to reflect the biological pathways occurring in individuals with heightened complement cascade activity and could be utilised to alert clinicians to individuals at high risk of multimorbidity. Elevated levels of C5 peptides have been associated with severe inflammatory, autoimmune, and neurodegenerative states (**Ma et al., 2019**; **Mantovani et al., 2014**; **Morgan and Harris, 2015**) and a range of C5-targetting therapeutic approaches are in development (**Alawieh et al., 2018**; **Brandolini et al., 2019**; **Hawksworth et al., 2017**; **Hernandez et al., 2017**; **Morgan and Harris, 2015**; **Ort et al., 2020**).

Though EpiScores such as C5 – which occupy central hubs in the disease prediction framework – may provide evidence of early methylation signatures common to the onset of multiple diseases, we did not observe associations between EpiScores and COVID-19 hospitalisation or long-COVID status. This is perhaps surprising, given that many of the morbidities that our EpiScores predicted are also known risk factors for increased risk of death due to COVID-19 (**Williamson et al., 2020**). Many of the proteins corresponding to EpiScores in our study were also associated with COVID-19 severity and progression in two previous studies that included a pilot sample (N = 199) from the Generation Scotland cohort at baseline as control data (**Demichev et al., 2021**; **Messner et al., 2020**). COVID-19 likely has multiple intersecting risk factors that impact severity and recovery, and the lack of associations we observe is likely to be in part due to the limited number of COVID-19 cases available in Generation Scotland. Additionally, there is a large lag time between baseline biological measurement and COVID-19 in our analyses, whereas the two studies that found protein marker associations integrated protein measurements longitudinally and from samples extracted during COVID-19 progression. With increased power available through continued data linkage, EpiScore relationships with COVID-19 outcomes may be observed in future work.

This study has several limitations. First, we demonstrate that EpiScores carry disease-relevant signals that may be clinically meaningful to delineate early disease risk when comparing relative differences within a cohort. However, projecting a new individual onto a reference set is complicated due to absolute differences in methylation quantification resulting from batch and processing effects. Second, future studies should assess paired protein and EpiScore contributions to traits, as inference from EpiScores alone, while useful for disease risk stratification, is not sufficient to determine mechanisms. This may also highlight EpiScores that outperform the measured protein equivalent in disease. Third, the epitope nature of the protein measurement in the SOMAscan panel may incur probe cross-reactivity and non-specific binding; there may also be differences in how certain proteins are measured across panels (**Pietzner et al., 2020**; **Sun et al., 2018**). Comparisons of multiple protein measurement technologies on the same samples would help to explore this in more detail. Fourth, there may be pQTLs with small effect sizes that were not regressed from the proteins prior to generating the EpiScores. Fifth, while training and testing was performed across multiple cohorts, it is likely that further development of EpiScores in larger proteomic samples with diverse ancestries will improve power to generate robust scores. Upper bounds for DNAm prediction of complex traits,

such as proteins, can be estimated by variance components analyses (***Hillary et al., 2020b***; ***Trejo Banos et al., 2020***; ***Zhang et al., 2019***). Finally, associations present between EpiScore measures and disease incidence may have been influenced by external factors such as prescription medications for comorbid conditions and comorbid disease prevalence.

We have shown that EpiScores for circulating protein levels predict the incidence of multiple diseases, up to 14 years prior to diagnosis. Our findings suggest that DNAm phenotyping approaches and data linkage to electronic health records in large, population-based studies have the potential to (1) capture inter-individual variability in protein levels; (2) predict incident disease risk many years prior to morbidity onset; and (3) highlight disease-relevant biological signals for further exploration. The EpiScore weights are publicly available, enabling any cohort with Illumina DNAm data to generate them and to relate them to various outcomes. Given the increasingly widespread assessment of DNAm in cohort studies (***McCartney et al., 2020***; ***Min et al., 2021***), EpiScores offer an affordable and consistent (i.e., array-based) way to utilise these signatures. This information is likely to be important in risk stratification and prevention of age-related morbidities.

Materials and methods

The KORA sample population

The KORA F4 study includes 3080 participants who reside in Southern Germany. Individuals 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 DNAm 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.

DNAm in KORA

Methylation data were generated for 1814 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 normalisation 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 (Version 3.2) (***Gold et al., 2010***) was used to quantify protein levels in undepleted plasma for 1129 SOMAmer probes (***Suhre et al., 2017***). Of the 1000 samples provided for analysis, two samples were excluded due to errors in bio-bank sampling and one based on quality control (QC) measures. Of the 997 samples available, there were 944 individuals with methylation and genotypic data. Of the 1129 probes available, five failed the QC, leaving a total of 1124 probes for the subsequent analysis. Protein measurements were transformed by rank-based inverse normalisation and regressed onto age, sex, known pQTLs, and 20 genetic principal components of ancestry derived from the Affymetrix Axiom Array to control for population structure. pQTLs for each protein were taken from a previous GWAS in the sample (***Suhre et al., 2017***).

The LBC1936 and LBC1921 sample populations

The Lothian Birth Cohorts of 1921 (LBC1921; N = 550) and 1936 (LBC1936; N = 1091) are longitudinal studies of aging in individuals who reside in Scotland (***Deary et al., 2012***; ***Taylor et al., 2018***). Participants completed an intelligence test at age 11 years and were recruited for these cohorts at mean ages of 79 (LBC1921) and 70 (LBC1936). They have been followed up triennially for a series of cognitive, clinical, physical, and social data, along with blood donations that have been used for genetic, epigenetic, and proteomic measurement. DNAm, proteomic (Olink platform), and genetic data for individuals from Waves 1 (n=875 at mean age 70 years and sd 0.8) and 2 (n=706 at mean age 73 years and sd 0.7) of the LBC1936 and Wave 3 of the LBC1921 (n=162 at mean age 87 years and sd 0.4) were available.

DNAm in LBC1936 and LBC1921

DNA from whole blood was assessed using the Illumina 450 K methylation array. Details of QC have been described elsewhere (***Shah et al., 2014***; ***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 DNAm-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 QC 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 et al., 2020b***).

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., 2013***). 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 9537 individuals with DNAm and phenotypic information available. The STRADL cohort is a subset of 1188 individuals from the GS cohort who undertook additional assessments approximately 5 years after the study baseline (***Navrady et al., 2018***).

DNAm in Generation Scotland and STRADL

In the GS cohort, blood-based DNAm was generated in two sets using the Illumina EPIC array. Set 1 comprised 5190 related individuals whereas Set 2 comprised 4583 individuals, unrelated to each other and to those in Set 1. During QC, 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 et al., 2018b***; ***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 9537 individuals ($n_{Set1} = 5087$, $n_{Set2} = 4450$). The same steps were also applied to process DNAm in STRADL.

Proteomics in STRADL

Measurements for 4235 proteins in 1065 individuals from the STRADL cohort were recorded using the SOMAscan technology; 793 epitopes matched between the KORA and STRADL cohorts and were included for training in KORA and testing in STRADL. There were 778 individuals with proteomics data and DNAm data in STRADL. Protein measurements were transformed by rank-based inverse normalisation and regressed onto age, sex, and 20 genetic principal components (derived from multi-dimensional scaling of genotype data from the Illumina 610-Quadv1 array).

Electronic health data linkage in Generation Scotland

Over 98% of GS participants consented to allow access to electronic health records via data linkage to GP records (Read 2 codes) and hospital records (ICD codes). Data are available prospectively from the time of blood draw, yielding up to 14 years of linkage. We considered incident data for 12 morbidities. Ten of the diseases are listed by the World Health Organization (WHO) as leading causes of either morbidity or mortality (***Hay et al., 2017***; ***World Health Organization, 2018***). Inflammatory bowel disease (IBD) (***Kassam et al., 2014***) and RA (***James et al., 2018***) are also included as traits

as they have been reported as leading causes of disability and morbidity and the global burdens of these diseases are rising (**Alatab et al., 2020**; **Safiri et al., 2019**). Prevalent cases (ascertained via retrospective ICD and Read 2 codes or self-report from a baseline questionnaire) were excluded. For IBD prevalent cases were excluded based on data linkage alone. Included and excluded terms can be found in **Supplementary files 1Q-1B1**. Alzheimer's dementia was limited to cases/controls with age of event/censoring ≥65 years. Breast cancer was restricted to females only. Recurrent, major and moderate episodes of depression were included in depression. Diabetes was comprised of predominantly type 2 diabetes codes and additional general diabetes codes such as diabetic retinopathy and diabetes mellitus with renal manifestation that often occur in individuals with type 2 diabetes. Type 1 and juvenile diabetes cases were excluded.

Elastic net protein EpiScores

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

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

Functional annotations for each of the proteins used to train the finalised set of 109 EpiScores were sourced from the STRING database (**Jensen et al., 2009**). GeneSet enrichment analysis against protein-coding genes was performed using the FUMA database, to quantify which canonical pathways were most commonly implicated across the 109 genes corresponding to the proteins used to train the 109 EpiScores (**Watanabe et al., 2017**). The background gene-set was specified as protein coding genes and a threshold of FDR $p < 0.05$ was applied for enrichment status, with the minimum overlapping genes with gene-sets set to ≥2.

The 109 selected EpiScores were then applied to Generation Scotland (n = 9537). DNAm at each CpG site was scaled to have a mean of zero and variance of one, with scaling performed separately for GS sets.

Associations with health linkage phenotypes in Generation Scotland

Mixed effects Cox proportional hazards regression models adjusting for age, sex, and methylation set were used to assess the relationship between 109 EpiScores and 12 morbidities in Generation Scotland. Models were run using coxme (**Therneau, 2020b**) (Version 2.2-16) with a kinship matrix accounting for relatedness in Set 1. Cases included those diagnosed after baseline who had died, in addition to those who received a diagnosis and remained alive. Controls were censored if they were disease free at time of death, or at the end of the follow-up period. EpiScore levels were rank-base inverse normalised. Fully adjusted models included the following additional covariates measured at baseline: alcohol consumption (units consumed in the previous week); deprivation assessed by the Scottish Index of Multiple Deprivation (**GovScot, 2016**); BMI (kg/m^2); educational attainment (an 11-category ordinal variable); and a DNAm-based score for smoking status (**Bollepalli et al., 2019**). A false discovery rate multiple testing correction $p < 0.05$ was applied to the 1308 EpiScore-disease associations (109 EpiScores by 12 incident disease traits). Proportional hazards assumptions were checked through Schoenfeld residuals (global test and a test for the protein-EpiScore variable) using the coxph and cox.zph functions from the survival package (**Therneau, 2020a**) (Version 3.2-7). For each association failing to meet the assumption (Schoenfeld residuals $p < 0.05$), a sensitivity analysis was run across yearly follow-up intervals.

Fully adjusted Cox proportional hazards models were run with Houseman-estimated white blood cell proportions as covariates (**Houseman et al., 2012**). A further sensitivity analysis added GrimAge

acceleration (***Lu et al., 2019***) as an additional covariate. Basic and fully adjusted Cox models were also run with estimated monocyte, B-cell, CD4T, CD8T, and natural killer cell proportions as predictors, in addition to models with GrimAge acceleration as the predictor of incident disease.

Correlation structures for EpiScores, DNAm-estimated white blood cell proportions, and phenotypic information were assessed using Pearson correlations and pheatmap (***Kolde, 2019***) (Version 1.0.12) and ggcorrplot packages (Version 0.1.3) (***Kassambara, 2019***). The psych package (Version 2.0.9) (***Revelle, 2020***) was used to perform principal components analysis on EpiScores. ***Figures 1 and 2*** were created with BioRender.com. Associations for EpiScores that were related to a minimum of three morbidities were subset from the fully adjusted Cox proportional hazards results and were visualised using the ggraph package (Version 2.0.5) (***Pedersen, 2021***). This network representation was used (***Figure 5***) to highlight protein EpiScores that were connected with multiple morbidities.

Consistency of disease associations between EpiScores and measured proteins

Comparisons were conducted between EpiScore-diabetes associations and type 2 diabetes associations with measured proteins using three previous large-scale proteomic studies (***Elhadad et al., 2020***; ***Gudmundsdottir et al., 2020***; ***Ngo et al., 2021***) In these studies, six cohorts were included (Study 1: KORA n = 993, HUNT n = 940 [***Elhadad et al., 2020***], Study 2: AGES-Reykjavik n = 5438 and QMDiab n = 356 [***Gudmundsdottir et al., 2020***], Study 3: Framingham Heart Study n = 1618 and the Malmo Diet and Cancer Study n = 1221). Study 1 included the KORA dataset, which we use in this study to generate SOMAscan EpiScores. We characterised which SOMAscan-based EpiScore-diabetes associations from our fully adjusted results reflected those observed with measured protein levels. We included basic (nominal $p < 0.05$) and fully adjusted results (with either FDR or Bonferroni-corrected $p < 0.05$), wherever available, across the lookup cohorts (***Supplementary file 1M***).

Relationship between EpiScores and COVID-19 outcomes

Associations between each of the 109 selected protein EpiScores and subsequent long-COVID or COVID-19 hospitalisation were tested in the Generation Scotland population. A binary variable was used for long-COVID based on self-reported COVID-19 duration from the CovidLife study survey 3 questionnaire (N = 2399 participating individuals) (***Fawns-Ritchie et al., 2021***). Participants were asked about the total overall time they experienced symptoms in their first/only episode of illness, as well as their COVID-19 illness duration. The dataset is correct as of February 2021 when the survey 3 was administered. Of the 9537 individuals with DNAm that were included in incident disease analyses, 173 indicated that they had COVID-19 and 56 of these individuals reported having long-COVID (>4 weeks duration of symptoms after infection). The mean duration from DNAm measurement to long-COVID for these 56 individuals was 11.2 years (sd 1.2). Hospitalisation information, derived from the Scottish Morbidity Records (SMR01), was used to obtain COVID-19 hospital admissions using ICD-10 codes U07.1 (lab-confirmed COVID-19 diagnosis), and U07.2 (clinically diagnosed COVID-19). This data linkage identified 268 of the 9537 individuals that had COVID-19 diagnoses and 29 had been recorded as being hospitalised due to COVID-19. The mean duration from DNAm measurement to hospitalisation for these 29 individuals was 11.9 years (sd 1.4). Logistic regression models with either hospitalisation or long-COVID status as binary outcomes were used, with the 109 scaled protein EpiScores as the independent variables. Sex and age at COVID testing were included as covariates. The latter was defined as the age at positive COVID-19 test or 1 January 2021 if COVID-19 test data were not available.

Acknowledgements

We are grateful to all study participants of KORA, LBC1936, LBC1921, and GS for their invaluable contributions to this study. This research was funded in whole, or in part, by the Wellcome Trust [104036/Z/14/Z, 220857/Z/20/Z, 108890/Z/15/Z, 203771/Z/16/Z, 216767/Z/19/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Additional information

Competing interests

Robert F Hillary: has received consultant fees from Illumina. Riccardo E Marioni: has received speaker fees from Illumina and is an advisor to the Epigenetic Clock Development Foundation. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Wellcome Trust	108890/Z/15/Z	Danni A Gadd Robert F Hillary
Wellcome Trust	203771/Z/16/Z	Anna J Stevenson
Alzheimer's Research UK	ARUK-PG2017B–10	Daniel L McCartney Riccardo E Marioni
Qatar Foundation	Biomedical Research Program at Weill Cornell Medicine	Shaza B Zaghlool Karsten Suhre
Qatar National Research Fund	NPRP11C-0115-180010	Shaza B Zaghlool Karsten Suhre
Bundesministerium für Bildung und Forschung	Helmholtz Zentrum München	Christian Gieger Annette Peters Melanie Waldenberger Johannes Graumann
Munich Center of Health Sciences	LMUinnovativ	Christian Gieger Annette Peters Melanie Waldenberger Johannes Graumann
Bavarian State Ministry of Health and Care	DigiMed Bayern	Christian Gieger Annette Peters Melanie Waldenberger Johannes Graumann
NIHR Biomedical Research Centre, Oxford		Liu Shi
Dementias Platform UK	MR/L023784/2	Liu Shi
Medical Research Council	MC_UU_00007/10	Caroline Hayward
Wellcome Trust	104036/Z/14/Z	Ian J Deary David J Porteous Andrew M McIntosh
Wellcome Trust	220857/Z/20/Z	Andrew M McIntosh
Wellcome Trust	216767/Z/19/Z	Chloe Fawns-Ritchie Cliff Nangle Archie Campbell Robin Flaig Ian J Deary David J Porteous Caroline Hayward Andrew M McIntosh Riccardo E Marioni
Chief Scientist Office of the Scottish Government Health Directorates	CZD/16/6	David J Porteous
Scottish Funding Council	HR03006	David J Porteous
Australian Research Council	Fellowship FT200100837	Allan F McRae

Funder	Grant reference number	Author
Australian Research Council	DP160102400	Peter M Visscher
National Health and Medical Research Council	1113400	Peter M Visscher
Medical Research Council and Biotechnology and Biological Sciences Research Council	MR/K026992/1	Ian J Deary
Biotechnology and Biological Sciences Research Council		Ian J Deary
Royal Society	Wolfson Research Merit Award	Ian J Deary
Chief Scientist Office (CSO) of the Scottish Government's Health Directorates		Ian J Deary
Age UK	(Disconnected Mind project)	Sarah E Harris Ian J Deary Simon R Cox
Medical Research Council	G0701120	Ian J Deary Simon R Cox
Biotechnology and Biological Sciences Research Council	BB/F019394/1	Ian J Deary
Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society	221890/Z/20/Z	Simon R Cox
National Institutes of Health	RF1AG073593	Elliot M Tucker-Drob
National Institutes of Health	R01AG054628	Elliot M Tucker-Drob Ian J Deary Simon R Cox
Health Data Research UK	substantive site award	Archie Campbell
Medical Research Council	MRC Human Genetics Unit core support	Caroline Hayward
Medical Research Council	MR/R024065/1	Ian J Deary Simon R Cox
Medical Research Council	MR/M013111/1	Ian J Deary Simon R Cox
Medical Research Council	G1001245	Ian J Deary Simon R Cox
Australian Research Council	FL180100072	Peter M Visscher
National Health and Medical Research Council	1010374	Peter M Visscher
National Institutes of Health	P30AG066614	Elliot M Tucker-Drob
National Institutes of Health	P2CHD042849	Elliot M Tucker-Drob
Alzheimer's Society	AS-PG-19b-010	Riccardo E Marioni

Funder	Grant reference number	Author
University of Edinburgh and University of Helsinki joint PhD programme in Human Genomics		Yipeng Cheng

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Danni A Gadd, Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review and editing; Robert F Hillary, Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization; Daniel L McCartney, Shaza B Zaghlool, Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation; Anna J Stevenson, Investigation, Methodology; Yipeng Cheng, Chloe Fawns-Ritchie, Data curation, Formal analysis; Cliff Nangle, Archie Campbell, Robin Flaig, Sarah E Harris, Rosie M Walker, Liu Shi, Elliot M Tucker-Drob, Christian Gieger, Annette Peters, Melanie Waldenberger, Johannes Graumann, Allan F McRae, Ian J Deary, David J Porteous, Caroline Hayward, Peter M Visscher, Simon R Cox, Kathryn L Evans, Andrew M McIntosh, Data curation, Investigation; Karsten Suhre, Conceptualization, Data curation, Investigation, Methodology; Riccardo E Marioni, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft, Writing – review and editing

Author ORCIDs

Danni A Gadd http://orcid.org/0000-0001-6398-5407
Cliff Nangle http://orcid.org/0000-0001-5432-1158
Sarah E Harris http://orcid.org/0000-0002-4941-5106
Karsten Suhre http://orcid.org/0000-0001-9638-3912
Riccardo E Marioni http://orcid.org/0000-0003-4430-4260

Ethics

Human subjects: 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.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.71802.sa1
Author response https://doi.org/10.7554/eLife.71802.sa2

Additional files

Supplementary files

- Supplementary file 1. Demographic information and supplementary datasets. (**A**) Demographic and array information for the cohorts and samples used in the study. (**B**) SomaScan panel EpiScore performance in the Stratifying Resilience and Depression Longitudinally (STRADL) test set. (**C**) Performance of Olink panel EpiScores in holdout, STRADL, and LBC1921 test sets. (**D**) Annotations for the proteins corresponding to the 109 selected EpiScores. (**E**) Predictor weights for the 109 EpiScores from Olink and SomaScan panels which passed testing in independent cohorts. (**F**) CpG feature counts for the 109 selected EpiScores. (**G**) Frequency of CpG sites selected for EpiScores with EWAS catalog annotations to phenotypic traits. (**H**) FUMA canonical pathway Gene set

enrichment for the genes encoding the 109 proteins EpiScores were trained on. (**I**) Basic Cox proportional hazards model results in Generation Scotland. (**J**) Fully adjusted and sensitivity analyses results for Cox proportional hazards models in Generation Scotland. (**K**) Schoenfeld residual Cox sensitivity analyses. (**L**) Schoenfeld residual Cox sensitivity analyses split by year of follow-up. (**M**) SOMAscan-EpiScore diabetes association lookup against three large-scale plasma protein-diabetes studies. (**N**) White blood cell sensitivity analyses. (**O**) GrimAge sensitivity analyses. (P) COVID-19 analyses. Q-1B1 Primary and secondary diagnosis codes for each of the 12 morbidities in this study that were used to assign case/control status of participants.

- Transparent reporting form

Data availability

Datasets generated in this study are made available in Supplementary file 1; this file includes the protein EpiScore weights for the 109 EpiScores we provide for future studies to use. Our MethylDetectR shiny app (Hillary and Marioni, 2020) has CpG weights for the 109 EpiScores integrated such that it automates the process of score generation for any DNAm dataset and is available at: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr. A video on how to use the MethylDetectR shiny app to generate EpiScores is available at: https://youtu.be/65Y2Rv-4tPU. All datasets used to create figures are included in Supplementary file 1 and specific locations for these are noted in figure legends. All code used in the analyses is available with open access at the following Gitlab repository: https://github.com/DanniGadd/EpiScores-for-protein-levels (copy archived at swh:1:rev:a5130fab3895a0d95f0dcc8826aa9fb5e8c0fa86). The source datasets analysed during the current study are not publicly available due to them containing information that could compromise participant consent and confidentiality. Data can be obtained from the data owners. Instructions for Lothian Birth Cohort data access can be found here: https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration. Dr Simon Cox must be contacted to obtain a Lothian Birth Cohort 'Data Request Form' by email: simon.cox@ed.ac.uk. Instructions for accessing Generation Scotland data can be found here: https://www.ed.ac.uk/generation-scotland/for-researchers/access; the 'GS Access Request Form' can be downloaded from this site. Completed request forms must be sent to access@generationscotland.org to be approved by the Generation Scotland access committee. Data from the KORA study can be requested from KORA-gen: https://www.helmholtz-munich.de/en/kora/for-scientists/cooperation-with-kora/index.html. Requests are submitted online and are subject to approval by the KORA board.

References

Alatab S, Sepanlou SG, Ikuta K, Vahedi H, Bisignano C, Safiri S, Sadeghi A, Nixon MR, Abdoli A, Abolhassani H, Alipour V, Almadi MAH, Almasi-Hashiani A, Anushiravani A, Arabloo J, Atique S, Awasthi A, Badawi A, Baig AAA, Naghavi M. 2020. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet. Gastroenterology & Hepatology* **5**:17–30. DOI: https://doi.org/10.1016/S2468-1253(19)30333-4, PMID: 31648971

Alawieh A, Langley EF, Tomlinson S. 2018. Targeted complement inhibition salvages stressed neurons and inhibits neuroinflammation after stroke in mice. *Science Translational Medicine* **10**:eaao6459. DOI: https://doi.org/10.1126/scitranslmed.aao6459, PMID: 29769288

Amador C, Huffman J, Trochet H, Campbell A, Porteous D, Wilson JF, Hastie N, Vitart V, Hayward C, Navarro P, Haley CS, Generation Scotland. 2015. Recent genomic heritage in Scotland. *BMC Genomics* **16**:437. DOI: https://doi.org/10.1186/s12864-015-1605-2, PMID: 26048416

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**:1363–1369. DOI: https://doi.org/10.1093/bioinformatics/btu049, PMID: 24478339

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**:1469–1486. DOI: https://doi.org/10.2217/epi-2019-0206, PMID: 31466478

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**:163–174. DOI: https://doi.org/10.1007/s12640-019-00026-w, PMID: 30953275

Conole ELS, Stevenson AJ, Muñoz Maniega S, Harris SE, Green C, Valdés Hernández MDC, Harris MA, Bastin ME, Wardlaw JM, Deary IJ, Miron VE, Whalley HC, Marioni RE, Cox SR. 2021. DNA Methylation and Protein Markers of Chronic Inflammation and Their Associations With Brain and Cognitive Aging. *Neurology* **97**:e2340–e2352. DOI: https://doi.org/10.1212/WNL.0000000000012997

Deary IJ, Gow AJ, Pattie A, Starr JM. 2012. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology* **41**:1576–1584. DOI: https://doi.org/10.1093/ije/dyr197, PMID: 22253310

Demichev V, Tober-Lau P, Lemke O, Nazarenko T, Thibeault C, Whitwell H, Röhl A, Freiwald A, Szyrwiel L, Ludwig D, Correia-Melo C, Aulakh SK, Helbig ET, Stubbemann P, Lippert LJ, Grüning N-M, Blyuss O, Vernardis S, White M, Messner CB, et al. 2021. A time-resolved proteomic and prognostic map of COVID-19. *Cell Systems* **12**:780–794. DOI: https://doi.org/10.1016/j.cels.2021.05.005, PMID: 34139154

Elhadad MA, Jonasson C, Huth C, Wilson R, Gieger C, Matias P, Grallert H, Graumann J, Gailus-Durner V, Rathmann W, von Toerne C, Hauck SM, Koenig W, Sinner MF, Oprea TI, Suhre K, Thorand B, Hveem K, Peters A, Waldenberger M. 2020. Deciphering the Plasma Proteome of Type 2 Diabetes. *Diabetes* **69**:2766–2778. DOI: https://doi.org/10.2337/db20-0296, PMID: 32928870

Fawns-Ritchie C, Altschul DM, Campbell A, Huggins C, Nangle C, Dawson R, Edwards R, Flaig R, Hartley L, Levein C, McCartney DL, Bell D, Douglas E, Deary IJ, Hayward C, Marioni RE, McIntosh AM, Sudlow C, Porteous DJ. 2021. CovidLife: a resource to understand mental health, well-being and behaviour during the COVID-19 pandemic in the UK. *Wellcome Open Research* **6**:176. DOI: https://doi.org/10.12688/wellcomeopenres.16987.1

Friedman J, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**:1–22. DOI: https://doi.org/10.18637/jss.v033.i01, PMID: 20808728

Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JRB, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Tajes JF, Highland HM, et al. 2016. The genetic architecture of type 2 diabetes. *Nature* **536**:41–47. DOI: https://doi.org/10.1038/nature18642, PMID: 27398621

Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, Sterling DG, Williams SA. 2016. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA* **315**:2532–2541. DOI: https://doi.org/10.1001/jama.2016.5951, PMID: 27327800

Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, Carter J, Dalby AB, Eaton BE, Fitzwater T, Flather D, Forbes A, Foreman T, Fowler C, Gawande B, Goss M, Gunn M, Gupta S, Halladay D, Heil J, et al. 2010. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLOS ONE* **5**:e15004. DOI: https://doi.org/10.1371/journal.pone.0015004, PMID: 21165148

GovScot. 2016. Scottish Government. The Scottish Index of Multiple Deprivation (SIMD). Accessed April 2021. http://www.gov.scot/Resource/0050/00504809.pdf [Accessed April 1, 2021].

Gudmundsdottir V, Zaghlool SB, Emilsson V, Aspelund T, Ilkov M, Gudmundsson EF, Jonsson SM, Zilhão NR, Lamb JR, Suhre K, Jennings LL, Gudnason V. 2020. Circulating Protein Signatures and Causal Candidates for Type 2 Diabetes. *Diabetes* **69**:1843–1853. DOI: https://doi.org/10.2337/db19-1070, PMID: 32385057

Hawksworth OA, Li XX, Coulthard LG, Wolvetang EJ, Woodruff TM. 2017. New concepts on the therapeutic control of complement anaphylatoxin receptors. *Molecular Immunology* **89**:36–43. DOI: https://doi.org/10.1016/j.molimm.2017.05.015, PMID: 28576324

Hay SI, Abajobir AA, Abate KH, Abbafati C, Abbas KM, Abd-Allah F, Abdulle AM, Abebo TA, Abera SF, Aboyans V, Abu-Raddad LJ, Ackerman IN, Adedeji IA, Adetokunboh O, Afshin A, Aggarwal R, Agrawal S, Agrawal A, Kiadaliri AA, Bryane CEG. 2017. 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* **390**:1260–1344. DOI: https://doi.org/10.1016/S0140-6736(17)32130-X, PMID: 28919118

Hernandez MX, Jiang S, Cole TA, Chu SH, Fonseca MI, Fang MJ, Hohsfield LA, Torres MD, Green KN, Wetsel RA, Mortazavi A, Tenner AJ. 2017. Prevention of C5aR1 signaling delays microglial inflammatory polarization, favors clearance pathways and suppresses cognitive loss. *Molecular Neurodegeneration* **12**:66. DOI: https://doi.org/10.1186/s13024-017-0210-z, PMID: 28923083

Hillary RF, McCartney DL, Harris SE, Stevenson AJ, Seeboth A, Zhang Q, Liewald DC, Evans KL, Ritchie CW, Tucker-Drob EM, Wray NR, McRae AF, Visscher PM, Deary IJ, Marioni RE. 2019. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. *Nature Communications* **10**:3160. DOI: https://doi.org/10.1038/s41467-019-11177-x, PMID: 31320639

Hillary RF, Marioni RE. 2020. MethylDetectR: a software for methylation-based health profiling. *Wellcome Open Research* **5**:283. DOI: https://doi.org/10.12688/wellcomeopenres.16458.2, PMID: 33969230

Hillary RF, Stevenson AJ, McCartney DL, Campbell A, Walker RM, Howard DM, Ritchie CW, Horvath S, Hayward C, McIntosh AM, Porteous DJ, Deary IJ, Evans KL, Marioni RE. 2020a. Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. *Clinical Epigenetics* **12**:115. DOI: https://doi.org/10.1186/s13148-020-00905-6, PMID: 32736664

Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, Patxot M, Ojavee SE, Zhang Q, Liewald DC, Ritchie CW, Evans KL, Tucker-Drob EM, Wray NR, McRae AF, Visscher PM, Deary IJ, Robinson MR, Marioni RE. 2020b. Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults. *Genome Medicine* **12**:60. DOI: https://doi.org/10.1186/s13073-020-00754-1, PMID: 32641083

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**:86. DOI: https://doi.org/10.1186/1471-2105-13-86, PMID: 22568884

James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, Abbastabar H, Abd-Allah F, Abdela J, Abdelalim A, Abdollahpour I, Abdulkader RS, Abebe Z, Abera SF, Abil OZ, Abraha HN, Abu-Raddad LJ, Abu-Rmeileh NME, Accrombessi MMK, Murray CJL. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**:1789–1858. DOI: https://doi.org/10.1016/S0140-6736(18)32279-7, PMID: 30496104

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research* **37**:D412–D416. DOI: https://doi.org/10.1093/nar/gkn760, PMID: 18940858

Kassam Z, Belga S, Roifman I, Hirota S, Jijon H, Kaplan GG, Ghosh S, Beck PL. 2014. Inflammatory bowel disease cause-specific mortality: a primer for clinicians. *Inflammatory Bowel Diseases* **20**:2483–2492. DOI: https://doi.org/10.1097/MIB.0000000000000173, PMID: 25185685

Kassambara A. 2019. ggcorrplot: Visualization of a Correlation Matrix using "ggplot2. 0.1.3. R Package. https://cran.r-project.org/web/packages/ggcorrplot/ggcorrplot.pdf

Kim SH, Park MJ. 2017. Effects of growth hormone on glucose metabolism and insulin resistance in human. *Annals of Pediatric Endocrinology & Metabolism* **22**:145–152. DOI: https://doi.org/10.6065/apem.2017.22.3.145, PMID: 29025199

Koenig W, Sund M, Fröhlich M, Löwel H, Hutchinson WL, Pepys MB. 2003. 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* **158**:357–364. DOI: https://doi.org/10.1093/aje/kwg135, PMID: 12915501

Kolde R. 2019. Pheatmap: Pretty Heatmaps. 1.0.12. R Package. https://cran.r-project.org/web/packages/pheatmap/index.html

Kwak SH, Park KS. 2016. Recent progress in genetic and epigenetic research on type 2 diabetes. *Experimental & Molecular Medicine* **48**:e220. DOI: https://doi.org/10.1038/emm.2016.7, PMID: 26964836

Lea AJ, Vockley CM, Johnston RA, Del Carpio CA, Barreiro LB, Reddy TE, Tung J. 2018. Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife* **7**:e37513. DOI: https://doi.org/10.7554/eLife.37513, PMID: 30575519

Liu Y, Buil A, Collins BC, Gillet LCJ, Blum LC, Cheng LY, Vitek O, Mouritsen J, Lachance G, Spector TD, Dermitzakis ET, Aebersold R. 2015. Quantitative variability of 342 plasma proteins in a human twin population. *Molecular Systems Biology* **11**:786. DOI: https://doi.org/10.15252/msb.20145728, PMID: 25652787

Lord J, Cruchaga C. 2014. The epigenetic landscape of Alzheimer's disease. *Nature Neuroscience* **17**:1138–1140. DOI: https://doi.org/10.1038/nn.3792, PMID: 25157507

Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Assimes TL, Ferrucci L, Horvath S. 2019. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**:303–327. DOI: https://doi.org/10.18632/aging.101684, PMID: 30669119

Ma Y, Liu Y, Zhang Z, Yang GY. 2019. Significance of Complement System in Ischemic Stroke: A Comprehensive Review. *Aging and Disease* **10**:429–462. DOI: https://doi.org/10.14336/AD.2019.0119, PMID: 31011487

Mantovani S, Gordon R, Macmaw JK, Pfluger CMM, Henderson RD, Noakes PG, McCombe PA, Woodruff TM. 2014. Elevation of the terminal complement activation products C5a and C5b-9 in ALS patient blood. *Journal of Neuroimmunology* **276**:213–218. DOI: https://doi.org/10.1016/j.jneuroim.2014.09.005, PMID: 25262158

McCartney DL, Stevenson AJ, Hillary RF, Walker RM, Bermingham ML, Morris SW, Clarke TK, Campbell A, Murray AD, Whalley HC, Porteous DJ, Visscher PM, McIntosh AM, Evans KL, Deary IJ, Marioni RE. 2018a. Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**:214–220. DOI: https://doi.org/10.1016/j.ebiom.2018.10.051, PMID: 30389506

McCartney DL, Stevenson AJ, Walker RM, Gibson J, Morris SW, Campbell A, Murray AD, Whalley HC, Porteous DJ, McIntosh AM, Evans KL, Deary IJ, Marioni RE. 2018b. Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. *Alzheimer's & Dementia* **10**:429–437. DOI: https://doi.org/10.1016/j.dadm.2018.05.006, PMID: 30167451

McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, Morris SW, Bermingham ML, Campbell A, Murray AD, Whalley HC, Gale CR, Porteous DJ, Haley CS, McRae AF, Wray NR, Visscher PM, McIntosh AM, Evans KL, Deary IJ, et al. 2018c. Epigenetic prediction of complex traits and death. *Genome Biology* **19**:136. DOI: https://doi.org/10.1186/s13059-018-1514-1, PMID: 30257690

McCartney DL, Min JL, Richmond RC, Lu AT, Sobczyk MK, Davies G, Broer L, Guo X, Jeong A, Jung J, Kasela S, Katrinli S, Kuo PL, Matias-Garcia PR, Mishra PP, Nygaard M, Palviainen T, Patki A, Raffield LM, Marioni RE. 2020. Genome-Wide Association Studies Identify 137 Loci for DNA Methylation Biomarkers of Ageing. [bioRxiv]. DOI: https://doi.org/10.1101/2020.06.29.133702

Messner CB, Demichev V, Wendisch D, Michalick L, White M, Freiwald A, Textoris-Taube K, Vernardis SI, Egger AS, Kreidl M, Ludwig D, Kilian C, Agostini F, Zelezniak A, Thibeault C, Pfeiffer M, Hippenstiel S, Hocke A, von Kalle C, Campbell A, et al. 2020. Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Systems* **11**:11–24. DOI: https://doi.org/10.1016/j.cels.2020.05.012, PMID: 32619549

Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, Carnero-Montoro E, Lawson DJ, Burrows K, Suderman M, Bretherick AD, Richardson TG, Klughammer J, Iotchkova V, Sharp G, Al Khleifat A, Shatunov A, Iacoangeli A, McArdle WL, Ho KM, et al. 2021. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nature Genetics* **53**:1311–1321. DOI: https://doi.org/10.1038/s41588-021-00923-x

Moldoveanu AI, Shephard RJ, Shek PN. 2000. Exercise elevates plasma levels but not gene expression of IL-1beta, IL-6, and TNF-alpha in blood mononuclear cells. *Journal of Applied Physiology* **89**:1499–1504. DOI: https://doi.org/10.1152/jappl.2000.89.4.1499, PMID: 11007588

Morgan BP, Harris CL. 2015. Complement, a target for therapy in inflammatory and degenerative diseases. *Nature Reviews Drug Discovery* **14**:857–877. DOI: https://doi.org/10.1038/nrd4657, PMID: 26493766

MRC-IEU. 2021. The MRC-IEU catalog of epigenome-wide association studies. http://www.ewascatalog.org [Accessed April 1, 2021].

Navrady LB, Wolters MK, MacIntyre DJ, Clarke T-K, Campbell AI, Murray AD, Evans KL, Seckl J, Haley C, Milburn K, Wardlaw JM, Porteous DJ, Deary IJ, McIntosh AM. 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**:13–14g. DOI: https://doi.org/10.1093/ije/dyx115, PMID: 29040551

Ngo D, Benson MD, Long JZ, Chen Z-Z, Wang R, Nath AK, Keyes MJ, Shen D, Sinha S, Kuhn E, Morningstar JE, Shi X, Peterson BD, Chan C, Katz DH, Tahir UA, Farrell LA, Melander O, Mosley JD, Carr SA, et al. 2021. Proteomic profiling reveals biomarkers and pathways in type 2 diabetes risk. *JCI Insight* **6**:e144392. DOI: https://doi.org/10.1172/jci.insight.144392, PMID: 33591955

NHS England. 2016. mproving Outcomes Through Personalised Medicine. https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf [Accessed April 1, 2021].

Ort M, Dingemanse J, van den Anker J, Kaufmann P. 2020. Treatment of Rare Inflammatory Kidney Diseases: Drugs Targeting the Terminal Complement Pathway. *Frontiers in Immunology* **11**:599417. DOI: https://doi.org/10.3389/fimmu.2020.599417, PMID: 33362783

Pedersen TL. 2021. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. 2.0.5. R Package Version. https://CRAN.R-project.org/package=ggraph

Pérez-Pérez R, García-Santos E, Ortega-Delgado FJ, López JA, Camafeita E, Ricart W, Fernández-Real JM, Peral B. 2012. Attenuated metabolism is a hallmark of obesity as revealed by comparative proteomic analysis of human omental adipose tissue. *Journal of Proteomics* **75**:783–795. DOI: https://doi.org/10.1016/j.jprot.2011.09.016, PMID: 21989264

Peters A, Nawrot TS, Baccarelli AA. 2021. Hallmarks of environmental insults. *Cell* **184**:1455–1468. DOI: https://doi.org/10.1016/j.cell.2021.01.043, PMID: 33657411

Petersen AK, Zeilinger S, Kastenmüller G, Römisch-Margl W, Brugger M, Peters A, Meisinger C, Strauch K, Hengstenberg C, Pagel P, Huber F, Mohney RP, Grallert H, Illig T, Adamski J, Waldenberger M, Gieger C, Suhre K. 2014. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Human Molecular Genetics* **23**:534–545. DOI: https://doi.org/10.1093/hmg/ddt430, PMID: 24014485

Pietzner M, Wheeler E, Carrasco-Zanini J, Raffler J, Kerrison ND, Oerton E, Auyeung VPW, Luan J, Finan C, Casas JP, Ostroff R, Williams SA, Kastenmüller G, Ralser M, Gamazon ER, Wareham NJ, Hingorani AD, Langenberg C. 2020. Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nature Communications* **11**:6397. DOI: https://doi.org/10.1038/s41467-020-19996-z, PMID: 33328453

R Development Core Team. 2020. R: A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org

Revelle W. 2020. Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. 2.0.9. Psych. https://CRAN.R-project.org/package=psych

Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, Dudbridge F. 2018. Estimation of a significance threshold for epigenome-wide association studies. *Genetic Epidemiology* **42**:20–33. DOI: https://doi.org/10.1002/gepi.22086, PMID: 29034560

Safiri S, Kolahi AA, Hoy D, Smith E, Bettampadi D, Mansournia MA, Almasi-Hashiani A, Ashrafi-Asgarabad A, Moradi-Lakeh M, Qorbani M, Collins G, Woolf AD, March L, Cross M. 2019. Global, regional and national burden of rheumatoid arthritis 1990-2017: a systematic analysis of the Global Burden of Disease study 2017. *Annals of the Rheumatic Diseases* **78**:1463–1471. DOI: https://doi.org/10.1136/annrheumdis-2019-215920, PMID: 31511227

Seeboth A, McCartney DL, Wang Y, Hillary RF, Stevenson AJ, Walker RM, Campbell A, Evans KL, McIntosh AM, Hägg S, Deary IJ, Marioni RE. 2020. DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936. *Clinical Epigenetics* **12**:49. DOI: https://doi.org/10.1186/s13148-020-00838-0, PMID: 32216821

Serban KA, Pratte KA, Bowler RP. 2021. Protein Biomarkers for COPD Outcomes. *Chest* **159**:2244–2253. DOI: https://doi.org/10.1016/j.chest.2021.01.004, PMID: 33434499

Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, Redmond P, Cox SR, Pattie A, Corley J, Murphy L, Martin NG, Montgomery GW, Starr JM, Wray NR, Deary IJ, Visscher PM. 2014. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Research* **24**:1725–1733. DOI: https://doi.org/10.1101/gr.176933.114, PMID: 25249537

Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, Deary IJ, MacIntyre DJ, Campbell H, McGilchrist M, Hocking LJ, Wisely L, Ford I, Lindsay RS, Morton R, Palmer CNA, Dominiczak AF, Porteous DJ, Morris AD. 2013. 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* **42**:689–700. DOI: https://doi.org/10.1093/ije/dys084, PMID: 22786799

Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, Walker RM, Evans KL, Boutin TS, Hayward C, McRae AF, McColl BW, Spires-Jones TL, McIntosh AM, Deary IJ, Marioni RE. 2020. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clinical Epigenetics* **12**:113. DOI: https://doi.org/10.1186/s13148-020-00903-8, PMID: 32718350

Stevenson AJ, Gadd DA, Hillary RF, McCartney DL, Campbell A, Walker RM, Evans KL, Harris SE, Spires-Jones TL, McRae AF, Visscher PM, McIntosh AM, Deary IJ, Marioni RE. 2021. Creating and Validating a DNA Methylation-Based Proxy for Interleukin-6. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* **76**:2284–2292. DOI: https://doi.org/10.1093/gerona/glab046, PMID: 33595649

Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, Sarwath H, Thareja G, Wahl A, DeLisle RK, Gold L, Pezer M, Lauc G, El-Din Selim MA, Mook-Kanamori DO, Al-Dous EK, Mohamoud YA, Malek J, Strauch K, Grallert H,

et al. 2017. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications* **8**:14357. DOI: https://doi.org/10.1038/ncomms14357, PMID: 28240269

Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, et al. 2018. Genomic atlas of the human plasma proteome. *Nature* **558**:73–79. DOI: https://doi.org/10.1038/s41586-018-0175-2, PMID: 29875488

Taylor AM, Pattie A, Deary IJ. 2018. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology* **47**:1042–1042r. DOI: https://doi.org/10.1093/ije/dyy022, PMID: 29546429

Therneau TM. 2020a. A Package for Survival Analysis in R. 3.2-7. R Package Version. https://CRAN.R-project.org/package=survival

Therneau TM. 2020b. coxme: Mixed Effects Cox Models. 2.2-16. R Package. https://CRAN.R-project.org/package=coxme

Trejo Banos D, McCartney DL, Patxot M, Anchieri L, Battram T, Christiansen C, Costeira R, Walker RM, Morris SW, Campbell A, Zhang Q, Porteous DJ, McRae AF, Wray NR, Visscher PM, Haley CS, Evans KL, Deary IJ, McIntosh AM, Hemani G, et al. 2020. Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications* **11**:2865. DOI: https://doi.org/10.1038/s41467-020-16520-1, PMID: 32513961

Watanabe K, Taskesen E, van Bochoven A, Posthuma D. 2017. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**:1–11. DOI: https://doi.org/10.1038/s41467-017-01261-5, PMID: 29184056

Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, Curtis HJ, Mehrkar A, Evans D, Inglesby P, Cockburn J, McDonald HI, MacKenna B, Tomlinson L, Douglas IJ, Rentsch CT, Mathur R, Wong AYS, Grieve R, Harrison D, et al. 2020. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**:430–436. DOI: https://doi.org/10.1038/s41586-020-2521-4, PMID: 32640463

World Health Organization. 2018. Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region. WHO.

Yao C, Chen G, Song C, Keefe J, Mendelson M, Huan T, Sun BB, Laser A, Maranville JC, Wu H, Ho JE, Courchesne P, Lyass A, Larson MG, Gieger C, Graumann J, Johnson AD, Danesh J, Runz H, Hwang S-J, et al. 2018. Author Correction: Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications* **9**:3853. DOI: https://doi.org/10.1038/s41467-018-06231-z, PMID: 30228274

Zaghlool SB, Kühnel B, Elhadad MA, Kader S, Halama A, Thareja G, Engelke R, Sarwath H, Al-Dous EK, Mohamoud YA, Meitinger T, Wilson R, Strauch K, Peters A, Mook-Kanamori DO, Graumann J, Malek JA, 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**:15. DOI: https://doi.org/10.1038/s41467-019-13831-w, PMID: 31900413

Zhang Y, Wilson R, Heiss J, Breitling LP, Saum KU, Schöttker B, Holleczek B, Waldenberger M, Peters A, Brenner H. 2017. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nature Communications* **8**:14617. DOI: https://doi.org/10.1038/ncomms14617, PMID: 28303888

Zhang Q, Marioni RE, Robinson MR, Higham J, Sproul D, Wray NR, Deary IJ, McRae AF, Visscher PM. 2018. Genotype effects contribute to variation in longitudinal methylome patterns in older people. *Genome Medicine* **10**:75. DOI: https://doi.org/10.1186/s13073-018-0585-7, PMID: 30348214

Zhang F, Chen W, Zhu Z, Zhang Q, Nabais MF, Qi T, Deary IJ, Wray NR, Visscher PM, McRae AF, Yang J. 2019. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biology* **20**:107. DOI: https://doi.org/10.1186/s13059-019-1718-z, PMID: 31138268