DNA methylation proxies for 16 plasma proteins predict the incidence of 7 leading causes of morbidity

Danni A Gadd[1,*], Robert F Hillary[1,*], Daniel L McCartney[1,*], Anna J Stevenson[1], Cliff Nangle[1], Archie Campbell[1], Robin Flaig[1], Sarah E Harris[2,3], Rosie M Walker[4], Liu Shi[5], Elliot M Tucker-Drob[6,7], Allan F McRae[8], Ian J Deary[2,3], David J Porteous[1], Caroline Hayward[1,9], Peter M Visscher[8], Simon R Cox[2,3], Kathryn L Evans[1], Andrew M McIntosh[1,10], Riccardo E Marioni[1,†]

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

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

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

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

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

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

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

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

[9] Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU

[10] 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

Chronic morbidities place longstanding burdens on our health as we age. Although protein biomarkers are critical for the early detection of such diseases, current studies are limited by low sample sizes, variability in proteomics methods and fluctuations in inflammatory protein expression. Here, we present a novel framework for protein-by-proxy analysis of incident disease. We show that DNA methylation proxies for nine inflammatory and seven neurology plasma proteins (generated in up to 875 individuals in the Lothian Birth Cohort 1936) predict the incidence of seven leading causes of morbidity in the Generation Scotland cohort (n=9,537), ascertained via electronic health data linkage over a follow-up period of up to 14 years. After correction for multiple testing and adjustment for common disease risk factors, these included proxy associations between CCL11 and depression (Hazard Ratio: HR = 1.45, P = 1.8 x 10^{-4}), VEGFA and ischaemic heart disease (HR = 1.16, P = 0.02) and associations between incident diabetes and FGF-21 (HR = 1.39, P = 9.7 x 10^{-7}), NEP (HR = 1.32, P = 2.8 x 10^{-6}) and N-CDase (HR = 1.16, P = 0.02). Several of the protein-proxy associations with disease pinpoint proteins that are already therapeutic targets for the diseases in question. These results provide new opportunities to identify circulating biomarkers for disease detection and candidate pathways for drug targeting.

1 Introduction

Ageing is associated with the increased incidence of many chronic morbidities which can raise an individual's risk of disability and mortality. The development of clinical tools for early detection and prevention of such morbidities is therefore a priority. Plasma protein biomarkers can help to achieve this goal, as their divergent expression provides insight into the likely causes and consequences of disease in an individual, oftentimes prior to the onset of clinically relevant symptoms [1]. The recent characterization of the genome-wide genetic and epigenetic architectures of plasma proteins highlights the potential for integrated omics data to uncover causal pathways contributing to protein expression [2,3] and health outcomes [2–7]. Developing robust ways to identify the relationships that exist between genetic architectures, circulating proteins and disease onset is therefore critical to uncovering both the mechanisms of disease and ways to stratify risk.

Despite the clear utility of protein biomarkers, there are several limitations that impede current research. The use of protein expression for risk prediction and causal inference relies on large and consistent datasets. Current work is limited by both the availability of suitable samples and the variability across the many methods used to generate proteomics data [8]. Additionally, proteins involved in the acute-phase inflammatory

response are known to be variable in their within-person expression, rendering single-time-point measures of these markers problematic [9–11].

We propose DNA methylation (DNAm) proxies as a means to circumvent these limitations. DNAm involves the addition of a methyl group to a cytosine residue, typically in the context of a CG dinucleotide (CpG), which can regulate gene expression [12]. Genome-wide DNAm is commonly measured by the Illumina Methylation BeadChip and is widely profiled in cohort studies that do not have proteomics data available; DNAm proxies for protein expression would subsequently create new opportunities for large and consistent analyses in many existing cohorts. Moreover, as previously demonstrated for C-reactive protein (CRP) and Interleukin-6 (IL6), single-time point DNAm proxies have the strong potential to be more stable than serological measures over multiple longitudinal measurements, [13–15] offering a means to minimise the variability known to affect acute-phase inflammatory protein measures. We therefore hypothesise that DNAm can offer an intermediate proxy for the expression of plasma proteins to advance disease biomarker detection.

Here, we developed blood-based DNA methylation proxies for plasma protein expression and related these proxies to 12 leading causes of morbidity and mortality (**Table 1, Fig. 1**). A panel of 160 inflammatory and neurology proteins in a group of up to 875 individuals from the Lothian Birth Cohort 1936 were used to generate DNAm proxies through elastic net penalised regression models. This initial cohort was first split into train and holdout test sets to evaluate proxy performance. The models were then re-run using all possible individuals as the training set and proxies were applied to two separate test sets in order to select the most robust measures. The selected proxies were then projected into an independent test sample of 9,537 individuals from Generation Scotland, the largest single cohort to have DNAm information available for such predictions. Taking advantage of retrospective and prospective data linkage to primary (general practice records) and secondary care (hospital) records, we relate protein proxies to the incidence of 12 diagnoses over a follow up period of up to 14 years. Ten of the traits are recognised by the World Health Organisation as leading causes of either morbidity or mortality and include diseases such as chronic obstructive pulmonary disease (COPD), diabetes, stroke and Alzheimer's dementia [16,17]. Rheumatoid arthritis and inflammatory bowel disease were also included as they are chronic conditions which constitute substantial chronic health burdens [18,19].

This work demonstrates that protein-by-proxy analyses can uncover early markers that may be critical to disease incidence, several years prior to diagnoses. As these diseases constitute a major global health burden and can profoundly affect the wellbeing of an individual as they age, markers that complement risk

stratification and deployment of preventative interventions are essential. The availability of our predictors in a Shiny App (MethylDetectR) allows for the projection of protein expression into any cohort which has Illumina DNAm data, creating opportunities for historic and existing datasets to be probed with no requirements for additional proteomics sampling. This approach facilitates a deeper understanding of the early mechanisms associated with disease onset and identify candidate pathways for therapeutic targeting.

2 Results

2.1 Creation and validation of DNAm proxies for protein expression

There were 18 neurology and 20 inflammatory proxies generated through elastic net regressions which correlated (r > 0.1, P < 0.1) with protein expression when applied to the initial holdout set in the Lothian Birth Cohort 1936 (LBC1936). Predictors for these 38 proteins were then created using the full LBC1936 cohort as the training dataset in elastic net regressions. This generated protein predictors for 37 proteins; one protein (TRAIL) contained only an intercept term (i.e. no nonzero features) and was therefore excluded. There were between 11 and 457 features for the 37 proxies (median = 84). Validation of these proxies in two independent test sets (the Stratifying Resilience and Depression Longitudinally: STRADL subset of Generation Scotland, n=778, and the Lothian Birth Cohort 1921: LBC1921, n=162) resulted in the selection of 14 neurology and nine inflammatory proxies which performed optimally (r > 0 and P < 0.05 in at least one test cohort, **Fig. 2**) (**Supplementary Tables 1-2**). A further three inflammatory proxies were included without comparisons available, based on their performance in the initial holdout set (P < 0.05). Though the IL6 proxy performed poorly in the GS:STRADL test set, it has been validated against ELISA measures previously and was therefore also included [14]. Predictor weights for these 27 proxies (13 inflammatory, 14 neurology) are provided in **Supplementary Table 3**. Across the selected 27 proxies, there were 223 CpG probes which were included in 2 or more proxies; a summary for each CpG site with annotations to the MRC-IEU EWAS catalog[20] traits (P < 3.6 x 10^{-8}) is presented in **Supplementary Table 4**. The smoking-related site cg05575921 was the most frequently selected and was present in 12 proxies. The highest proportion of *cis* associations – CpGs within 10Mb of the transcription start site of the gene encoding the protein – were found for MDGA1 and IL-18R1 (11% and 6%, respectively**)**.

2.2 Proxy associations with incident diseases in Generation Scotland

There were 72 associations between the selected DNAm protein proxies and time-to-event disease incidence in the basic mixed effects Cox proportional hazards regression models with P < 0.05 after FDR correction (**Supplementary Table 5**). This equated to a maximum uncorrected P-value 0.01. Of the 72 associations, 26

remained significant with $P < 0.05$ in the fully-adjusted model accounting for age, sex and common risk factors, including alcohol consumption, body mass index, deprivation, educational attainment, and a DNAm-based proxy for smoking (**Supplementary Tables 6-7**). Mean attenuation of log hazard ratios was -60.5% (ranging from -163.2% to 26.5%, with 69 attenuated, two enhanced and one unchanged), across the 72 proxies and -33.4% (ranging from -78.9% to 26.5%, with 23 attenuated, two enhanced and one unchanged) across the 26 significant associations. Globally, the proportional hazards assumption was satisfied for all models. However, six of the 26 fully-adjusted associations failed the proportional hazards assumptions for the protein-proxy variable ($P < 0.05$ for the association between the Schoenfeld residuals and time; **Supplementary Table 8**). Restricting the time-to-event/censor period by possible years of follow-up, there were minimal differences in the proxy-disease hazard ratios between follow-up periods which did not violate the assumption and those that did (**Supplementary Table 9**). No associations were therefore excluded.

A hazard ratio-weighted network summary of the 26 proxy-disease relationships with $P < 0.05$ in the fully-adjusted model is presented in **Fig. 3**. A summary of the hazard ratios and confidence intervals for each proxy-disease relationship is presented in **Fig. 4**. These findings involved 16 unique protein proxies (nine inflammatory and seven neurology) and seven disease outcomes: depression, diabetes, ischaemic heart disease, stroke, rheumatoid arthritis, COPD and lung cancer. A one standard deviation increase in DNAm proxy CCL11 levels at baseline was associated with risk of incident depression (HR = 1.45, 95% CI = [1.19, 1.75], $P = 1.8 \times 10^{-4}$). Incident diabetes was associated with elevated baseline proxy measures of three proxies: NEP (HR = 1.32, 95% CI = [1.18, 1.49], $P = 2.8 \times 10^{-6}$), FGF-21 (HR = 1.39, 95% CI = [1.22, 1.58], $P = 9.7 \times 10^{-7}$) and N-CDase (HR = 1.16, 95% CI = [1.03, 1.31], P = 0.02). Of the 16 protein proxies which had significant associations, there were nine which had relationships with multiple disease outcomes. For example, in addition to the association found for diabetes, FGF-21 proxy levels at baseline were associated with stroke (HR = 1.24, 95% CI = [1.08, 1.43], P = 0.002). Whereas higher levels of these proxies were associated with increased risk, higher levels of the SIGLEC1 proxy were linked to a reduced incidence of lung cancer (HR = 0.79, 95% CI = [0.68, 0.92], P = 0.003). Higher GZMA levels were associated with a decreased risk of both COPD (HR = 0.83, 95% CI = [0.73, 0.96], P = 0.01) and rheumatoid arthritis (HR = 0.67, 95% CI = [0.49, 0.92], P = 0.01). Additional associations for COPD, ischaemic heart disease and lung cancer were also observed (**Supplementary Table 7**).

Full correlation structures for the 27 proxies included in the Cox analyses and the 16 which were associated with incident diseases are presented in **Supplementary Fig. 1**. There were no correlations between protein proxy measures and common risk factor covariates with $r > 0.30$, with the exception of the EpiSmokEr DNAm-derived measure of smoking and age (**Supplementary Fig. 2**). There were four diseases which were associated with multiple proxy predictors; correlation structures and principal components analyses for these proteins are presented in **Supplementary Fig. 3**. Filtering by an eigenvalue > 1, there were three components for the 12 COPD-associated proxies and two components across the five lung cancer proxies,

suggesting that DNAm proxies were reflecting several independent proteomic signatures. One component with eigenvalue > 1 was present for both ischaemic heart disease and diabetes. A sensitivity analysis which removed controls who died after the study baseline from the Cox models did not considerably alter the 26 associations (**Supplementary Table 7**).

2.3 White blood cell influences on proxies

There were correlations of up to r = 0.77 between protein proxies and estimated white blood cell (WBC) proportions in the Generation Scotland cohort (**Supplementary Fig. 2**). A sensitivity analysis was therefore conducted, which adjusted for estimated WBC proportions in the fully-adjusted Cox proportional hazards models. In this analysis, 19 of the 26 fully-adjusted associations remained significant (**Supplementary Table 10**). In the 19 associations, there was an overall mean attenuation in log hazard ratios of -12.9%, ranging from -31.3% to 33.8% in relation to the fully-adjusted model, with 15 attenuated, 3 enhanced and one unchanged. In a further sensitivity analysis, relationships between estimated WBC proportions and incident diseases were assessed in the Cox model structure, independently of proxies. Four inverse relationships (higher cell proportions linked to decreased disease risk) were found between natural killer cells and the incidence of COPD, rheumatoid arthritis, diabetes and pain (**Supplementary Table 11**).

To explore the interplay between measured white blood cells and our proxies, the elastic net penalised regression models used in the train/test optimisation phase in LBC1936 were re-run, with measured WBCs included as features. Of the six immune cells available (**Methods**), neutrophil and monocyte features were selected for and increased the correlation strength between the proxy and measured proteins in the test set for seven of the nine inflammatory proxies and three of the seven neurology proxies, providing further evidence of a tightly interlinked relationship (**Supplementary Table 12**).

2.4 Protein Quantitative trait Locus (pQTL) mapping

To determine if pQTL mapping was possible with the proxy measures, GWAS analyses were run for the proxies corresponding to 7 proteins with GWAS significant SNPs in previous Lothian Birth Cohort 1936 analyses [2,3] (N=9 genome-wide significant SNPs; **Supplementary Note 1**). We replicated 7/9 sites from previous studies at $P < 5 \times 10^{-8}$, six of which were *cis* associations for the protein coding gene, and one of which was *trans* (rs46876657; **Supplementary Table 13**) [2,3]. Moreover, all 7 SNPs were within 75kb of a CpG that was included in the corresponding protein proxy, six of which were previously reported as

methylation quantitative trait loci (mQTLs) for protein proxy CpGs [PMID 27036880] [21]. The proxies therefore largely capture mQTLs.

3 Discussion

By projecting DNA methylation proxies for protein expression into a large cohort with extant data linkage, we have identified 16 protein proxies that predict the incidence of seven leading causes of morbidity, after controlling for common risk factors such as smoking and deprivation.

Here, we developed DNAm proxies for protein expression to advance disease biomarker detection and offer a means to identify therapeutic targets. We were able to validate several proxies, demonstrating that DNAm can proxy for the expression of these proteins. Further to this, many of the proxy-disease relationships that we identify have been found linking true protein measurements and diseases, suggesting that our methylomic proxies capture clinically relevant facets of these pathways. Proteins corresponding to proxies associated with incident disease in our study are the targets for current therapeutic approaches; a recent trial demonstrated a reduction in the rate of decline in renal function in individuals that had type 2 diabetes and heart failure resulting from the inhibition of NEP [22,23]. NEP inhibition has also been shown to associate with improved insulin sensitivity in those with obesity and hypertension [24], which has led to this pathway being proposed as a candidate for type 2 diabetes therapy [25]. Gene therapies targeting VEGFA to promote localised angiogenesis are also in ongoing trials for the treatment of ischaemic heart disease [26,27] and several trials are in progress for anti-SIGLEC antibody therapies in cancer due to the role of this receptor family in the modulation of tumour-associated macrophages [28,29]. Taken together, these examples suggest that our DNAm proxies are able to identify disease-relevant pathways for therapeutic targeting. Though our associations are in some cases contradictory to these therapeutic strategies, such as the inverse association found between SIGLEC1 and lung cancer, these instances may reflect a time-critical, systemic variation in protein expression during the window prior to diagnosis and their causality should therefore be explored further.

Nine of the proxies were identified as potential biomarkers for the onset of more than one disease. For example, the FGF-21 proxy measure associated with incident stroke and diabetes. Serum and plasma measures of FGF-21 have been shown to be predictive of type 2 diabetes and metabolic syndrome [30–35], with FGF-21 implicated in the response to metformin [36]. FGF-21 has also been characterised as a blood-based marker for poor cardiovascular health, both generally and in those with type 2 diabetes [37–40]. DNAm proxies

may therefore uncover important, but as yet, undefined nodes relevant to multiple diseases that are linked through the CpG features contributing to them. There were also four diseases which were associated with multiple, often intercorrelated proxies. Many of the proteins implicated in proxy-COPD relationships have been identified as potential markers for the disease and are thought to contribute to destruction of lung tissue as part of an ongoing, inflammatory state [41–44]. Principal components analysis suggests that the proxies are capturing multiple, distinct signatures of inflammatory protein expression in those that subsequently receive a COPD diagnosis. Given that anti-inflammatory therapy for COPD is highly desired but has as yet been challenging to achieve [45,46], proxy-driven insight into the interrelatedness of the wider inflammasome may offer valuable context for therapeutic strategies.

Though the known disparities between blood and brain DNA methylation [47,48] may have limited the detection of markers relevant to neurological diseases with unique pathology in the brain, a relationship was found between CCL11 and depression. The mechanisms by which CCL11 may be related to depression are unclear; however, CCL11 is thought to mediate peripheral and central nervous system inflammation, with evidence that it has microglial and astrocytic targets [49]. As CCL11 is suggested as one of several cytokine plasma markers that may identify those with depression [50–53], circulating DNAm proxies could prove to be relevant markers for the stratification of psychiatric illness risk.

We found that the genetic architecture of the proxy proteins largely captures either mQTLs, or pQTLs which constitute mQTLs for the CpG sites contributing to the protein proxies. The proxies are therefore unlikely to identify novel pQTL findings. Though testing in a holdout set and two external cohorts suggested that many of our 27 optimal proxies were robustly capturing protein expression, there were many proteins for which we did not achieve reasonable proxies. As the training sample increases, we expect convergence between proxy projections and measured proteins; however, our previous work indicates that there is a threshold for variance explained in protein expression by genome-wide DNAm [2,3]. Nevertheless, even where DNAm proxies CRP and IL6 correlate ~0.2 with measured protein levels, they provide a more stable measure of expression than proteomic measures when averaged longitudinally; they often outperform the measured proteins in relation to associations with health outcomes and lifestyle factors [13–15]. As with CRP and IL6, many of the proteins we have created proxies for are involved in or associated with the acute-phase response; GZMA is thought to promote the release of IL6, IL8 and TNF-alpha, thereby inducing a cytokine syndrome in those with sepsis (a condition hallmarked by a rapid alterations in the inflammatory state)[11,54]. The proxies we have created for GZMA and IL8 may therefore be more stable than longitudinal serum measurements. Consequently, the protein proxy-phenotyping approach may augment insights into inflammatory pathways which can be difficult to quantify due to natural and disease-associated variability [55,56].

This study has a number of limitations. First, the size of the protein training dataset constrained predictor generation. Second, due to missing covariate information, cases were excluded from the fully-adjusted models; however, this had a marginal influence on the main associations. Third, whereas there was a degree of attenuation in proxy-trait relationships upon adjustment for white blood cell proportions, our analyses highlight the interrelated nature of these measures with the proxies. Though many of the strongest relationships withstood adjustment for white blood cell proportions, measured white blood cells were selected as contributing features for many of the proxies in our elastic net sensitivity analyses. It is therefore challenging to establish directionality between the proxies, immune cells and diseases; however, the selection of immune cells as features contributing to proxies presents an interesting avenue for further exploration. Fourth, Cox model effect sizes should be interpreted with the caveats that hazard ratios reflect a relatively arbitrary scale (per SD of the DNAm score) and that DNAm scores were generated using relative protein measures, rather than absolute quantification. Fifth, the associations present between proxy measures and disease incidence may have been influenced by external factors such as prescription medications and disease prevalence at baseline, which should be investigated future analyses. Sixth, though we show that many of the proxies trained in an age-homogenous cohort performed well when applied to cohorts of differing age distributions, it is likely that protein measurements in our cohorts are somewhat age-sensitive. It is therefore possible that our proxies may not generalise optimally beyond a cohort of healthy ageing individuals. Finally, our proxies were trained and tested on individuals from relatively homogeneous Scottish genetic heritage, which may limit their applicability to individuals from other genetic ancestries.

This current application is particularly valuable for cohorts such as Generation Scotland, which does not currently have protein data available but is the largest single-cohort DNAm resource in the world. We have created a Shiny app (MethylDetectR [57]) to enable any study with Illumina-based DNA methylation data to easily generate and visualise projections for the 27 protein proxies in addition to DNAm predictors of lifestyle [58] and chronological age [59]. These proxies can be filtered by age and sex and visualised for an individual (or group of individuals e.g., disease cases) relative to the rest of the input cohort (**Fig. 5**). Another strength is the extensive data linkage capacity in Generation Scotland that allowed us to investigate time-to-event for several common disease outcomes. Whereas, the number of incident cases was modest for some traits, the extant nature of the linkage means that we will continue to acquire cases across all disease areas. Our findings suggest that proxy phenotyping approaches and data linkage to electronic health records in large, population-based studies have the potential to (1) capture clinically relevant facets of true protein expression; (2) highlight novel disease-associated proteins and mechanisms, many of which have existing drug targets; and (3) augment risk prediction years prior to disease onset. This knowledge is integral to the

early detection and improved risk stratification of complex diseases, which are central aims for both biomedical research and public health [60,61].

In conclusion, we validate a novel framework for the large-scale identification of protein biomarkers associated with disease. Through this framework, we show that DNA methylation proxies for the expression of 16 plasma proteins predict the incidence of seven leading causes of morbidity and mortality. This work highlights the potential for methylomics approaches to uncover the drivers of multimorbidity as we age and provides context relevant to preventative interventions.

4 Methods

4.1 Lothian Birth Cohorts of 1936 and 1921

The Lothian Birth Cohorts of 1936 (LBC1936; N=1,091) and 1921 (LBC1921; N=550) are longitudinal studies of ageing in individuals living in Edinburgh and the surrounding areas [62,63]. Participants completed a childhood intelligence test at age 11 years in 1947 and were recruited for these cohorts at mean ages of 79 (LBC1921) and 70 (LBC1936). Participants in both cohorts have been followed up approximately every 3-4 years since baseline [64]. A series of cognitive, clinical, physical and social data, along with blood donations that have been used for genetic, epigenetic, and proteomic measurement were collected at the majority of visits.

Separate panels of 92 inflammatory and 92 neurology proteins (Olink® antibody-based technology – measurement details in **Supplementary Note 2**) were assessed in plasma samples from the first and second waves of the LBC1936 study, respectively (mean age 69.6 years for inflammatory n=875 and 72.5 years for neurology n=706). The Olink® neurology panel was also assessed in plasma samples from wave 3 of the LBC1921 cohort in 162 individuals (mean age 86.7 years). Protein levels were rank-based inverse normal transformed and regressed on age, sex and four genetic principal components. This was performed separately for the train and test sets used in the penalised regression models. two neurology proteins, MAPT and HAGH, and twenty-one inflammatory proteins were excluded due to >40% of observations being below the lower limit of detection; one further inflammatory protein, BDNF, failed quality control and was also removed from the study. This resulted in 160 protein measurements across both panels for use.

Blood-based DNA methylation was assessed using the Illumina 450k array. Quality control details are reported in **Supplementary Note 2**. There were 459,308 methylation sites measured in the LBC1936. To permit comparison across platforms, sites that overlapped between the Illumina 450k and Illumina EPIC

arrays were used as the features (93% of the original sites, n=428,489) for the penalised regression models. CpG features were scaled to have a mean of zero and variance of one prior to projections into cohorts.

White blood cell measures in the LBC1936 were acquired using the same blood samples taken for methylation and were collected and processed on the same day. Technical details for these measures have been outlined previously [65]. Monocytes, granulocytes, natural killer cells, B cells and both CD8T and CD4T cells were available for inclusion in the elastic net sensitivity analyses.

4.2 Generation Scotland and STRADL

Generation Scotland: the Scottish Family Health Study (GS) is a large, family-structured, population-based cohort study of >24,000 individuals from across Scotland. 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) [66].

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 [67]. Measurements for 4,236 proteins in 1,065 individuals from the STRADL cohort were recorded (SomaScan® technology – measurement details in **Supplementary Note 2**). Of the original 160 Olink® proteins present on inflammatory and neurology panels, 56 matched the SomaScan SOMAmer IDs and were used for test assessments of proxies where possible. The final test set was comprised of 778 individuals (mean age 60.1 ± 8.81 years) with both protein measurements and DNA methylation available. The methylation data were assessed in two separate batches (n_{batch1}= 504, n_{batch2} = 306 – details in **Supplementary Note 2**).

In the main GS cohort, blood-based DNA methylation has been generated in two separate sets using the Illumina EPIC array. Prior to quality control, Set 1 comprised 5,190 related individuals whereas Set 2 comprised 4,583 individuals, unrelated to each other and to those in Set 1. Quality control details have been reported previously and are also detailed in **Supplementary Note 2**. Briefly, probes were removed based on (i) outliers from visual inspection of the log median intensity of the methylated versus unmethylated signal per array, (ii) a bead count <3 in more than 5% of samples, and (iii) ≥0.5% of samples having a detection P value >0.05 in Set 1 and ≥1% of samples having a detection P value >0.01 in Set 2. Samples were removed (i) if there was a mismatch between their predicted sex and recorded sex and/or (ii) if ≥1% of CpGs had a detection P value >0.05 in Set 1 and >0.5% of CpGs had a detection P value >0.01 in Set 2. Ten saliva samples were excluded from Set 1, along with three individuals who had answered “yes” to all self-reported health conditions. One person with suspected XXY genotype and seven genetic outliers were also removed [68]. The quality-controlled dataset comprised 9,537 individuals (n_{Set1}=5,087, n_{Set2}=4,450).

Over 98% of GS participants consented to allow access to electronic health records via data linkage. This includes GP records (Read 2 codes), prescription data, and hospital records (ICD codes). These data are available both retrospectively and prospectively from the time of initial blood draw, yielding up to approximately 14 years of follow-up data. For the current analysis, we considered incident disease data for 12 outcomes that are leading causes of mortality and morbidity (**Supplementary Note 3**). For each outcome, prevalent cases (ascertained via retrospective ICD and Read 2 codes or self-report from a baseline questionnaire) were excluded from the analyses. Self-report data was not available for the inflammatory bowel disease (IBD) outcome which meant that prevalent cases (i.e. recorded as having disease at baseline) were only excluded based on data linkage codes. Codes were excluded if they were not closely related to the 12 diseases and a summary of the included and excluded terms can be found in **Supplementary Tables 14-25**. Alzheimer's dementia was limited to those cases/controls with age of event/censoring greater than or equal to 65 years. Breast cancer analyses was restricted to females only. Recurrent and both major and moderate episodes of depression were included in the depression trait, whereas single episodes of depression were excluded. The diabetes trait was comprised of type 2 diabetes and more general diabetes codes such as diabetic retinopathy and diabetes mellitus with renal manifestation. All type 1 and juvenile cases of diabetes were excluded from the diabetes trait.

4.3 Ethics declarations

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.

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.

4.4 Elastic Net Protein Predictors

Penalised regression models were generated for each of the 90 neurology and 70 inflammatory proteins in the Lothian Birth Cohort 1936 using Glmnet (Version 4.0-2) [69] in R (Version 3.6.0) [70]. Protein levels were the outcome and there were 428,489 CpG features per model. An elastic net penalty was specified (alpha=0.5) and cross validation was applied. To reduce the possibility of overfitting in the cross-fold

validation step, each fold represented a single methylation processing batch or a combination of batches, with a range of 50-68 and 62-85 individuals per fold for the neurology and inflammatory analyses, respectively. Two folds were set aside as the test data and 10-fold cross validation was carried out on the remaining data (n_{train}=576, n_{test}=130 for neurology and n_{train}=725, n_{test}=150 for inflammatory). The optimal predictors, based on lambda values that minimised the mean cross-validation errors, were applied to the test data and we retained proxies with $r > 0.1$ and $P < 0.1$ against measured proteins in the holdout set (n=18 neurology and 20 inflammatory proxies). We generated new elastic net predictors for these 38 proteins, using 12-fold cross validation in order to maximise the sample size of the training dataset. All except one of the inflammatory folds represented a single batch. Of the 12 neurology folds, three were assigned to a singular batch and the remainder were composed of either 2-3 batches. Individuals per fold ranged from 62-85 and 49-81 for the inflammatory and neurology analyses, respectively.

Of the 38 DNAm proxies chosen from the optimisation step, 37 generated sufficient features in the elastic net regressions on the full Lothian Birth Cohort 1936. The remaining protein (TRAIL) only contained an intercept term (i.e. no nonzero features) and was therefore excluded. The 37 DNAm protein proxies from the 12-fold cross validation were then tested externally through correlations with STRADL (n=778, for both inflammatory and neurology panels) and LBC1921 (n=162, for the neurology panel) protein measurements. Comparisons were available for all 18 neurology proxies and 14 of the 20 inflammatory proxies. We identified 14 neurology proxies and 9 inflammatory proxies with $P < 0.05$ and $r > 0$ in at least one of the external test sets. Of the 5 inflammatory proxies which had no comparison available, 3 were included based on their performance in the holdout set ($P < 0.05$) and two (CCL11 and TNFB) were excluded (holdout set $P > 0.05$). IL6 did not achieve the required thresholds but has been shown to perform well against ELISA measures previously and was therefore included [14]. The 27 chosen proxies (13 inflammatory and 14 neurology) were then applied to the DNAm dataset in Generation Scotland (n=9,537). In both the STRADL and GS cohorts DNAm at each CpG site was scaled to have a mean of zero and variance of one prior to the projections.

A sensitivity analysis was performed where we re-ran the elastic net penalised regression models used in the initial optimisation step within LBC1936 with the addition of measured white blood cell proportions as features.

4.5 Associations with health in Generation Scotland

Cox proportional hazards regression models adjusting for age, sex, and methylation set were used to assess the relationship between the 27 selected DNAm protein proxies and 12 leading causes of morbidity and mortality in Generation Scotland. Models were run using the coxme package [71] (Version 2.2-16) in R version 3.6.3 with a kinship matrix specified to account for relatedness in the Set 1 methylation data. Time

to first event was ascertained through data linkage. Disease cases included those who had been diagnosed after baseline and subsequently died, in addition to those who received a diagnosis and remained alive. Control subjects were censored if disease free at time of death or at the end of the data linkage follow-up period. Fully-adjusted models included the following additional covariates measured at baseline: alcohol consumption (both units consumed in the previous week and a variable charting if this was more, less, or about the same as usual consumption); deprivation (assessed by the Scottish Index of Multiple Deprivation [72]); body mass index (kg/m^2); educational attainment (an 11-category ordinal variable: How many years altogether did you attend school or study full-time? (0: 0, 1: 1-4,2: 5-9, 3: 10-11, 4: 12-13, 5: 14-15, 6: 16-17, 7: 18-19, 8: 20-21, 9: 22-23, 10: 24+)); and a DNAm-based proxy for smoking status [73] which was well-correlated with the number of pack years individuals had smoked in the Generation Scotland cohort (r = 0.55, n = 9,311). Covariate phenotypes were prepared according to previous methodology [74]. A false discovery rate multiple testing correction was applied to the 324 protein-proxy:trait associations (27 proxies by 12 incident disease traits).

Proportional hazards assumptions were checked by running the fully-adjusted models and extracting Schoenfeld residuals (global test and a test for the protein-proxy variable) using the coxph and cox.zph functions from the survival package (Version 3.2-3) [75]. These models did not account for relatedness and random effects. For each association failing to meet the assumption (Schoenfeld residuals $P < 0.05$), a sensitivity analysis was run across yearly follow-up intervals. To ensure that underlying health conditions which had not been diagnosed in controls who had died post-baseline were not influencing the main findings, a sensitivity analysis was run which excluded these individuals from the Cox models.

In a further sensitivity analysis, adjusted Cox proportional hazards models were re-run with Houseman-estimated white blood cell proportions as covariates [77]. A further sensitivity analysis then assessed WBC-trait relationships independently of proxies by running the basic and fully-adjusted Cox models with WBC estimates as predictors, for each WBC-trait combination.

The correlation structures of the DNAm proxies with DNAm-based estimated white cell proportions and phenotypic information [77] were assessed using Pearson correlations and heatmaps using pheatmap (Version 1.0.12) [78]. The ggcorrplot package (Version 0.1.3) [79] was used to generate correlation structures between DNAm proxies. The psych package (Version 1.9.12) [80] was used to perform principal components analysis on multiple proxy measures which were associated with the same diseases. A network visualisation was produced using the igraph package (Version 1.2.5) [76].

4.6 Data availability

Lothian Birth Cohort 1936 data are available on request from the Lothian Birth Cohort Study, University of Edinburgh (simon.cox@ed.ac.uk). Lothian Birth Cohort 1936 data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

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

4.7 Code availability

All code is available at the following Gitlab repository: https://gitlab.com/marioni-group/dnam-protein-proxies. Access will be provided upon request.

5 Acknowledgements

The LBC1921 was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), a Royal Society–Wolfson Research Merit Award to I.J.D., and the Chief Scientist Office (CSO) of the Scottish Government's Health Directorates. The LBC1936 is supported by Age UK (Disconnected Mind project, which supports S.E.H.), the Medical Research Council (G0701120, G1001245, MR/M013111/1, MR/R024065/1, which supports S.R.C.), and the University of Edinburgh. Genotyping was supported by the Biotechnology and Biological Sciences Research Council (BB/F019394/1). Methylation typing in both the LBC1921 and LBC1936 was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. Proteomic analyses in LBC1936 and LBC1921 were supported for by the Age UK grant and NIH Grants R01AG054628 and R01AG05462802S1. This work was conducted in the Centre for Cognitive Ageing and Cognitive Epidemiology, which was supported by the Medical Research Council and Biotechnology and Biological Sciences Research Council (MR/K026992/1) and which supports I.J.D. We acknowledge Grant P2CHD042849 for supporting the Population Research Center at the University of Texas. This research was supported by Australian National Health and Medical Research Council (grants 1010374, 1046880 and 1113400) and by the Australian Research Council (DP160102400). P.M.V., and N.R.W. are supported by the NHMRC Fellowship Scheme (1078037, 1078901). A.F.M. is supported by the Australian Research Council Fellowship (FT200100837). P.M.V. was also funded by the Australian Research Council (DP160102400 and FL180100072). Lothian Birth Cohort 1921 and 1936 proteomic analyses were supported by a National Institutes of Health (NIH) research grant (R01AG054628) which also supports E.M.T-D and S.R.C.

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

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

6 Author contributions

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

7 Competing interests

R.E.M. has received payment from Illumina for presentations. All other authors declare no competing interests.

References

1. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13**, 942 (2017).

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

3. Hillary, R. *et al.* Multi-method genome- and epigenomewide studies of inflammatory protein levels in healthy older adults. *Genome Med.* **12**, (2020).

4. Zhao, T., Hu, Y., Zang, T. & Wang, Y. Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Front. Genet.* **10**, 1–8 (2019).

5. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 1–12 (2019).

6. Yao, C. *et al.* Genome□wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, (2018).

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

8. Aslam, B., Basit, M., Nisar, M. A., Khurshid, M. & Rasool, M. H. Proteomics: Technologies and their applications. *J. Chromatogr. Sci.* **55**, 182–196 (2017).

9. Gabay, C. & Kushner, I. Acute-phase proteins and other systemic responses to inflammation. *N. Engl. J. Med.* **340**, 448–54 (1999).

10. Sproston, N. R. & Ashworth, J. J. Role of C-reactive protein at sites of inflammation and infection. *Front. Immunol.* **9**, 1–11 (2018).

11. Matsumoto, H. *et al.* The clinical importance of a cytokine network in the acute phase of sepsis. *Sci. Rep.* **8**, 1–4 (2018).

12. Mazzio, E. A. & Soliman, K. F. A. Basic concepts of epigenetics impact of environmental signals on gene expression. *Epigenetics* **7**, 119–130 (2012).

13. Stevenson, A. *et al.* Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin. Epigenetics* **12**, 113 (2020).

14. Stevenson, A. J., Gadd, D. A., Hillary, R. F., Mccartney, D. L. & Campbell, A. Creating and validating a DNA methylation-based proxy for Interleukin-6. (2020).

15. Conole, E. L. S. *et al.* An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. 1–27 (2020).

16. World Health Organization. Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. (2018).

17. Hay, S. I. *et al.* Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1260–1344 (2017).

18. Alatab, S. *et al.* 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. *Lancet Gastroenterol. Hepatol.* **5**, 17–30 (2020).

19. Safiri, S. *et al.* Global, regional and national burden of rheumatoid arthritis 1990-2017: a systematic analysis of the Global Burden of Disease study 2017. *Ann. Rheum. Dis.* **78**, 1463–1471 (2019).

20. The MRC-IEU catalog of epigenome-wide association studies. Avaialable at: http://www.ewascatalog.org.

21. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).

22. Packer, M. *et al.* Effect of neprilysin inhibition on renal function in patients with type 2 diabetes and chronic heart failure who are receiving target doses of inhibitors of the renin-angiotensin system: a secondary analysis of the PARADIGM-HF trial. *Lancet Diabetes Endocrinol.* **6**, 547–554 (2018).

23. Nougué, H. *et al.* Effects of sacubitril/valsartan on neprilysin targets and the metabolism of natriuretic peptides in chronic heart failure: a mechanistic clinical study. *Eur. J. Heart Fail.* **21**, 598–605 (2019).

24. Jordan, J. *et al.* Improved Insulin Sensitivity With Angiotensin Receptor Neprilysin Inhibition in Individuals With Obesity and Hypertension. *Clin. Pharmacol. Ther.* **101**, 254–263 (2017).

25. Esser, N. & Zraika, S. Neprilysin inhibition: a new therapeutic option for type 2 diabetes? *Diabetologia* **62**, 1113–1122 (2019).

26. Ylä-Herttuala, S., Bridges, C., Katz, M. G. & Korpisalo, P. Angiogenic gene therapy in cardiovascular diseases: Dream or vision? *Eur. Heart J.* **38**, 1365–1371 (2017).

27. Anttila, V. *et al.* Synthetic mRNA Encoding VEGF-A in Patients Undergoing Coronary Artery Bypass Grafting: Design of a Phase 2a Clinical Trial. *Mol. Ther. - Methods Clin. Dev.* **18**, 464–472 (2020).

28. Angata, T., Nycholat, C. M. & Macauley, M. S. Therapeutic Targetting of Siglecs using Antibody- and Glycan- based Approaches. *Trends Pharmacol* **36**, 645–660 (2015).

29. Cassetta, L. *et al.* Human Tumor-Associated Macrophage and Monocyte Transcriptional Landscapes

Reveal Cancer-Specific Reprogramming, Biomarkers, and Therapeutic Targets. *Cancer Cell* **35**, 588-602.e10 (2019).

30. Panahi, Y. *et al.* Serum levels of fibroblast growth factor 21 in type 2 diabetic patients. *Acta Endocrinol. (Copenh).* **12**, 257–261 (2016).

31. Zhang, X. *et al.* Serum FGF21 levels are increased in obesity and are independently associated with the metabolic syndrome in humans. *Diabetes* **57**, 1246–1253 (2008).

32. Cheng, X., Zhu, B., Jiang, F. & Fan, H. Serum FGF-21 levels in type 2 diabetic patients. *Endocr. Res.* **36**, 142–148 (2011).

33. Shafaei, A., Khoshnia, M. & Marjani, A. Serum level of fibroblast growth factor 21 in type 2 diabetic patients with and without metabolic syndrome. *J. Med. Sci.* **15**, 80–86 (2015).

34. Chen, C. *et al.* High plasma level of fibroblast growth factor 21 is an independent predictor of type 2 diabetes: A 5.4-year population-based prospective study in Chinese subjects. *Diabetes Care* **34**, 2113–2115 (2011).

35. Li, L. *et al.* Plasma FGF-21 levels in type 2 diabetic patients with ketosis. *Diabetes Res. Clin. Pract.* **82**, 209–213 (2008).

36. Kim, K. H. *et al.* Metformin-induced inhibition of the mitochondrial respiratory chain increases FGF21 expression via ATF4 activation. *Biochem. Biophys. Res. Commun.* **440**, 76–81 (2013).

37. Ong, K. L. *et al.* The relationship of fibroblast growth factor 21 with cardiovascular outcome events in the Fenofibrate Intervention and Event Lowering in Diabetes study. *Diabetologia* **58**, 464–473 (2015).

38. Lenart-Lipińska, M., Matyjaszek-Matuszek, B., Gernand, W., Nowakowski, A. & Solski, J. Serum fibroblast growth factor 21 is predictive of combined cardiovascular morbidity and mortality in patients with type 2 diabetes at a relatively short-term follow-up. *Diabetes Res. Clin. Pract.* **101**, 194–200 (2013).

39. Xiao, Y. *et al.* Serum fibroblast growth factor 21 levels are related to subclinical atherosclerosis in patients with type 2 diabetes. *Cardiovasc. Diabetol.* **14**, 1–8 (2015).

40. Lee, C. H. *et al.* Role of circulating fibroblast growth factor 21 measurement in primary prevention of coronary heart disease among chinese patients with type 2 diabetes mellitus. *J. Am. Heart Assoc.* **6**, 1–9 (2017).

41. Lee, J. S. *et al.* Inverse association of plasma IL-13 and inflammatory chemokines with lung function impairment in stable COPD: A cross-sectional cohort study. *Respir. Res.* **8**, 1–10 (2007).

42. Ju, C. R. & Chen, R. C. Serum myostatin levels and skeletal muscle wasting in chronic obstructive pulmonary disease. *Respir. Med.* **106**, 102–108 (2012).

43. Mitra, A. *et al.* Association of Elevated Serum GM-CSF, IFN- γ, IL-4, and TNF- α Concentration with Tobacco Smoke Induced Chronic Obstructive Pulmonary Disease in a South Indian Population. *Int. J. Inflam.* **2018**, (2018).

44. Palange, P. *et al.* Circulating haemopoietic and endothelial progenitor cells are decreased in COPD. *Eur. Respir. J.* **27**, 529–541 (2006).

45. Watz, H. Next generation of anti-inflammatory therapy for COPD? *Eur. Respir. J.* **50**, 9–10 (2017).

46. Barnes, P. J. New anti-inflammatory targets for chronic obstructive pulmonary disease. *Nat. Rev. Drug Discov.* **12**, 543–559 (2013).

47. Hannon, E., Lunnon, K., Schalkwyk, L. & Mill, J. Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* **10**, 1024–1032 (2015).

48. Edgar, R. D., Jones, M. J., Meaney, M. J., Turecki, G. & Kobor, M. S. BECon: A tool for interpreting DNA methylation findings from blood in the context of brain. *Transl. Psychiatry* **7**, e1187-10 (2017).

49. Teixeira, A. L., Gama, C. S., Rocha, N. P. & Teixeira, M. M. Revisiting the role of eotaxin-1/CCL11 in psychiatric disorders. *Front. Psychiatry* **9**, 1–6 (2018).

50. García-Marchena, N. *et al.* Inflammatory mediators and dual depression: Potential biomarkers in plasma of primary and substance-induced major depression in cocaine and alcohol use disorders. *PLoS One* **14**, 1–17 (2019).

51. Simon, N. M. *et al.* A detailed examination of cytokine abnormalities in Major Depressive Disorder. *Eur. Neuropsychopharmacol.* **18**, 230–233 (2008).

52. Leighton, S. P. *et al.* Chemokines in depression in health and in inflammatory illness: A systematic review and meta-Analysis. *Mol. Psychiatry* **23**, 48–58 (2018).

53. Magalhaes, P. V. S. *et al.* Peripheral eotaxin-1 (CCL11) levels and mood disorder diagnosis in a population-based sample of young adults. *J. Psychiatr. Res.* **48**, 13–15 (2014).

54. Garzón-Tituaña, M. *et al.* The Multifaceted Function of Granzymes in Sepsis: Some Facts and a Lot to Discover. *Front. Immunol.* **11**, 1–12 (2020).

55. Dibbs, Z., Thornby, J., White, B. G. & Mann, D. L. Natural variability of circulating levels of cytokines and cytokine receptors in patients with heart failure: Implications for clinical trials. *J. Am. Coll. Cardiol.* **33**, 1935–1942 (1999).

56. Moldoveanu, A. I., Shephard, R. J. & Shek, P. N. Exercise elevates plasma levels but not gene expression of IL-1β, IL-6, and TNF-α in blood mononuclear cells. *J. Appl. Physiol.* **89**, 1499–1504 (2000).

57. Hillary, R. F. & Marioni, R. E. MethylDetectR - a software for methylation-based health profiling. *Wellcome Open Res* (2020).

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

59. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 1–11 (2019).

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

61. Loomans-Kropp, H. A. & Umar, A. Cancer prevention and screening: the next step in the era of precision medicine. *npj Precis. Oncol.* **3**, (2019).

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

63. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort profile update: The Lothian birth cohorts of 1921 and 1936. *Int. J. Epidemiol.* **47**, 1042–1060 (2018).

64. Seeboth, A. *et al.* DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936. *Clin. Epigenetics* **12**, 1–13 (2020).

65. McIlhagger, R. *et al.* Differences in the haematological profile of healthy 70 year old men and women: Normal ranges with confirmatory factor analysis. *BMC Blood Disord.* **10**, 2–9 (2010).

66. Smith, B. H. *et al.* Generation Scotland: The Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med. Genet.* **7**, 1–9 (2006).

67. Navrady, L. B. *et al.* Cohort profile: Stratifying Resilience and Depression Longitudinally (STRADL): A questionnaire follow-up of Generation Scotland: Scottish Family Health Study (GS: SFHS). *Int. J. Epidemiol.* **47**, 13-14g (2018).

68. Amador, C. *et al.* Recent genomic heritage in Scotland. *BMC Genomics* **16**, 1–17 (2015).

69. J, F., T, H. & R, T. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

70. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2020).

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

72. Scottish Government. The Scottish Index of Multiple Deprivation (SIMD); 1-20. (2016). Available from: http://www.gov.scot/ Resource/0050/00504809.pdf. Accessed 03 September 2020.

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

74. Hillary, R. *et al.* Epigenetic clocks predict prevalence and incidence of leading causes of death and disease burden. 1–12 (2020) doi:10.1101/2020.01.31.928648.

75. Therneau, T. M. A Package for Survival Analysis in R. R package version 3.2-3, https://CRAN.R-project.org/package=survival. 98784 (2020).

76. Csardi, G. & Nepusz, T. The igraph software package for complex network research. InterJournal, Complex Systems, 1695. http://igraph.org. (2006).

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

78. Kolde, R. Pheatmap : Pretty Heatmaps. R package version 1.0.12. *R Packag. version 1.0.8* (2019).

79. Kassambara, A. ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. R package version 0.1.3. https://CRAN.R-project.org/package=ggcorrplot. 2 (2019).

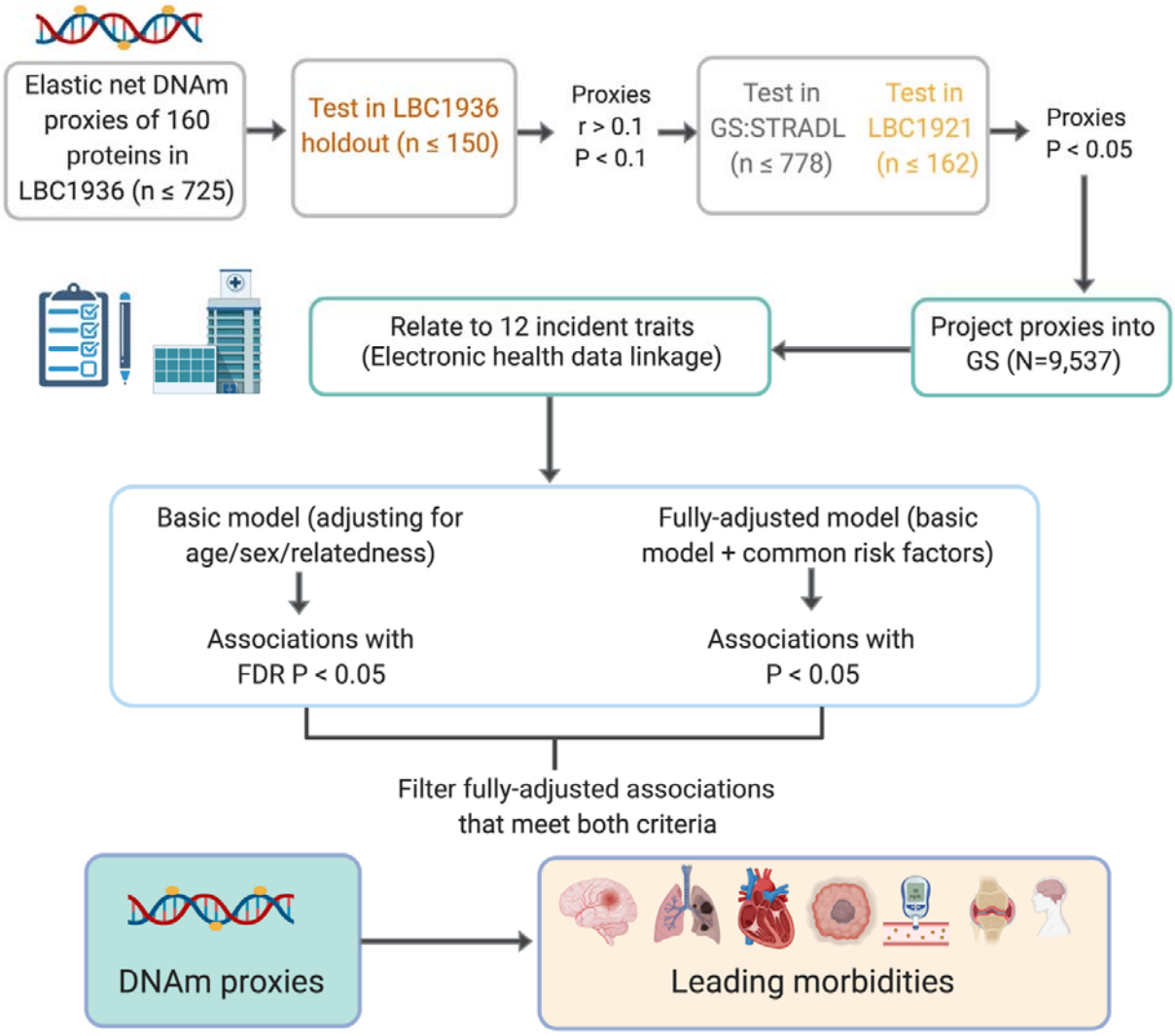
80. Revelle, W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.9.12. (2019).

Table 1: Summary of incident phenotypes in Generation Scotland.

Trait	N cases	N controls	Years to event (mean, sd)
Rheumatoid arthritis	65	9,281	6.1 (3.5)
Alzheimer's dementia	69	3,764	8.3 (2.7)
Bowel cancer	77	9,398	6.4 (3.2)
Depression	101	8,306	3.9 (3.3)
Breast cancer	129	5,355	6 (3.4)
Lung cancer	201	9,265	5.2 (3.1)
Inflammatory bowel disease	203	9,083	5 (3.5)
Stroke	317	9,023	6.5 (3.4)
COPD	346	8,939	6.2 (3.4)
Ischaemic heart disease	395	8,646	5.8 (3.3)
Diabetes	428	8,756	5.7 (3.4)
Pain	1494	5,341	5.2 (3.5)

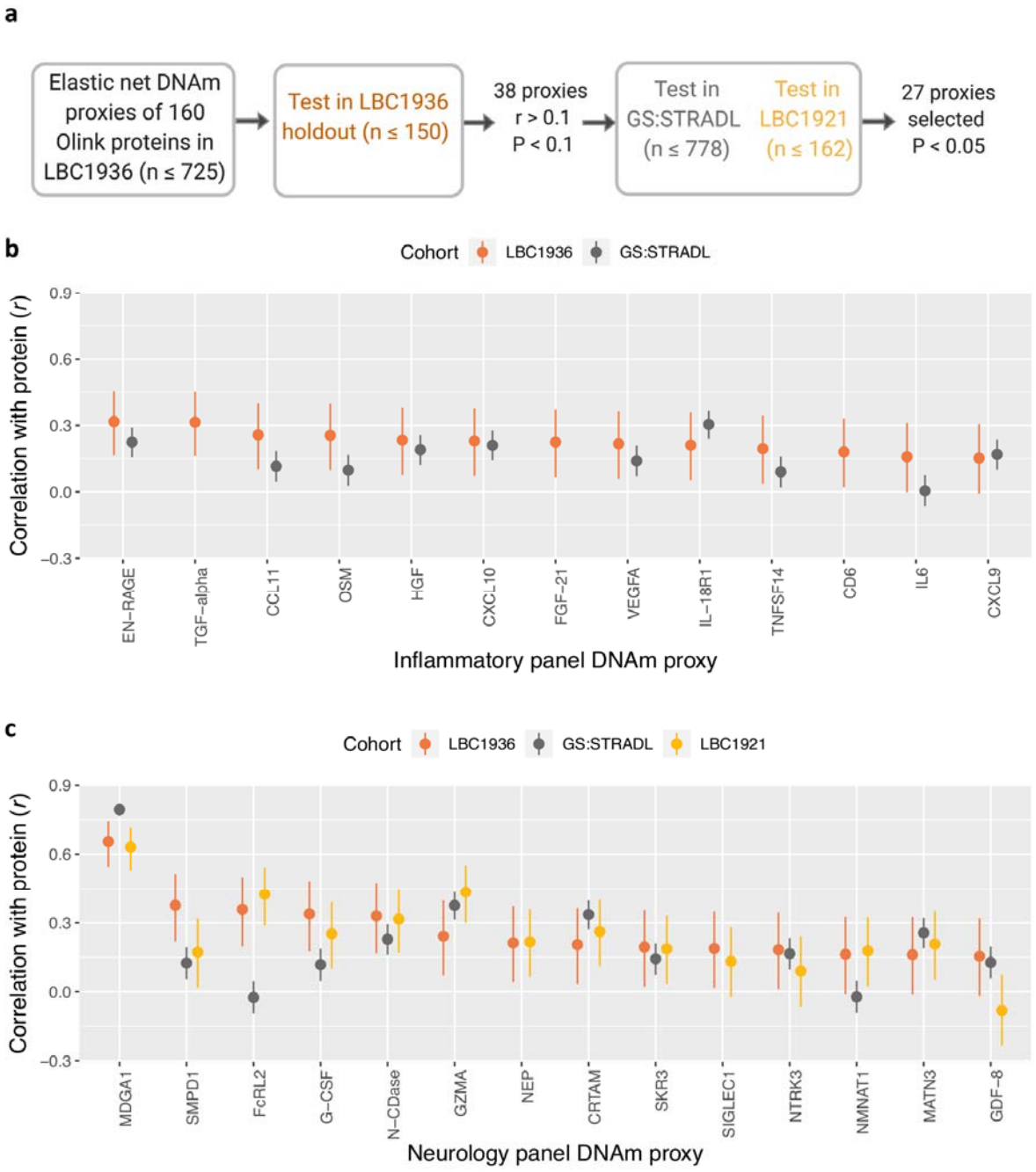
Counts are provided for the number of cases and controls for each incident trait in the Generation Scotland cohort (n=9,537). Mean time-to-event is summarised in years for each phenotype.

Fig. 1: A new framework for protein-by-proxy biomarker analyses.



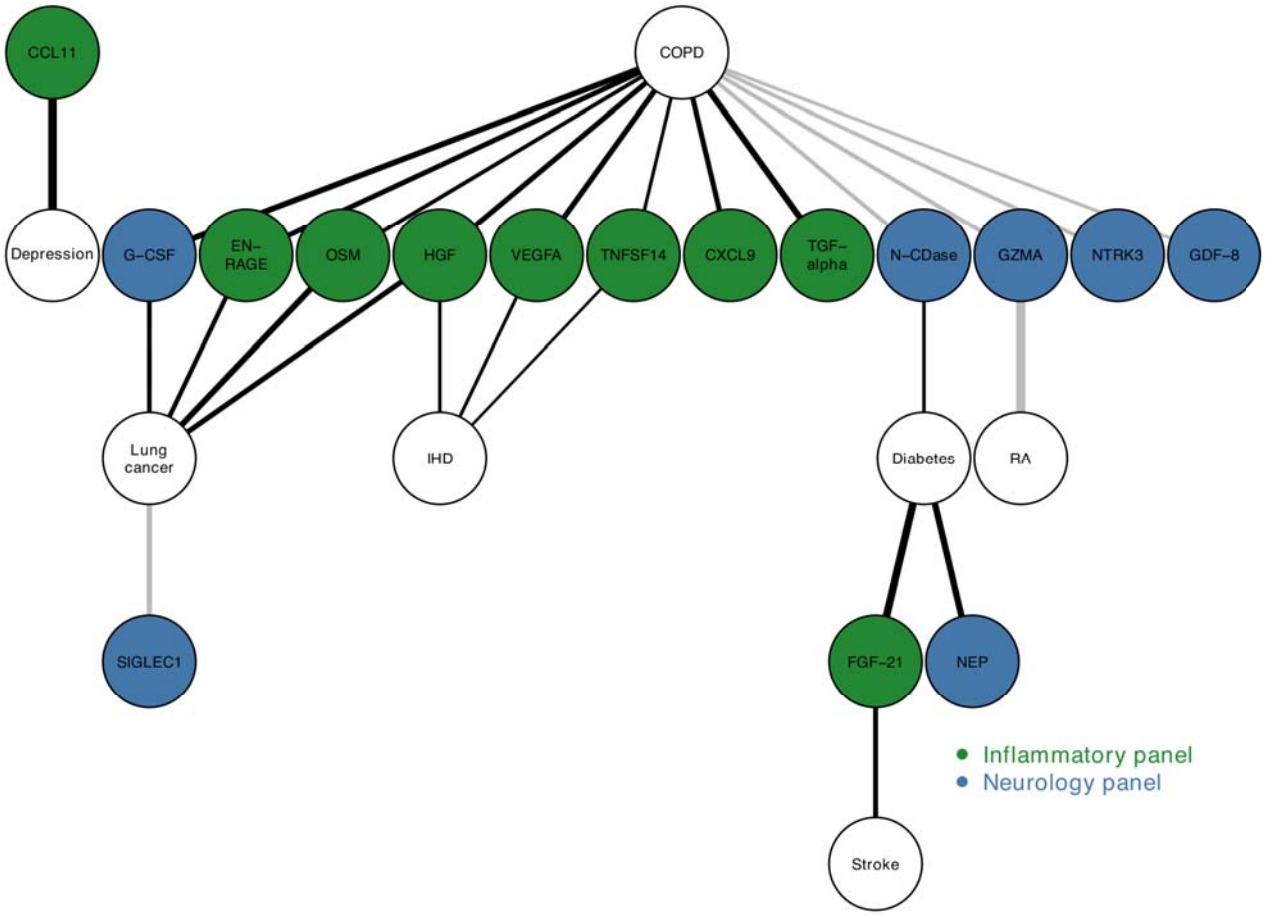
Elastic net penalized regression models were used to create DNAm proxies for protein expression. After evaluation in a holdout set, the optimal predictors ($r > 0.1$, $P < 0.1$) were trained on the full LBC1936 cohort and tested in a further two external test sets: GS:STRADL (n=778) and LBC1921 (n=162). Selected proxies were projected into Generation Scotland (N=9,537) and related to the incidence of 12 leading causes of morbidity and mortality, through GP and hospital electronic health data linkage over a period of up to 14 years. Proxy-disease associations with FDR-adjusted $P < 0.05$ in the basic and $P < 0.05$ in the fully-adjusted Cox mixed effects proportional hazards models were identified as significant. All protein measurements were rank-based inverse normal transformed and regressed on age, sex and four genetic principal components. All CpGs were scaled to have a mean of 0 and standard deviation of 1. Image created with BioRender.com.

Fig. 2: Validation of DNAm proxies for protein expression.



a, Results from the evaluation of DNAm proxies for protein expression in the LBC1936 cohort for Olink® neurology (train n=576, test n=130) and inflammatory (train n=725, test n=150) protein panels. Of the 160 proteins used in the proxy generation step, there were 38 proxies with $r > 0.1$ and $P < 0.1$ in the initial holdout set which were re-run using the full LBC1936 data as the training set. Of these 38, the inflammatory (**b**) and neurology (**c**) proxies shown performed well ($P < 0.05$) in either one or both of the subsequent GS:STRADL (n=778) and LBC1921 (n=162) test sets, wherever protein data was available for comparison. IL6 was previously validated against ELISA measurements and was therefore included despite poor performance. The correlation coefficients are plotted here for the final 27 proxies (13 inflammatory and 14 neurology) with upper and lower confidence intervals. Proxies are ordered by performance in the initial LBC1936 holdout set.

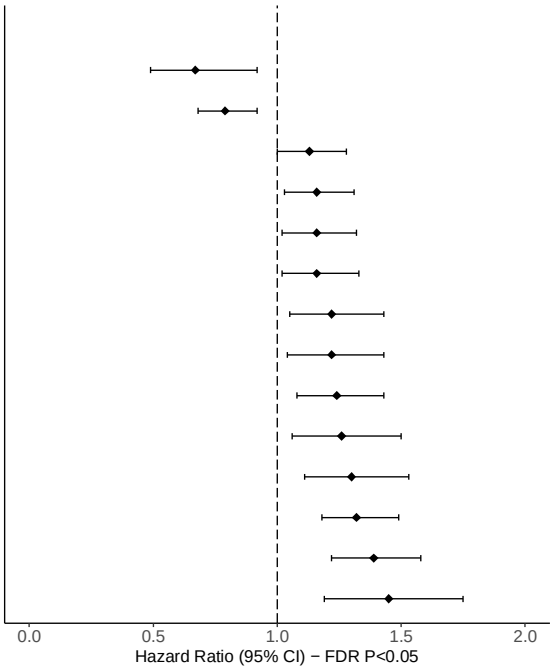
Fig. 3: DNAm proxy associations with incident disease.



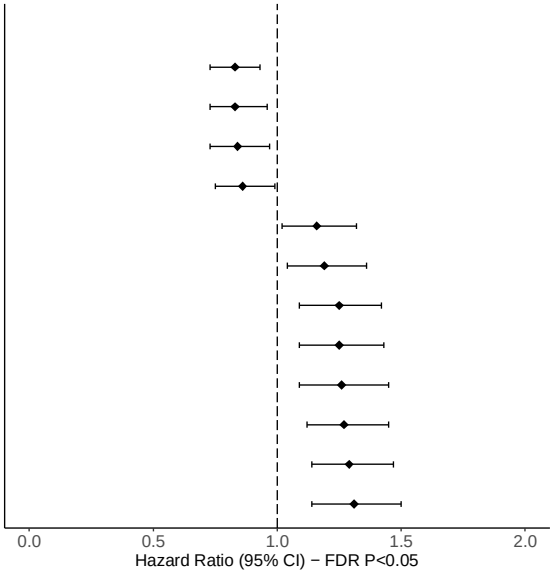
DNAm proxy measures for inflammatory (green) and neurology (blue) panel proteins which predicted incident diseases (white) in the Generation Scotland cohort (N=9,537). All 26 relationships between proxies and diseases presented here were significant with FDR-corrected $P < 0.05$ in the basic Cox mixed effects proportional hazard model and were also significant with $P < 0.05$ in the fully-adjusted model after accounting for age, sex and common risk factors. The connecting edges of the network represent proxy-disease relationships and their thickness is weighted according to log hazard ratios. Positive associations indicating increased risk of disease are shown as black edges. Protective, negative associations are shown as grey edges. IHD: ischaemic heart disease. RA: rheumatoid arthritis. COPD: chronic obstructive pulmonary disease.

Fig. 4: Hazard ratios for DNAm proxy-disease relationships.

DNAm Protein	Outcome	HR
GZMA	RA	0.67
SIGLEC1	Lung cancer	0.79
TNFSF14	IHD	1.13
N−CDase	Diabetes	1.16
VEGFA	IHD	1.16
HGF	IHD	1.16
G−CSF	Lung cancer	1.22
EN−RAGE	Lung cancer	1.22
FGF−21	Stroke	1.24
HGF	Lung cancer	1.26
OSM	Lung cancer	1.3
NEP	Diabetes	1.32
FGF−21	Diabetes	1.39
CCL11	Depression	1.45

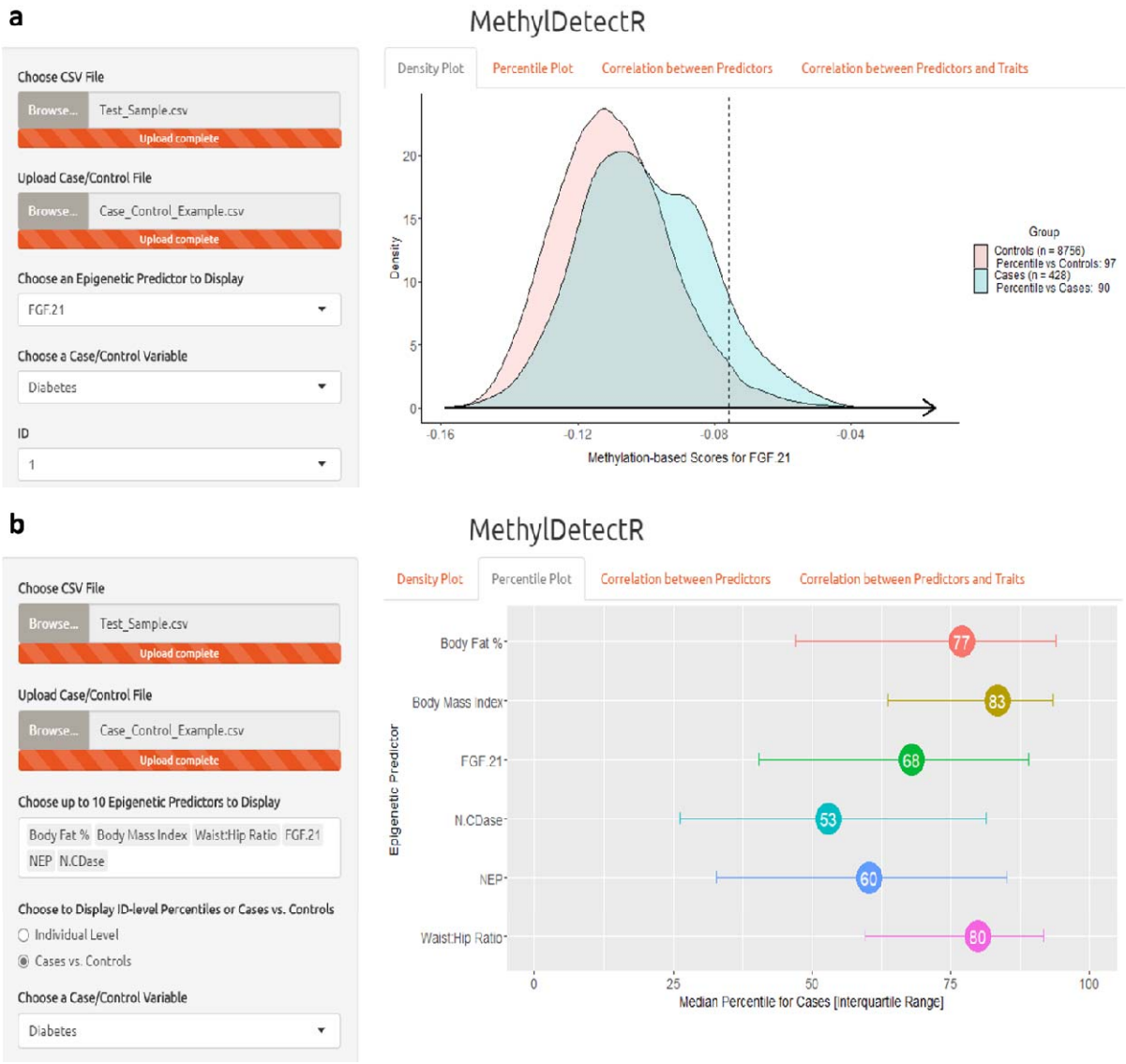


DNAm Protein	Outcome	HR
N−CDase	COPD	0.83
GZMA	COPD	0.83
NTRK3	COPD	0.84
GDF−8	COPD	0.86
TNFSF14	COPD	1.16
OSM	COPD	1.19
CXCL9	COPD	1.25
EN−RAGE	COPD	1.25
HGF	COPD	1.26
VEGFA	COPD	1.27
G−CSF	COPD	1.29
TGF−alpha	COPD	1.31



Hazard ratios for the Olink® inflammatory and neurology protein proxies (per SD increase) related to health outcomes in the Generation Scotland cohort (n=9,537). The 26 proxy-disease associations which were significant with FDR-corrected $P < 0.05$ in the basic Cox mixed effects proportional hazards model and significant with $P < 0.05$ in the fully-adjusted model after accounting for age, sex and common risk factors are presented with confidence intervals. IHD: ischaemic heart disease. RA: rheumatoid arthritis. COPD: chronic obstructive pulmonary disease.

Fig. 5: The MethylDetectR shiny app for protein proxy generation and visualisation.



(a) In MethylDetectR (https://shiny.igmm.ed.ac.uk/MethylDetectR_Demo/), users can obtain DNAm-based estimates of 27 blood protein levels, as well as epigenetic age, lifestyle and biochemical traits for all individuals in their sample. In this panel, the distributions of estimated FGF-21 levels are shown for individuals with incident diabetes (cases; blue) and for those who remained free of the disease (controls; pink) in the present study. The score for a selected individual is illustrated by the dotted vertical line. The user can subset the sample by age range and sex. (**b**) In this panel, the user can choose to view percentile ranks for a selected individual when compared to the remainder of the sample for up to 10 traits simultaneously. Alternatively, as shown here, the user can select an option to show the median percentile for cases in their sample with respect to a binary trait of interest that is uploaded by the user. Interquartile ranges for case percentile ranks are shown by horizontal bars. In this example, the median percentile for diabetes cases are plotted for a number of physical traits and the three protein proxies which were associated with incident diabetes in a fully-adjusted Cox model: FGF-21, N-CDase and NEP.